



**HAL**  
open science

# Phraséologie et genres textuels. Perspectives synchroniques et diachroniques

Julie Sorba

► **To cite this version:**

Julie Sorba. Phraséologie et genres textuels. Perspectives synchroniques et diachroniques : Mémoire de synthèse présenté pour l'obtention de l'Habilitation à diriger des recherches. Linguistique. Université Grenoble Alpes (UGA), 2022. tel-03891983

**HAL Id: tel-03891983**

**<https://hal.science/tel-03891983>**

Submitted on 9 Dec 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

**Phraséologie et genres textuels**  
**Perspectives synchroniques et diachroniques**

Mme Julie SORBA

Mémoire de synthèse

Volume 1

présenté pour l'obtention de l'Habilitation à diriger des recherches

le 10 juin 2022

à l'Université Grenoble Alpes – Laboratoire LiDiLEM – École doctorale LLSH

Spécialité : Sciences du Langage

MEMBRES DU JURY

M. Peter BLUMENTHAL, Professeur émérite, Universität zu Köln (Rapporteur)

Mme Dominique LAGORGETTE, Professeure, Université Savoie Mont Blanc, LLSETI (Examinatrice)

M. Dominique LEGALLOIS, Professeur, Université Paris 3 Sorbonne Nouvelle, LATTICE (Rapporteur)

M. Dominique LONGREE, Professeur, Université de Liège, LASLA (Rapporteur)

Mme Iva NOVAKOVA, Professeure, Université Grenoble Alpes, LiDiLEM (Garante)

Mme Agnès TUTIN, Professeure, Université Grenoble Alpes, LiDiLEM (Présidente)



*À Lucien Sorba, mon père trop tôt disparu.*



# Sommaire

Préambule : une discipline, un objet d'étude .....	9
Introduction .....	21
<b>Partie 1. La phraséologie : le saut du syntagme au texte .....</b>	<b>23</b>
Chapitre 1. L'approche lexicologique à l'interface de la sémantique et de la syntaxe.....	25
1.1 Des paroles ailées aux paroles entravées .....	25
1.1.1 La question des contraintes et du figement linguistique .....	25
1.1.2 Le principe de l'idiomaticité.....	28
1.1.3 Le cas des formules homériques .....	31
1.2. Les unités phraséologiques : quelle(s) typologie(s) ?.....	34
1.2.1 Qu'est-ce qu'une unité phraséologique ? .....	34
1.2.2 Comment classer les unités phraséologiques ? .....	36
1.2.3 Le cas de la collocation : perspectives synchroniques et diachroniques.....	41
Chapitre 2. Au-delà du syntagme : « Vers une extension du domaine de la phraséologie » .....	47
2.1. Le tournant phraséologique .....	47
2.1.1 La théorie du <i>lexical priming</i> .....	48
2.1.2 La contrainte dépasse le cadre du syntagme .....	50
2.2 La « portée textuelle » d'une unité lexicale.....	51
2.2.1 Actualiser une acception .....	51
2.2.2 Structurer l'environnement textuel .....	53
2.3 Modélisation du lien entre UP et texte : le cas des lexies d'affect .....	55
2.3.1 L'impact des lexies d'affect sur l'organisation discursive .....	55
2.3.2 Le rôle des lexies d'affect dans l'argumentation .....	57
<b>Partie 2. Pour un renouveau de l'approche générique .....</b>	<b>65</b>
Chapitre 3. La question du genre : comment saisir le générique ? .....	67
3.1 Où saisir le générique ? .....	67
3.1.1 Le texte, « un objet de haute complexité ».....	67
3.1.2 Sur quels critères se fonde le jugement de textualité ? .....	70

3.2. Comment définir le genre ? .....	72
3.2.1 Le genre : la mise en œuvre d'une compétence cognitive.....	73
3.2.2 Le genre : une partie constitutive du discours.....	75
3.2.3 Le genre : une articulation des contraintes linguistiques et situationnelles.....	79
Chapitre 4. La typologie des genres : une entreprise vouée à l'échec ? .....	83
4.1 Le cas de la typologie des genres littéraires .....	84
4.1.1 Aperçu historique de la conceptualisation du genre littéraire .....	85
4.1.2 Les genres littéraires en transformation permanente.....	88
4.1.3. Les genres littéraires comme « réalité interdiscursive globale » .....	90
4.2 Pourquoi classer les productions textuelles ? .....	92
4.2.1 Classer pour accéder à une unité structurante .....	93
4.2.2 Classer pour accéder à la variation linguistique .....	95
4.2.3 Classer pour accéder au changement linguistique.....	97
4.3 Comment classer les productions textuelles ? .....	99
4.3.1 Classer les textes selon la spécificité de leurs unités linguistiques .....	100
4.3.2 Classer les textes en combinant plusieurs critères linguistiques et extralinguistiques...	103
4.3.3 Construire une typologie ouverte pour rendre compte de la généricité .....	107
<b>Partie 3. Ce que la phraséologie dit des genres textuels grâce à la linguistique de corpus outillés</b> .....	113
Chapitre 5. Les enjeux de la linguistique de corpus .....	114
5.1 Qu'est-ce que la linguistique de corpus ? .....	114
5.1.1 Une discipline fondée sur l'observation .....	115
5.1.2 Les deux approches des corpus : <i>corpus-based</i> et <i>corpus-driven</i> .....	116
5.2 Les corpus : comment et pourquoi les constituer ?.....	120
5.2.1 Le corpus comme ensemble de textes authentiques réunis en vue d'un objectif.....	121
5.2.2 La corpus comme ensemble représentatif .....	122
5.2.3 Le corpus comme un ensemble de choix .....	124
5.3. Trois exemples de corpus pour l'étude de la phraséologie en diachronie .....	126
5.3.1 Agrandir un corpus romanesque : le corpus PhraseoRom_19e_fr (1830-1899) .....	126

5.3.2 Créer un corpus romanesque pour des états de langue anciens : le corpus PhraseoRoChe (1200-1650) .....	130
5.3.3 Créer un corpus épistolaire : le corpus PhraseoCorr (1640-1960) .....	135
Chapitre 6. L'émergence de nouveaux observables phraséologiques dans les corpus outillés .....	143
6.1 Explorer des corpus outillés .....	144
6.1.1. Le degré zéro de l'outillage : le concordancier .....	145
6.1.2 La fouille de corpus élaborée : le Lexicoscope .....	149
6.2 Étudier la phraséologie sur corpus outillé pour caractériser les genres textuels .....	154
6.2.1 De l'arbre lexicosyntaxique récurrent au motif .....	155
6.2.2 Des motifs pour contraster des sous-genres romanesques : études de cas .....	160
Conclusion .....	167
Annexes .....	169
I. Liste des travaux scientifiques .....	170
II. Références bibliographiques .....	182
III. Références sitographiques .....	198
IV. Curriculum vitae .....	199
V. Une expérience de la valorisation en SHS : le projet Certiredac et son prolongement dans écrit + .....	208



## **Préambule : une discipline, un objet d'étude**

« Tout seul on va plus vite ; ensemble, on va plus loin. »

Proverbe africain

Le préambule a été supprimé de la version mise en ligne sur HAL. Néanmoins, la pagination est conforme à celle du mémoire original tel qu'il a été présenté et soutenu.

*Page originale supprimée*





# Phraséologie et genres textuels

## Perspectives synchronique et diachronique

« If I have seen further it is by standing on the shoulders of Giants »

Isaac Newton letter to Robert Hook, February 5, 1675<sup>1</sup>

---

<sup>1</sup> Turnbull, H.W. ed. (1959). *The Correspondence of Isaac Newton: 1661-1675*, Volume 1, London, UK: Published for the Royal Society at the University Press. p. 416.



La rédaction de ce travail suit « Les rectifications de l'orthographe » publiées dans les Documents administratifs du *Journal officiel de la République française* le 6 décembre 1990.



## Introduction

Ce mémoire d'habilitation à diriger des recherches est l'occasion de présenter ma trajectoire intellectuelle depuis la préparation de ma thèse, soutenue en février 2010, mais aussi de revisiter une quinzaine d'années de recherche à la lumière des nouvelles connaissances et compétences acquises. Mon travail qui portait initialement sur le lexique, les relations sémantiques et le fonctionnement des associations lexicales a progressivement étendu son champ d'analyse aux plans textuel et générique tout en s'inscrivant dans une approche de linguistique de corpus outillés.

La saisie conjointe de ces objets me permet d'opérer désormais dans une perspective interdisciplinaire puisque mon objectif est d'observer comment la régularité linguistique entre en relation fonctionnelle avec la régularité générique et de mettre ainsi au jour les liens qu'elles entretiennent l'une avec l'autre sur le temps long. J'ai choisi de traiter cette problématique par l'entrée de la phraséologie en me demandant comment les unités phraséologiques permettent de caractériser les genres textuels et de les distinguer les uns des autres. Pour réaliser ce programme, j'ai opté pour une démarche de linguistique de corpus outillés afin de montrer l'émergence possible de nouveaux observables.

Ce travail est donc l'occasion de poser des jalons théoriques concernant les deux notions clés complexes que sont la phraséologie (partie 1) et le genre textuel (partie 2) en présentant l'apport de la linguistique de corpus outillés à ces recherches croisant deux champs disciplinaires (partie 3). Je montrerai comment mes travaux s'inscrivent dans ce cadre et la contribution qu'ils y ont apportée. Les exemples présentés tout au long de ce mémoire sont donc issus des différents corpus que j'ai explorés depuis le début de ma recherche (indo-aryen ancien, grec ancien, latin, ancien français, français classique et moderne ; corpus poétique, romanesque, journalistique, épistolaire)<sup>2</sup>. Pour que le lecteur puisse retrouver facilement une référence, la liste de mes travaux (annexe I) a été dissociée de la bibliographie générale (annexe II) et de la sitographie qui répertorie les sites des outils et corpus numériques mentionnés au fil de ce mémoire (annexe III).

---

<sup>2</sup> Toutes les traductions proposées sont personnelles (sauf mention contraire). Les références des éditions des sources utilisées sont indiquées en note de bas de page.



## Partie 1. La phraséologie : le saut du syntagme au texte

« La littérature sur la phraséologie est énorme, ce qui exclut toute tentative d'en rendre compte ici même de façon superficielle. » (Mel'čuk 2013, p. 129)

Selon Sinclair (2008), la phraséologie a été un champ de recherche longtemps ignoré, du moins par les chercheurs occidentaux, pour deux raisons : « One is that phraseology does not make a sharp division between grammar and lexis/semantics. [...] Secondly, phraseology, in sharp contrast to most grammars, prioritises syntagmatic patterns over paradigmatic ones. » (2008, p. xv). L'absence d'un périmètre clairement délimité et une préoccupation d'ordre essentiellement syntagmatique auraient ainsi longtemps été des obstacles pour penser ce phénomène linguistique complexe.

De fait, la phraséologie est traversée d'une double tension que González Rey (2015) résume, de manière pertinente, sous la forme d'une série de questions :

Qu'est-ce donc que ce corps étrange à la langue, composé d'éléments polylexicaux, figés et parfois incohérents du point de vue sémantique, qu'il faut apprendre par cœur comme s'ils s'agissait de mots isolés ? Fait-il partie d'un lexique qui le relègue à l'article des bizarreries de la langue ou bien d'une grammaire où sont entassés les cas d'anomalies syntaxiques ? Ou s'agit-il, au contraire, d'un domaine linguistique essentiel dont aucune langue ne peut faire l'économie pour expliquer son identité face aux autres langues et face à elle-même ? (González Rey 2015, p. 15-16)

Dans cet extrait, la phraséologie est présentée à la fois comme un intrus dans la langue (« étrange », « incohérents », « bizarreries », « anomalies ») d'un point de vue intralinguistique, sur les niveaux sémantique et syntaxique, et comme un élément qui façonne l'identité de celle-ci dans une perspective interlinguistique. La question du statut de la phraséologie est également posée : constitue-t-elle un élément du lexique<sup>3</sup> parmi d'autres ou un

---

<sup>3</sup> Nous désignons conventionnellement par *lexique* « l'ensemble indéterminé des éléments signifiants stables (mots, locutions...) d'une langue, considéré abstraitement comme une des composantes formant le code de cette langue. » (*Le Petit Robert de la langue française*, s.v.). Ces « éléments signifiants stables », autrement nommés *unités lexicales* (par ex., Mortureux 2008, p. 7), *lexème* (par ex., Greimas 1986 [1966], p. 27), *signe* (par ex., Polguère 2008, p. 21) ou plus commodément *mot* (par ex., Lehmann & Martin-Berthet 1998, p. 3), forment une classe ouverte. En effet, « toute langue d'une part voit constamment une partie de son vocabulaire tomber en désuétude, d'autre part acquiert sans cesse des lexèmes nouveaux, soit qu'elle les construise avec ses moyens

« domaine » autonome de la linguistique ? C'est la seconde voie que défend Pamies Bertrán (2018) considérant que ce domaine a progressivement émergé « par élimination, à partir de ses propriétés “négatives”, englobant ce que rejettent d'autres domaines métalinguistiquement antérieurs, la syntaxe et le lexique. » (p. 221).

Dans cette première partie, je montrerai pour ma part que la phraséologie en vient progressivement à se constituer comme une nouvelle discipline linguistique à partir d'une approche lexicologique à l'interface de la sémantique et de la syntaxe (chapitre 1). La linguistique a en effet opéré un « tournant phraséologique » (Legallois & Gréa 2006, p. 5) pour étendre son domaine au-delà des frontières du syntagme qui lui étaient originellement assignées comme je le préciserai dans le deuxième chapitre.

---

propres, soit qu'elle les emprunte à une autre langue. » (Apothéloz 2002, p. 11). Le terme *vocabulaire* est également employé en tant que synonyme de *lexique* comme c'est le cas dans cette citation.

## **Chapitre 1. L'approche lexicologique à l'interface de la sémantique et de la syntaxe**

L'approche lexicologique de la phraséologie traite ce phénomène en tant qu'élément du lexique, à l'interface de la sémantique et de la syntaxe. Cette conception, tout particulièrement représentée dans la recherche française, propose de définir la phraséologie comme « la congruence à la fois syntaxique et sémantique qui lie les unités lexicales entre elles pour donner lieu à des unités polylexicales qui se distinguent par une fixité d'emploi conditionnant leur fonctionnement interne et leur combinatoire externe. » (Grossmann, Mejri & Sfar 2017, p. 7). Dans ce cadre, l'interface entre sémantique et syntaxe opère aussi bien lors de la genèse de l'unité polylexicale – la « congruence » indique la pertinence de l'association des éléments de ladite unité sur les deux plans à un moment donné et pour une raison donnée – que lors de son usage, la « fixité d'emploi » impliquant des contraintes. Dans ce premier chapitre, nous interrogerons donc cette notion de contraintes linguistiques au coeur de l'approche lexicologique de la phraséologie (section 1.1) avant d'aborder la question des unités phraséologiques et de leur typologie (section 1.2).

### **1.1 Des paroles ailées aux paroles entravées**

Mon objectif est ici de montrer, d'une part, que la question de la contrainte linguistique interroge depuis longtemps les linguistes qui travaillent sur des états de langue aussi bien anciens que contemporains (1.1.1) et, d'autre part, la façon dont ceux-ci la traitent dans le cadre de la réflexion sur l'idiomaticité (1.1.2). C'est en somme relier la célèbre formule homérique des ἔπεα πτερόεντα à la formulation moderne du figement comme « parole entravée », pour reprendre le titre du volume coordonné par Anscombe & Mejri (2011). J'illustrerai d'ailleurs ma démonstration en m'appuyant sur mes premiers travaux réalisés sur les formules homériques (1.1.3).

#### **1.1.1 La question des contraintes et du figement linguistique**

La présence de contraintes linguistiques a été théorisée en particulier dans le cadre moderne des recherches sur le figement linguistique. Comme le rappelle Anscombe (2011), « on ne peut définir une catégorie linguistique 'figement' que dans et par rapport à une théorie de la langue » et c'est précisément à la grammaire générative transformationnelle qu'on doit de voir les phénomènes de figement « comme une catégorie linguistique (i.e. faisant partie du système et devant donc être engendrés dans ce système). » (p. 18). Mejri (1997) a souligné la complexité

de ce phénomène qui relève à la fois de la syntaxe et de la sémantique mais aussi de la morphologie ou encore de la prosodie (p. 17-18). Les études sur les textes antiques issus de différentes traditions orales se sont intéressées à ce phénomène sous ces différents angles : le figement se trouve aussi bien dans le carcan métrique de l'hexamètre dactylique<sup>4</sup> des épopées grecques homériques que dans les hymnes védiques en indo-aryen ancien dont la composition est assujettie à des normes rigoureuses. Pinault (1995-1996) définit d'ailleurs ces derniers comme « des productions figées de 'performances' différentes, par des poètes d'une même école, qui recourent aux mêmes procédés pour s'adresser à tel ou tel dieu du panthéon védique. » (p. 307-308). Tout le travail fondateur de Parry (1928a) sur les formules homériques a montré comment le figement était une nécessité pour les aèdes grecs :

Pour [Homère], comme pour tous les aèdes, versifier c'était se souvenir. C'était se souvenir des mots, des expressions, des phrases entendus dans le récit des aèdes qui lui avaient enseigné le style traditionnel de la poésie héroïque. C'était se souvenir de la place ou des places que les mots et les expressions traditionnels occupaient dans le moule complexe de l'hexamètre. C'était enfin se souvenir des artifices innombrables qui permettaient de combiner ces mots et ces expressions en phrases parfaites et en vers dactyliques de six pieds pour exprimer les idées convenant aux récits des gestes légendaires. Car de génération en génération les aèdes avaient conservé des mots et des expressions qui, heureusement trouvés, pouvaient resservir dans la poésie héroïque. (Parry 1928a, p. 6).

Dans ces deux cas (grec et indo-aryen anciens), on ne peut que souscrire aux propos d'Anscombe sur la définition du figement comme « catégorie linguistique » inhérente aux systèmes qui l'engendrent.

G. Gross (1996) considère ce phénomène de figement comme l'« un des plus importants dans les langues » (p. 9) et comme « constitutif de la langue elle-même » (p. 144). Il en définit l'origine de la manière suivante :

[Le figement linguistique] est caractérisé par deux niveaux. Il s'agit, d'abord, avec des mots existants par ailleurs dans la langue, de créer des unités nouvelles, des désignations nouvelles. Pour qu'on puisse dénommer un concept nouveau à partir d'un agglomérat d'autres mots, il faut que ces mots perdent d'abord leur actualisation, c'est-à-dire ce qui les lie à une situation donnée. Le groupe n'est plus alors un syntagme régulier mais peut être appelé une *locution*. À ce niveau,

---

<sup>4</sup> Le rythme dactylique a connu un franc succès au sein de l'hexamètre dactylique, le mètre de l'épopée. Celui-ci est constitué de 6 pieds dont chaque élément comprend un temps marqué (une syllabe longue) suivi d'un temps faible (une syllabe longue pouvant se résoudre en deux brèves), le dernier demi-pied faible étant une syllabe unique et indifférente (voir Dain 1965, p. 51).

il n'est pas nécessaire que la suite obtenue soit figée sémantiquement. [...] Le second niveau est caractérisé par le fait que les mots polylexicaux peuvent être opaques, c'est-à-dire forgés de façon non reproductible. (G. Gross 1996, p. 144-145).

Cette approche lexicologique s'intéresse à la formation de nouvelles unités polylexicales à partir de mots préexistants.

Comme G. Gross l'expose dans cet extrait, le critère sémantique ne suffit pas à identifier un fait lexical relevant du figement : c'est l'« absence de libre actualisation » de chaque élément lexical qui est déterminante. Pour mettre en évidence la propriété de « libre actualisation », G. Gross prend l'exemple de la locution verbale *prendre une veste* : une lecture compositionnelle permet l'actualisation de *veste* dans *Nicolas a pris (une, sa, cette, ta) veste avant de sortir* tandis que dans la suite figée *Nicolas a pris une veste aux élections présidentielles*, « le substantif *veste* ne réfère à aucun vêtement, ce que met en évidence la contrainte sur le déterminant. » (1996, p. 14). Ces recherches proposent une approche lexicologique de la phraséologie à l'interface de la syntaxe et de la sémantique. De même, la locution *un cordon bleu*, apparue au début du 19<sup>e</sup> siècle pour désigner « [un] cuisinier, [ou une] cuisinière très habile », a bien été créée à partir de mots existants antérieurement (le nom masculin *cordons* et l'adjectif *bleu*) mais l'adjectif *bleu* ne peut pas recevoir d'actualisation (*\*un cordon très bleu, \*un cordon bleu roi*) et le sens est complètement opaque (*un cordon bleu* n'est ni un cordon ni bleu). En revanche, une suite comme *vin blanc* n'est pas complètement figée : sémantiquement, la locution *vin blanc* désigne bien un vin et sa couleur mais celle-ci tire plutôt sur le jaune que le blanc. Ces deux exemples servent ainsi à montrer que le figement peut être considéré comme un phénomène scalaire.

Dans ce cadre, les unités polylexicales plus ou moins figées sont mises en évidence par la comparaison avec ce qui serait un état régulier et libre de la langue : les faits linguistiques relevant de ce phénomène de figement se caractériseraient comme un écart, sur les plans sémantique ou syntaxique, par rapport à cette norme<sup>5</sup>. Mais comment déterminer les caractéristiques de cet état régulier et libre qui permettrait de définir en creux ces unités ?

---

<sup>5</sup> Dans les grandes lignes, cette conception n'est pas très éloignée de celle des « figures du discours » développée par Fontanier au 19<sup>e</sup> s. : « Les figures du discours sont les traits, les formes ou les tours plus ou moins remarquables et d'un effet plus ou moins heureux, par lesquels le discours, dans l'expression des idées, des pensées ou des sentiments, s'éloigne plus ou moins de ce qui en eût été l'expression simple et commune. » (Fontanier 1977 [1821], p. 64, nous soulignons). En effet, une unité polylexicale, de même qu'une figure, s'écarterait d'une norme antérieure et cet écart est susceptible de degré (« s'éloigne plus ou moins »). De plus, la figure est ici définie

Au fondement de cette question se trouve le vif débat sur les notions chomskyennes de « grammaticalité », d'« acceptabilité » et d'« intuition des locuteurs ». Ces notions sont aujourd'hui vivement critiquées. Selon Cori & David 2008 (p. 115-116), les critiques de la grammaticalité et de l'acceptabilité portent sur leur caractère binaire (acceptable/non-acceptable), sur leur caractère subjectif (le jugement introspectif des linguistes n'est pas fiable) ainsi que sur l'absence de précision quant à leur objet (« on ne sait (jamais) précisément ce sur quoi porte le jugement, étant donné qu'interviennent et des facteurs contextuels et notre capacité à imaginer des contextes favorables. »). De son côté, même si la notion d'usage ne va pas de soi, le linguiste britannique Quirk s'est attaché à montrer l'inadéquation fréquente entre les croyances des locuteurs et leurs usages réels : les locuteurs n'arrivent pas à distinguer nettement les usages qu'ils acceptent de ceux qu'ils produisent. Ces résultats l'ont amené à considérer de fait l'acceptabilité « comme un phénomène multifacé et continu qui ne peut se réduire à un jugement strict en oui ou non. » (Léon 2008, p. 21-22).

La notion de figement est toujours âprement discutée par les nombreuses études qui mettent en évidence la variation dans des expressions traditionnellement considérées comme figées (voir par ex., Mogorrón Huerta 2017 à propos des constructions verbales figées), voire abandonnée « à la suite d'études de la préformation dans d'importants corpus [...] du fait qu'aucune modification ne semble au fond exclue. » (Schmale 2013, p. 31). Pour notre part, à l'instar de Mel'čuk (2013), nous préférons désormais parler de *contraintes* qui pèsent sur les unités lexicales et leur combinatoire et garder le terme de *figement* pour désigner le phénomène scalaire. Cette question des contraintes linguistiques est au cœur de la réflexion britannique contextualiste sur le principe de l'idiomaticité.

### **1.1.2 Le principe de l'idiomaticité**

Lorsqu'il expose les deux principes mis en œuvre dans le processus herméneutique, « the open-choice principle » et « the idiom principle », Sinclair souligne que le premier ne suffit pas à rendre compte de la construction du sens dans sa totalité car c'est seulement virtuellement que n'importe quelle unité lexicale peut apparaître à n'importe quelle place d'un énoncé. C'est bien le second principe qui permet de rendre compte des « bizarreries de la langue » mentionnées plus haut :

---

comme un fait de style (« tours plus ou moins remarquables et d'un effet plus ou moins heureux ») et c'est bien dans un *Traité de stylistique* que Bailly théorise, en premier, la notion de phraséologie.

The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. (Sinclair 1991, p. 110)

Dans le premier cas, celui de l'*open choice principle*, le locuteur dispose d'un grand choix de possibilités lexicales seulement restreintes par les contraintes grammaticales ; dans le second cas, *idiom principle*, le locuteur ne dispose que d'un choix limité de syntagmes préconstruits. Ainsi, l'impression d'« anomalie », notée plus haut par González Rey, proviendrait, d'une part, du fait que le locuteur n'a pas vraiment le choix de ses mots car il est souvent contraint d'utiliser une unité lexicale préconstruite<sup>6</sup>, et, d'autre part, du fait que le fonctionnement interne de cette unité préconstruite ne correspond pas systématiquement à celui des éléments dont elle est constituée. À titre d'exemple, Sinclair (1991) prend la locution anglaise *of course* dans laquelle l'unité *of* ne se comporte pas comme la préposition décrite par les grammaires, en principe placée après le nom-tête d'un syntagme nominal (*the principle of idiom*) ou au sein dans un quantifieur (*a pint of*). Ces « anomalies » ne sont pas seulement d'ordre syntaxique, comme dans cet exemple commenté par Sinclair, mais peuvent aussi être d'ordre sémantique. Par exemple, l'adjectif de couleur *bleu* n'a plus vraiment son sens usuel dans la locution française *cordons bleus* lorsqu'elle désigne, de nos jours, dans la langue familière « [un] cuisinier, [ou une] cuisinière très habile » ou, dans la langue spécialisée, « [une] escalope garnie de jambon et de fromage fondu, puis panée » (*Le Petit Robert de la langue française*, s.v.).

C'est cette entrée sémantique que Bally a adoptée, au début du 20<sup>e</sup> siècle, lorsqu'il fut le premier<sup>7</sup> à théoriser la notion de *phraséologie* en traitant de ce qu'il nommait alors la « délimitation des faits d'expression » et « l'illusion du mot » (1909, § 77-78). Dans ce cadre, Bally distingue trois cas de figure pour classer les faits linguistiques issus de « l'observation des rapports entre la parole et la pensée » (1909, § 78) : premièrement, le mot graphique « recouvre l'unité de pensée [...] comme *table, route, maison* » ; deuxièmement, le mot graphique « renferme deux unités de pensée : c'est le cas [...] illustré par les exemples *relire* et *cuillerée*. » ; troisièmement :

---

<sup>6</sup> La *préconstruction* désigne les restrictions qui pèsent sur la liberté combinatoire des éléments, c'est-à-dire « la possibilité d'utiliser librement les éléments des classes d'objets autorisées par le schéma d'arguments d'un prédicat donné [...]. Cela vaut pour les relations entre un prédicat et ses arguments mais aussi, au niveau inférieur, entre les éléments constitutifs des groupes. » (G. Gross 2012, p. 197).

<sup>7</sup> Même si le terme *phraséologie* est attesté bien avant Charles Bally – la première attestation en français daterait du 18<sup>e</sup> siècle –, le disciple de Ferdinand de Saussure est le premier à proposer une contribution significative à la théorie de la phraséologie. Pour un bref historique de la notion, voir González Rey 2015 (p. 17-28).

L'unité psychologique excède les limites de l'unité graphique et s'étend sur plusieurs mots ; le mot n'est qu'un élément de l'unité réelle, qui se trouve être alors une locution composée ; si, dans un groupe de mots, chaque unité graphique perd une partie de sa signification individuelle ou n'en conserve aucune, si la combinaison de ces éléments se présente seule avec un sens bien net, on peut dire qu'il s'agit d'une locution composée. Un exemple suffira pour caractériser le phénomène : l'expression adverbiale *tout de suite*, qui correspond à l'allemand *gleich*, n'a conservé le sens d'aucun des trois mots qui la composent ; en revanche, elle en a acquis un par elle-même [...]. C'est l'ensemble de ces faits que nous comprenons sous le terme général de phraséologie. » (Bally 1909, § 78)

C'est ainsi le critère sémantique de *compositionnalité* du sens qui est utilisé par Bally dans son approche des « locutions phraséologiques » pour les distinguer des associations libres : lesdites locutions ont un sens qui ne résulte pas de la combinaison des sens des mots les constituant. À l'instar de Legallois & Gréa (2006, p. 12), nous définissons la *compositionnalité* comme « la relation entre le sens d'une expression et le sens de ses parties en vertu de leur combinaison grammaticale » et le terme désigne « la participation des parties constitutives de l'[expression idiomatique] dans le calcul interprétatif, sans pour autant que le tout soit égal à la somme des parties. » (*ibid.*, p. 13). Cette définition est formulée au moyen de l'opérateur logique de l'union linguistique ( $\oplus$ ) par Mel'čuk : « Un signe linguistique complexe AB est dit compositionnel ssi  $AB = A \oplus B$ . » (2013, p. 131). On peut reprendre l'exemple donné par l'auteur pour exemplifier cette définition : « l'énoncé *pomme d'Espagne* est sémantiquement compositionnel : son signifié 'pomme d'Espagne' se compose des signifiés de *pomme* 'pomme', de *de* 'provenant de' et d'*Espagne* 'Espagne' : 'pomme d'Espagne' = 'pomme'  $\oplus$  'de'  $\oplus$  'Espagne'. En revanche, l'énoncé *château en Espagne*  $\approx$  'projet chimérique' n'est pas sémantiquement compositionnel : son signifié 'projet chimérique' ne se compose pas des signifiés de *château* 'château', de *en* 'localisé dans' et de *Espagne* 'Espagne' : 'projet chimérique'  $\neq$  'château'  $\oplus$  'localisé dans'  $\oplus$  'Espagne'. » (Mel'čuk 2013, p. 131). Néanmoins, de nombreuses études ont montré que le critère de compositionnalité demeurerait très délicat à manipuler en raison du peu d'accord observé entre les locuteurs<sup>8</sup>.

Le propos de Sinclair selon lequel la phraséologie a été un champ de recherche longtemps ignoré par les chercheurs occidentaux mérite d'être donc d'être nuancé et ce d'autant plus, si l'on s'intéresse aux travaux des hellénistes du début du 20<sup>e</sup> siècle avec le cas particulier des formules homériques auxquelles j'ai consacré plusieurs travaux.

---

<sup>8</sup> Pour des références détaillées sur ce point, voir Legallois & Tutin 2013 (p. 7).

### 1.1.3 Le cas des formules homériques

La réflexion sur le figement est au cœur des recherches sur les formules homériques. Dès 1923, Meillet est formel :

L'épopée homérique est toute faite de formules que se transmettaient les poètes. Qu'on prenne un morceau quelconque, on reconnaît vite qu'il se compose de vers ou de fragments de vers qui se retrouvent textuellement dans un ou plusieurs passages. (Meillet 1923, p. 61).

En 1928, M. Parry consacra sa thèse pour le doctorat ès Lettres à l'étude de la diction formulaire dans les épopées homériques. Celui-ci voit dans les formules « un fonds commun dans lequel tout poète épique pouvait puiser [...] pour exprimer d'une manière convenable les idées propres à l'épos et pour atténuer les difficultés de la versification. » (1928b, p. 9-10). Il définit la formule comme « une expression qui est régulièrement employée, dans les mêmes conditions métriques, pour exprimer une certaine idée essentielle. » (p. 16). Parry considère ainsi le critère de récurrence comme définitoire dans l'identification de la formule.

Les exemples les plus célèbres de formules sont les syntagmes nominaux associant un nom propre et son épithète (« formule nom-épithète », *ibid.*, p. 20) comme ποδάρκης δῖος Ἀχιλλεύς « le divin Achille aux pieds agiles » (voir par ex. *Iliade* 1.92) ou Ποσειδάων ἐνοσίχθων « Poséidon l'ébranleur de terre » (voir par ex. *Odyssée* 1.44). On peut considérer que dans ces syntagmes, l'« absence de libre actualisation » de chaque élément lexical, pour reprendre les mots de G. Gross, est avérée. En effet, quand Poséidon ou Achille sont désignés par cette formule, le sens de l'épithète n'est pas actualisé : Poséidon n'est pas en train de provoquer un tremblement de terre et Achille n'utilise pas précisément ses pieds au moment où le poète le nomme ainsi. Les formules homériques sont ainsi des unités préconstruites réutilisables tel quel pour exprimer une « idée » (l'idée 'Achille', l'idée 'Poséidon', selon Parry, c'est-à-dire « ce qui en reste après qu'elle a été débarrassée de toute superfluité stylistique », *ibid.*, p. 16). Leur caractère figé provient également de la contrainte métrique exercée par le cadre rigide de l'hexamètre dactylique. En effet, ce critère métrique est un indice précieux pour détecter le caractère figé d'une unité car comme le souligne Habert et *al.* (1997), les corpus dans des états de langue anciens sont typiquement des corpus « sur lesquels l'analyste n'est pas forcément à même de porter des jugements d'acceptabilité » (p. 9). En l'absence de locuteur, le critère métrique est donc fort utile pour dépasser l'impossible jugement d'acceptabilité. On peut ainsi considérer ces études des formules homériques comme une première approche lexicologique du figement et de la phraséologie.

Pour ma part, j'ai mis en œuvre cette approche des formules homérique pour analyser le traitement de la synonymie. L'étude de la combinatoire lexico-syntaxique<sup>9</sup> d'un pivot au sein des formules permet en effet de distinguer les emplois des lexies présentées comme synonymes. Par exemple, l'analyse des épithètes récurrentes – dont la fréquence est supérieure à 10% des emplois dans la formule nom-épithète – est très utile pour contraster les emplois de 4 noms de la mer (ἡ ἄλς, ἡ θάλασσα, τὸ πέλαγος, ὁ πόντος) dans le corpus homérique (Sorba 2010a). Ainsi, la lexie ἄλς présente la particularité d'être le seul nom de la mer à se combiner avec l'adjectif δῖος « à l'éclat divin »<sup>10</sup>. Cette formule nom-épithète apparaît systématiquement sous la forme d'un syntagme prépositionnel à l'accusatif //εἰς ἄλα δῖαν# remplissant les deux derniers pieds de l'hexamètre (/ - ~ ~ / - ~ #)<sup>11</sup>. L'adjectif fonctionne, à chaque fois, comme épithète dite *ornementale ou fixe*, c'est-à-dire « au point de s'être fondu avec [le substantif] dans l'expression d'une idée unique. » (Parry 1928b, p. 193), en l'occurrence celle de 'mer'. Pour leur part, les adjectifs πολύφλοισβος « au grondement puissant » et εὐρύπορος « aux larges passages »<sup>12</sup> se combinent exclusivement avec la lexie θάλασσα (épithètes dites *spéciales*) au sein d'une formule correspondant à un patron métrique bien spécifique : les syntagmes nominaux au génitif //πολυφλοίσβοιο θαλάσσης# et //θαλάσσης εὐρυπόροιο# remplissent systématiquement la fin de l'hexamètre dactylique après la coupe trochaïque (soit après la

---

<sup>9</sup> Dans la lignée de Novakova et Tutin (2009, p. 7), nous considérons que « la combinatoire syntaxique renvoie à la structure actancielle des mots (ou structure valencielle), alors que la combinatoire lexicale intègre les cooccurrences lexicales privilégiées (ou collocations), ainsi que les relations lexicales paradigmatiques ». L'étude de cette combinatoire permet d'élaborer le *profil combinatoire* d'une lexie, c'est-à-dire la « structure schématique de son voisinage syntaxique et sémantique » (Blumenthal 2002b, p. 115-116), et s'inscrit dans une approche lexicologique de la phraséologie à l'interface de la sémantique et de la syntaxe.

<sup>10</sup> Notre traduction tente de concilier les deux notions 'luminosité' et 'divin'. Chantraine (1968) traduit δῖος par « divin » lorsque l'adjectif qualifie des noms de personne et par « brillant » lorsque celui-ci qualifie des éléments, conformément à son origine étymologique (\**dej-* « hell, scheinen », Wodtko et al. 2008, s.v.). Pour une liste des emplois de l'épithète δῖος avec d'autres lexies, voir Parry (1928b, p. 106).

<sup>11</sup> Les 12 occurrences du syntagme prépositionnel εἰς ἄλα δῖαν se trouvent dans *Iliade* 1.141 ; 2.152 ; 14.76 ; 15.161 ; 177 ; 223 ; 21.219 et *Odyssée* 3.153 ; 4.577 ; 5.261 ; 8.34 ; 11.2. Les éditions consultées sont les suivantes : Monro D.B. & Allen T.W. (1920). *Iliadis* (2 vol.). Clarendon Press ; Muehl P. von der (1984). *Homeri Odysseae*. B.G. Teubner.

<sup>12</sup> Nous suivons ici les traductions proposées par Chantraine (1968).

première brève du troisième pied, // ~ / - - / - ~ ~ / - - #)<sup>13</sup>. L'étude des occurrences montre qu'il s'agit, dans ce cas, d'épithètes dites *particularisées*, employées d'après leur signification et non pour leur seul patron métrique : dans le premier cas, la formule désigne la 'mer en mouvement' et, dans le second cas, la 'mer comme voie de navigation'. Ainsi, les formules homériques nom-épithètes peuvent s'analyser sur une échelle de compositionnalité variable.

Par ailleurs, le profil combinatoire des adjectifs dérivés des noms de la mer en grec ancien (ἄλιος, ἐνάλιος, θαλάσσιος, πελάγιος, πόντιος, ὑπερπόντιος), étudié dans un corpus élargi aux tragédies d'Eschyle<sup>14</sup>, apporte également des données pour distinguer les emplois de ces synonymes (Sorba 2015b). D'une part, on constate l'hétérogénéité des syntagmes nominaux comprenant l'adjectif πόντιος dans les tragédies d'Eschyle, leur seul point commun étant le nom non humain ainsi qualifié (τὸ κῦμα et ὁ κλύδων « vague », ὁ μυχός « golfe », ἡ δρόσος « eau », ἡ ἀκτὴ « rivage », ἡ βᾶρις « galiote persane », ἡ πλαγὰ « coup », etc.) ; d'autre part, l'adjectif ἄλιος apparaît majoritairement spécialisé dans l'expression de l'origine-filiation au sein de la formule ὁ γέρων ἄλιος désignant « le Vieux de la mer »<sup>15</sup> et au sein des différentes unités polylexicales désignant ses filles, les Néréides (dat. ἀθανάτησ' ἀλίησι « avec les [déesses] immortelles marines » en *Iliade* 18.86 ; *Odyssée* 24.47, 24.55 ; dat. ἀλίησι κασιγνήτησι « à [ses] sœurs marines » en *Iliade* 18.139 ; gén. ἀλλάων ἀλιάων « parmi les autres [déesses] marines » en *Iliade* 18.432 et nom. ἄλαι θεαί « les déesses marines » en *Iliade* 24.84).

Dans cette section 1.1, mon objectif était de définir succinctement l'approche lexicologique de la phraséologie à l'interface de la sémantique et de la syntaxe mais aussi l'ancienneté de la réflexion sur les contraintes linguistiques qui pèsent sur l'emploi des unités lexicales. J'ai choisi de montrer comment elle était mise en œuvre dans le cadre de mes premiers travaux sur la poésie homérique. L'intérêt de cette approche lexicologique de la phraséologie réside également dans l'influence notable qu'elle a exercée dans mes réflexions sur la modélisation des unités phraséologiques comme nous allons le voir dans la section suivante.

---

<sup>13</sup> Les 8 occurrences de //πολυφλοίσβοιο θαλάσσης# se trouvent dans *Iliade* 1.34 ; 2.209 ; 6.347 ; 9.182 ; 13.798 ; 23.59 et *Odyssée* 13.85 ; 220 ; les 3 occurrences de //θαλάσσης εὐρυπόροιο# se trouvent dans *Iliade* 15.381 et *Odyssée* 4.432 ; 12.2.

<sup>14</sup> Les éditions consultées sont les suivantes : Mazon P. (1931). *Eschyle. Tome I* (2<sup>e</sup> éd.). Les Belles Lettres ; Mazon P. (1925). *Eschyle. Tome II*. Les Belles Lettres.

<sup>15</sup> La formule désigne majoritairement Nérée (*Iliade* 1.538, 1.556, 18.141, 20.107, 24.562 ; *Odyssée* 4.349, 4.365, 4.384, 4.401, 4.542, 13.345, 24.58) mais aussi Protée (*Odyssée* 17.140) et Phorkys (*Odyssée* 13.96).

## 1.2. Les unités phraséologiques : quelle(s) typologie(s) ?

Comme le notent Zaharieva & Kaladieva-Zaharieva (2017), le domaine de la phraséologie se caractérise par une terminologie foisonnante. Le classement des unités phraséologiques s'apparenterait parfois plus à une « véritable quête », pour reprendre le mot de Pecman (2004), qu'à une entreprise scientifique. Celle-ci souligne que c'est précisément parce que les unités phraséologiques constituent un ensemble très complexe et très peu homogène (1.2.1) que leur classification est problématique (1.2.2). Nous nous intéresserons ensuite à l'une d'entre elles en particulier, la collocation, unité dont nous avons décrit le fonctionnement dans plusieurs de nos travaux (1.2.3). Mais avant d'espérer classer les unités phraséologiques, il convient de cerner les contours de celles-ci.

### 1.2.1 Qu'est-ce qu'une unité phraséologique ?

À cette question, Frath & Gledhill (2005) répondent qu'une unité phraséologique se définit comme « une entité référentielle à géométrie variable au sein d'un continuum d'expressions référentielles qui vont de l'unité lexicale à la phrase en passant par l'unité phraséologique, puis au paragraphe et au texte en entier. » (p. 24). L'unité phraséologique est donc ici une expression parmi d'autres (unité lexicale, phrase, paragraphe, texte) avec lesquelles elle partage la propriété de référer. Toutefois, ces expressions se distinguent les unes des autres par la place qu'elles occupent au sein d'un « continuum », terme à comprendre aussi bien dans son sens physique (« ensemble d'éléments homogènes ») que dans son sens didactique (« objet ou phénomène progressif dont on ne peut considérer une partie que par abstraction », *Le Petit Robert s.v.*). Le continuum implique ainsi une conception scalaire du phénomène considéré.

Dans son approche Sens-Texte, Mel'čuk propose une approche binaire pour distinguer une unité phraséologique – ou « phrasème » dans sa terminologie – du reste des autres unités :

un phrasème est un syntagme NON LIBRE, c'est-à-dire, de façon informelle, un syntagme qui ne peut pas être construit selon les règles générales de la langue. (Mel'čuk 2011, p. 42).

Avec sa définition du phrasème comme un « énoncé multilexémique non libre [soit] une configuration de deux ou plus lexèmes syntaxiquement liés », Mel'čuk (2013) formule l'idée d'une contrainte opérant sur l'axe paradigmatique : le locuteur ne sélectionne pas le sens ni chacune des composantes lexicales de l'unité strictement pour ses propriétés linguistiques, c'est-à-dire indépendamment des autres composantes. Dans son approche, Mel'čuk propose d'utiliser les termes *contrainte* et *contraint* dans la définition du concept de phrasème et de

garder les termes *figé* et *figement* pour caractériser chaque phrasème par le degré et la nature de son figement.

L'une des définitions les plus complètes de l'unité phraséologique a été proposée par Gries (2008, p. 4-6) et s'articule en 6 dimensions ou paramètres définitoires :

	Paramètre définitoire	Attendus
(1)	the nature of the elements involved in a phraseologism.	the co-occurrence of a form or a lemma of a lexical item and any other kind of linguistic element.
(2)	the number of elements involved in a phraseologism.	phraseologisms can contain more than two elements.
(3)	the number of times an expression must be observed before it counts as a phraseologism.	an expression [is considered as] a phraseologism if its observed frequency of occurrence is larger than its expected one.
(4)	the permissible distance between the elements involved in a phraseologism.	the more widespread broader perspective which also recognizes discontinuous phraseologisms.
(5)	the degree of lexical and syntactic flexibility of the elements involved.	studies (also) including completely lexically unspecified and thus maximally flexible expressions, such as the English ditransitive pattern [V OBJ1 OBJ2] (...). My definition of phraseologisms excludes only the last of these because they do not involve at least one lexically specified element (as required by the first parameter).
(6)	the role that semantic unity and semantic non-compositionality / non-predictability play in the definition.	I will require semantic unity, but not non-compositional semantics

Tableau 1 : Les six paramètres définitoires des phraséologismes selon Gries (2008).

In sum, a phraseologism is defined as the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance. (Gries 2008, p. 6)

La définition de Gries a le mérite d'éviter l'écueil du tout phraséologique dans la langue en englobant dans son périmètre les expressions complètement figées en synchronie (angl. *of course, at least*) et les associations entre unités lexicales et structures grammaticales

particulières (ou *collostructions* selon Stefanowitsch & Gries 2003, par ex. angl. *The [X think nothing of V<sub>gerund</sub>] construction*) mais en excluant les expressions récurrentes qui ne fonctionnent pas comme unité de sens (angl. *of the, in the*) et les patterns sans aucune réalisation lexicale<sup>16</sup>. L'avantage de cette définition est, finalement, de ne pas spécifier le niveau de granularité auquel les observations se font et de laisser cette liberté au chercheur.

Nous proposons d'ajouter un dernier élément dans la définition de l'unité phraséologique (ou phraséologisme) pour y inclure le lien que cette unité entretient avec l'environnement textuel dans lequel elle s'insère. Cette conception, qui s'appuie sur les travaux précurseurs de Blumenthal (voir par exemple 2002a), incite à envisager « comment l'unité polylexicale prend racine (dans le sens presque littéral du terme) dans le discours, c'est-à-dire [à] chercher les liens subtils sur les plans sémantiques, syntaxiques et pragmatiques que ce genre de séquences tisse avec son voisinage. » (Soutet, Sfar & Mejri 2018, p. 8). Le phraséologisme peut ainsi être considéré comme « une forme que le discours pourrait générer ou épouser pour se configurer [... et qui possède la faculté d']essaimer du sens dans le discours. » (p. 7-9). Toutefois, cet ajout déborde du cadre strictement lexicologique de la définition de l'unité phraséologique. J'en examinerai les implications dans le chapitre 2 en présentant nos travaux qui s'inscrivent précisément dans le dépassement de ce cadre. Aussi, avant d'envisager ce tournant phraséologique, nous présentons ici quelques réflexions concernant la manière de classer les unités phraséologiques ainsi que les implications des choix opérés.

### 1.2.2 Comment classer les unités phraséologiques ?

À la suite de Pecman (2004), nous pouvons affirmer que le problème de classification des unités phraséologiques se trouve au cœur des études en phraséologie. Notre objectif est bien de montrer que les multiples classifications en phraséologie sont complémentaires et que « la typologie n'est pas uniquement un moyen d'observation des faits de langue mais aussi matière à réflexion en soi » (*ibid.*, p. 144). Cette proposition rend compte du modèle dominant qu'elle

---

<sup>16</sup> Voir Gries (2008) : « the definition of phraseologism proposed above serves as a convenient cover term for co-occurrence phenomena at the *syntax-lexis interface* (since at least one lexical element must be specified) as opposed to the *syntax-semantics interface*, to which lexically unspecified patterns from Pattern Grammar or argument structure constructions from Construction Grammar would be associated. On the other hand, the present definition does cover particular words' significant attraction to argument structure constructions (cf. Stefanowitsch & Gries's (2003) *collostructions*) and completely lexically filled and frozen expressions which, although diachronically derived from collocations, are synchronically single lexemes (e.g. *of course, at least*). » (p. 8).

a dégagé de l'examen de différentes typologies et qui se fonde sur le principe du continuum, en l'élargissant pour y inclure des classes manquantes :

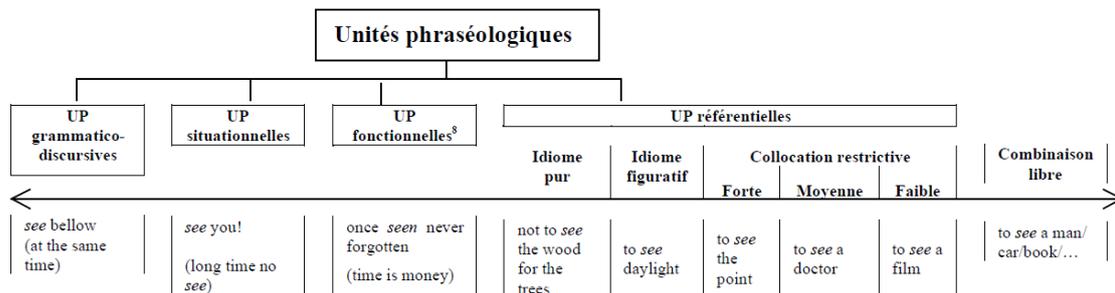


Figure 1 : Typologie des unités phraséologiques selon Pecman (2004, p. 143).

Sa typologie est fondée sur une organisation en trois niveaux. Au premier niveau, sont placées, de part et d'autre d'un continuum, l'*unité phraséologique* (UP) et la *combinaison libre*. Cette dernière ne présente aucune contrainte combinatoire hormis la compatibilité sémantique entre les unités mises en relation. C'est ce que Pecman nomme le « modèle dominant » formulé, entre autres, par Mel'čuk (1993) pour le français ou Howarth (1996) pour l'anglais.

Au deuxième niveau, Pecman propose quatre catégories d'unités phraséologiques : les *UP grammatico-discursives*, les *UP situationnelles*, les *UP fonctionnelles* et les *UP référentielles*, en s'inspirant du classement de Burger (1998) qui distinguait *referentiell – structurell – kommunikativ Phraseologismen*. Chacune de ces quatre catégories englobe des UP particulières :

(i) les *UP grammatico-discursives* regroupent les mots outils (conjonctions, locutions prépositionnelles, etc.) ;

(ii) les *UP situationnelles* (ou les *énoncés liés* de Fonagy 1982 ou les *pragmatèmes* selon la terminologie de Mel'čuk 1993, entre autres) regroupent des « séquences qui sont gelées par rapport à un contexte situationnel (par exemple, selon Mel'čuk, *c'est pour toi* est la seule variante possible à utiliser dans une situation où on demande quelqu'un au téléphone) » (Pecman 2004, p. 134) ;

(iii) les *UP fonctionnelles* regroupent les séquences comme les proverbes, les maximes, les slogans, les citations, etc. qui « sont souvent décrits comme des éléments de la langue ayant une fonction spécifique qui les distingue de tous les autres éléments lexicaux. » (Pecman 2004, p. 143).

Le troisième niveau concerne exclusivement la catégorie des *UP référentielles* organisée en trois sous-ensembles (*ibid.*, p. 142) :

(i) Les *idiomes* sont « des expressions complètement figées » ;

(ii) Les *idiomes figuratifs* connaissent un double emploi : « un emploi libre avec un sens compositionnel et un emploi figé où le sens de l'ensemble est figuratif. Ainsi, *c'est dans la poche* peut être employé au sens propre (l'expression a alors un sens compositionnel et signifie que "qch est (effectivement) dans la poche") et au sens figuré (elle signifie alors "c'est réussi"). » ;

(iii) Les *collocations restrictives* « concernent les relations préférentielles qui s'établissement entre les unités lexicales. ». Cet ensemble d'UP est lui-même subdivisé en 3 groupes selon la force de l'attraction entre les unités lexicales (*collocation restrictive forte, moyenne, faible*). Cette modélisation est inspirée des travaux de Benson, Benson & Ilson (1997).

Cette vision des UP en parallèle des constructions libres a l'avantage de permettre « la prise en compte des unités comportant des paradigmes ouverts qui donnent parfois lieu à de véritables schémas collocationnels et qui sont souvent délaissés par ce type d'examen car considérées comme étant en marge de l'univers phraséologique. » (Pecman 2004, p. 144). Le modèle de Pecman propose ainsi un élargissement des modèles précédents en s'en inspirant afin de rendre compte des données linguistiques. Il propose de distinguer les différentes UP par des critères largement fonctionnels ce qui permet de résoudre, au moins en partie, le problème de la perception empirique du figement<sup>17</sup>.

Même si elle s'inscrit dans le « modèle dominant » décrit dans les paragraphes précédents, la typologie développée par Mel'čuk dans une perspective lexicographique ne propose pas exactement le même découpage des UP (ou *phrasèmes* dans ce cadre théorique) comme on peut le voir dans la figure 2 :

---

<sup>17</sup> Ce « modèle dominant » est aussi celui mis en œuvre dans la tripartition des UP proposée par Frath & Gledhill (2005) qui postulent une conception scalaire du figement mais utilise surtout le critère de la référence pour distinguer : (1) les **UP lexicales** très figées (ex. *kick the bucket*) réfèrent d'un seul tenant de la même manière que des dénominations monolexicales (« Ce qui compte, ce n'est donc pas l'aspect mono- ou polylexical d'une séquence de mots, c'est sa capacité à référer ensemble ou séparément » *ibid.*, p. 18) ; (2) les **UP semi-figées** (ex. *messe noire, colère noire, spill the beans*) qui acceptent un paradigme commutatif restreint construit à partir de l'élément séminal ; (3) les **UP ouvertes** (ex. *cheveux noirs, strong tea*) qui acceptent un paradigme commutatif plus large et dont les éléments sont liés sémantiquement à l'élément séminal (ou élément tête de la structure soit *tea* ou *cheveux* dans les exemples précédents). Le principal obstacle à la mise en œuvre de cette typologie réside dans la difficulté à tracer des frontières nettes entre seulement 3 catégories d'UP (UP lexicales et semi-figées et UP semi-figées et ouvertes) sur les seuls critères de référence et d'échelle de figement.

## 8 Typologie générale des phrasèmes et leur place dans un dictionnaire de la langue

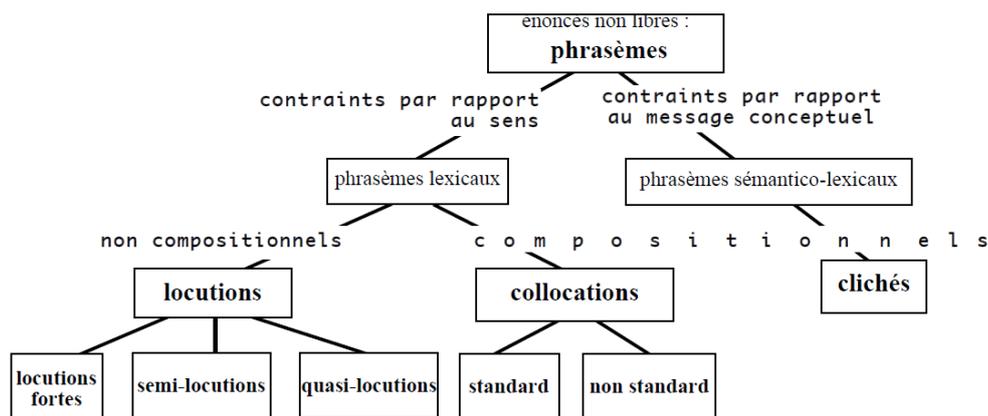


Figure 2 : Typologie des phrasèmes selon Mel'čuk (2013, p. 145).

En effet, le classement des phrasèmes mis en œuvre ici est hiérarchisé en 3 niveaux. À l'intérieur de chacun de ceux-ci, les ensembles s'opposent les uns aux autres selon plusieurs types de contraintes.

Au premier niveau, sont regroupés *phrasèmes lexicaux* et *phrasèmes sémantico-lexicaux* dont la distinction se fait ici selon l'axe paradigmatique, c'est-à-dire selon les contraintes de sélection de leurs composants. Dans le cas du *phrasème lexical*, « le sens 'σ' est construit par le Locuteur librement pour n'importe quelle situation désignée, mais le choix des lexèmes pour exprimer 'σ' est contraint. » (ex. *porter son attention sur, être au bout du rouleau*) (*ibid.*, p. 131). Dans le cas du *phrasème sémantico-lexical*, « le sens 'σ' n'est pas construit par le Locuteur mais sélectionné comme un tout de façon contrainte – en fonction du contenu conceptuel à verbaliser, donc en fonction de la situation désignée ; le choix des lexèmes pour exprimer 'σ' est, le plus souvent, également contraint (ex. *Vous dites ?, Défense de stationner*). » (*ibid.*, p. 131).

Au deuxième niveau, Mel'čuk distingue les *phrasèmes sémantiquement compositionnels* des *phrasèmes sémantiquement non compositionnels* selon l'axe syntagmatique, c'est-à-dire selon les contraintes de combinaison de leurs composantes ou leur compositionnalité. En combinant ces deux types de contrainte, il obtient ainsi 3 classes majeures de phrasèmes<sup>18</sup> :

<sup>18</sup> La quatrième classe qui serait celle des *\*phrasèmes sémantico-lexicaux non compositionnels* ne peut pas exister car « un énoncé non libre quant à la construction de son sens et, en même temps, non compositionnel est impossible pour la raison suivante : s'il est non compositionnel, il a, par définition, un sens "uni", qui lui est associé comme à un tout ; ce sens n'est donc pas construit par le Locuteur pour l'occasion (mais est sélectionné comme un tout) et, par conséquent, on ne peut pas parler du caractère contraint de sa construction. » (Mel'čuk 2013, note 4).

(i) les *locutions* (ou *phrasèmes lexicaux non compositionnels*, en anglais *idiom*, ex. *être au bout du rouleau* ‘ne plus avoir de ressources vitales’);

(ii) les *collocations* (ou *phrasèmes lexicaux compositionnels*, ex. *porter son attention sur*);

(iii) les *clichés* (ou *phrasèmes sémantico-lexicaux compositionnels*, ex. *Quel âge avez-vous ?*).

Dans la modélisation de Mel’čuk, le troisième niveau hiérarchique concerne exclusivement les *phrasèmes lexicaux* (*locutions* et *collocations*) qui sont répartis en plusieurs sous-ensembles s’opposant les uns aux autres. Tout d’abord, les *locutions* peuvent être *locutions fortes*, *semi-locutions* et *quasi-locutions* selon leur degré d’opacité, c’est-à-dire en fonction de l’inclusion du sens des composantes A et B dans le sens de la locution AB. Ainsi, une *locution forte* ou *complète* n’inclut dans son sens aucun des sens de ses composantes A et B (ex. *au bout du rouleau* ‘sans ressources vitales’). Pour distinguer *semi-locution* et *quasi-locution* ou *locution faible*, Mel’čuk modélise un découpage du sens linguistiques des composantes de ces locutions en séparant un sens linguistique ‘ $\sigma$ ’ en deux parties : ‘ $\sigma$ ’ = ‘ $\sigma_1$ ’  $\oplus$  ‘ $\sigma_2$ ’. Il appelle *pivot sémantique* du sens ‘ $\sigma$ ’ sa partie ‘ $\sigma_1$ ’ si et seulement si on a ‘ $\sigma_2(\sigma_1)$ ’. Autrement dit, le *pivot sémantique* est « la partie du sens ‘ $\sigma$ ’ qui est l’argument de l’autre partie de ce sens, qui est un prédicat. Par exemple, dans le sens de VENDRE, ‘X cède Y à Z en échange d’une somme d’argent W’, le pivot sémantique est ‘céder’ : le sens prédicatif ‘en échange d’une somme d’argent – le reste du sens de ‘vendre’ – porte sur ‘céder’ comme un prédicat sur son argument. » (*ibid.*, p. 134). Dans ce cadre, une *semi-locution* (ex. *loup de mer* ‘homme dont la profession est de naviguer sur mer et qui est très expérimenté’) inclut dans son sens le sens d’une de ses composantes A (‘mer’) mais pas en tant que pivot sémantique et n’inclut pas le sens de l’autre composante B (‘loup’), tout en incluant encore un sens additionnel C qui est son pivot sémantique (‘homme’). Une *quasi-locution* ou *locution faible* (ex. *donner le sein* ‘nourrir un bébé en mettant le sein dans la bouche’) inclut dans son sens le sens de toutes ses composantes A et B (‘sein, ‘donner’  $\approx$  ‘mettre à la bouche’) mais pas en tant que pivot sémantique, en incluant un sens additionnel C qui constitue le pivot sémantique (‘nourrir’). Mel’čuk reconnaît néanmoins que les locutions faibles sont parfois difficilement distinguables des collocations.

En tant que *phrasème lexical compositionnel*, la *collocation* est décrite comme la réunion de deux composantes : l’une (*la base*) est sélectionnée librement, juste pour son sens tandis que l’autre (*le collocatif*) doit être choisie en fonction du sens à exprimer et de la première composante. Mel’čuk répartit celles-ci en deux catégories selon que le lien sémantique qui unit la base et le collocatif peut être systématisé (*collocation standard*) ou non (*collocation non*

*standard*). On peut rendre compte du sens des *collocations standards* par une modélisation du lien sémantique unissant la base et le collocatif : par exemple, l'intensivité est marquée par des collocatifs de différentes natures grammaticales (la base étant indiquée entre parenthèses) : (trempé) *comme une soupe*, raide (bourré), (respirer) à *pleins poumons*, (rôle) *crucial*, (rire) à *s'en décrocher la mâchoire*, etc. Ce caractère systématique du lien sémantique unissant base et collocatif est modélisé au moyen des fonctions lexicales dans la théorie Sens-Texte développée par l'auteur. Dans cet exemple, c'est l'opérateur Magn qui est mobilisé : Magn(*trempé*) = *comme une soupe*. À l'inverse, le lien unissant base et collocatif ne peut pas être systématisé dans une *collocation non standard* : « ce lien s'applique à très peu de bases, très souvent à une seule base ; et il n'implique que très peu de collocatif, le plus souvent un seul » (*ibid.*, p. 137). Par exemple, le lien sémantique qui unit la base *année* à son collocatif *bissextile* 'qui comporte 366 jours' est unique.

La typologie proposée par Mel'čuk dans le cadre de son approche lexicographique a été fondatrice. En effet, elle a été à l'origine de nombreuses recherches qui ont progressivement permis de faire sortir les UP de ce cadre initial. C'est ce que je propose d'examiner désormais en interrogeant l'UP singulière qu'est la collocation.

### 1.2.3 Le cas de la collocation : perspectives synchroniques et diachroniques

La *collocation* est une UP à laquelle de nombreuses recherches ont été consacrées dans les différentes traditions<sup>19</sup>. Comme l'a montré Tutin (2010), la modélisation de la collocation a remis en question deux dichotomies classiques en linguistique. Tout d'abord, la collocation présentée comme le produit de la performance et de l'usage, tout en étant préconstruite dans la mémoire sémantique des locuteurs, rend l'opposition performance / compétence caduque. Ensuite, définie comme une association lexicale privilégiée mise en œuvre dans une relation syntaxique, elle bouscule le clivage entre lexique et syntaxe (*ibid.*, p. 13-14). C'est pourquoi sa saisie s'opère à l'interface de la sémantique et de la syntaxe.

La notion de *collocation* et de « *meaning by collocation* » est centrale dans les travaux de Firth dès les années 1950 :

« [...] élaborée dans un texte de 1951, elle est conçue dans l'approche polysystémique de Firth comme sens lexical, l'un des cinq niveaux où se détermine le sens – les autres étant les niveaux phonétique, morphologique, syntaxique et sémantique. [...] Le sens lexical réside dans l'usage

---

<sup>19</sup> Le terme apparaît pour la première fois sous la plume de Palmer en 1933 dans une étude en didactique du lexique. Pour un historique de la question, voir Tutin 2010.

des mots en contexte, et non dans une sémantique du mot *a priori*, conceptualiste, logique ou psychologique. [... Ce sont] les collocations d'un mot pivot regroupées en configurations qui permettent d'établir le sens de ce mot. » (Léon 2008, p. 15-16).

Toutefois, cette notion de collocation évolue dans l'œuvre de Firth. Après avoir désigné la cooccurrence de deux éléments linguistiques quelconques (morphème, mot, syntagme ou expression) dans toute étendue de texte, elle se focalise ensuite sur « l'attente mutuelle » qu'un mot fait porter sur un autre (« *mutual expectancy* » ou « *collocability* ») :

Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of *night* is its collocability with *dark*, and of *dark*, of course, collocation with *night*. (Firth 1957 [1951], p. 196).

Comme le rappelle Tutin (2010), Firth utilise surtout la notion de collocation dans une perspective textuelle qui s'émancipe de la tradition lexicographique. C'est le sens tel qu'il est actualisé par son contexte qui intéresse Firth et c'est également dans cette perspective que j'ai inscrit mes travaux depuis mon doctorat<sup>20</sup>.

Dans le cadre du projet EMOLEX<sup>21</sup>, j'ai étudié les collocations construites autour des lexies d'affect en considérant la collocation, à l'instar de Tutin (2013), comme un « sous-ensemble productif d'expressions lexicalisées binaires, organisées autour d'une structure prédicat-argument. » (*ibid.*, p. 61). La lexie d'affect<sup>22</sup> (verbe, nom, adjectif) constitue alors le prédicat autour duquel s'organise la collocation. Afin de mener à bien ces études, dans nos travaux collectifs, nous avons systématiquement analysé la combinatoire lexico-syntaxique des lexies d'affect, et tout particulièrement leurs structures actanciennes et leurs associations lexicales privilégiées, en combinant ainsi les deux volets indissociables que sont la syntaxe et la sémantique (Novakova & Sorba 2013a ; 2013b ; 2014a ; 2014b ; 2014c ; 2018 ; Sorba & Goossens 2016 ; Sorba & Novakova 2018). Ces éléments sont devenus les constituants centraux

---

<sup>20</sup> En effet, cette conception du sens de Firth se rapproche de celle de la sémantique interprétative développée par Rastier (1987) que j'ai mise en œuvre dans mon doctorat (Sorba 2010a) et dans cinq publications en m'intéressant aux associations lexicales récurrentes construites autour des noms de la 'mer' en indo-aryen ancien (Sorba 2008-2009 ; 2010-2011), grec ancien (Sorba 2008b ; 2015a) et latin (Sorba 2010b).

<sup>21</sup> Pour la présentation du corpus et des outils développés dans le projet EMOLEX, voir Diwersy *et al.* 2014.

<sup>22</sup> Le terme *affect* a été choisi de manière conventionnelle pour regrouper ce que certaines classifications distinguent parfois difficilement (nom/verbes/adjectifs d'émotions, de sentiment, d'état, etc. ; voir Grossmann & Tutin 2005).

de ce que nous avons nommé *profil discursif*<sup>23</sup>. Nos études s'inscrivent dans une conception de la collocation à la fois quantitative et qualitative puisqu'elles combinent des éléments statistiques avec une analyse fine des contraintes qui pèsent sur les lexies (sur ce point, voir Hausmann & Blumenthal 2006, p. 3).

L'intérêt de cette méthode apparaît tout d'abord dans le fait qu'elle relie une acception à une configuration lexico-syntaxique. Celle-ci est donc particulièrement opératoire pour observer le traitement de la synonymie à laquelle je m'intéresse depuis mon doctorat. Par exemple, notre étude des lexies françaises synonymes *rage* et *fureur* dans le corpus journalistique contemporain du projet EMOLEX (Sorba & Goossens 2016) a pu montrer une nette préférence de la lexie d'affect *fureur* à se combiner avec des collocatifs verbaux exprimant la causativité (par ex., *susciter / provoquer la fureur*), tandis que son synonyme *rage* se combine plutôt avec des collocatifs verbaux exprimant la manifestation physique subie (par ex. *trembler de rage*). Nous avons pu remarquer également une différence de nature grammaticale dans les associations privilégiées par *fureur* qui se combine préférentiellement avec des collocatifs adjectivaux (*noire, soudaine, sauvage*) alors que *rage* tend à préférer les collocatifs verbaux (*pleurer, trembler, libérer, ravalé*, etc.).

Un deuxième résultat mis en évidence par nos études est la possibilité de contraster non pas des lexies une à une mais des classes entières de lexies (Novakova & Sorba 2013a ; 2013b ; 2014a ; 2014b ; 2014c ; 2018 ; Sorba & Novakova 2018). En effet, nous avons pu observer des différences dans la combinatoire lexico-syntaxique des lexies d'affects interpersonnels avec celles d'affects causés<sup>24</sup>. Sur le plan syntaxique, les premières (affects interpersonnels) réalisent plus volontiers leur structure actancielle complète (ex. 1) alors que les secondes (affects causés) préfèrent un schéma actantiel plus dépouillé (ex. 2) :

---

<sup>23</sup> Le *profil discursif* d'une lexie décrit son fonctionnement linguistique en discours. Il réunit des phénomènes linguistiques relevant aussi bien du niveau local (paliers en-deçà de la phrase) que du niveau global (paliers au-delà de la phrase). Par l'observation des propriétés formelles des unités lexicales, il permet de « construire leur identité sémantique sans recourir au monde conceptuel » pour reprendre les propos de Novakova & Tutin (2009, p. 9) à propos de l'approche distributionnelle harrissienne. Ce que nous entendons par *profil discursif* est très proche de la définition du *profil lexico-discursif* proposée par Veniard (2013, p. 55) : « Le profil rassemble les caractéristiques préférentielles de la combinatoire et du fonctionnement discursif de ce mot sur les plans sémantique, syntaxique, syntagmatique (cooccurrence, collocations, cotexte), textuel, énonciatif et interdiscursif ».

<sup>24</sup> Les classifications traditionnelles distinguent entre *affect interpersonnel* dont la source est interne (ou « endogène », voir Anscombe 1995 ; par ex., *amour, dédain, jalousie*) et *affect réactif* dont la cause est externe (ou « exogène » *ibid.* ; par ex., *jubilation, stupeur, déception, surprise*).

- 1) Vous (X) êtes souvent **jaloux** de vos collègues (Y) et vous avez du mal à être sincèrement heureux pour eux *lorsqu'ils réussissent* (Z). (*Ouest France* 14/06/2007)
- 2) *L'attaque* (Z) **stupéfi**a. (*Le Figaro* 06/01/2007)

Dans l'exemple 1, la réalisation intégrale des actants autour de l'adjectif *jaloux* (le jaloux X, le jaloué Y, la cause Z = la réussite) confère à l'énoncé toute sa complétude en discours, ce qui s'oppose à l'emploi monovalent de *stupeur* dans l'exemple 2 dans lequel la cause (Z) est seule exprimée. Sur le plan sémantique, les lexies d'affects causés se combinent plus que les lexies d'affects interpersonnels avec des collocatifs intensifs (par ex. *jubilation irrésistible, puissante, féroce* ; *déception immense, profonde, énorme, terrible*).

L'application de cette méthode a également produit un troisième résultat pour mettre en évidence la différence de comportement combinatoire de collocations construites autour d'un même verbe. Par exemple, dans le cas du verbe *prendre*, l'étude des variations syntagmatiques permet de distinguer un profil combinatoire distinct (Cavalla & Sorba 2017) : la collocation *prendre la fuite* connaît peu de variations syntagmatiques alors que *prendre un bain* et *prendre des risques* en connaissent davantage (ex. *prendre un bon bain, ~ un bain de soleil, ~ le bain, etc.* ; *prendre un/des risques, ~ le plus grand risque, ~ à court terme le risque de, etc.*). Ces trois collocations connaissent donc un degré de figement différent. De ces quelques résultats, il apparaît clairement que la collocation est un fait linguistique dont la description permet de répondre au principe saussurien de la langue comme système au sein duquel les éléments s'opposent.

Toutefois, cette définition de la collocation ne serait pas complète si nous ne l'envisagions pas aussi comme un phénomène dynamique. En effet, comme le souligne Dubreil (2008), la collocation est soumise, en synchronie, à des contraintes sociolinguistiques, par exemple de registres de langue (familier *péter un cable* ; soutenu *hanter sa conscience*) ou de domaines de spécialité (*une dose d'attaque*, en pharmacologie). En diachronie, on observe une tendance des collocations à devenir sémantiquement opaque y compris pour le locuteur natif (*fort comme un Turc*<sup>25</sup>). C'est dans une perspective diachronique que j'ai mené quatre études sur les collocations afin d'apporter des éléments inédits à l'analyse du changement linguistique.

Ainsi, mon étude des collocations autour de la lexie *feu* dans un corpus issu de Frantext (Sorba 2018a) a montré comment la collocation *donner + feu* s'était maintenue au fil des siècles

---

<sup>25</sup> Un locuteur contemporain pourra trouver opaque cette référence à la force des Turcs s'il ne l'inscrit pas dans le contexte historique du combat de la chrétienté contre les Ottomans et de la peur que ces derniers inspiraient (sur ce point, voir Sonntag 2018).

en changeant de sens (« faire chauffer » → « tirer au moyen d'une arme à feu » → « fournir un moyen pour allumer du tabac » → « donner l'autorisation ») alors que les emplois d'autres collocations plus ou moins synonymes s'étiolaient (*apporter + feu ; fournir + feu*). L'explication de ce maintien pourrait résider dans la plasticité sémantique de la base *donner* qui a su épouser les changements de sens de son collocatif *feu* au fil du temps.

Par ailleurs, dans nos deux études diachroniques des collocations construites autour des verbes *prendre* (Cavalla & Sorba 2017) et *donner* (Cavalla & Sorba 2018), nous avons analysé le lien entre le noyau sémique des verbes<sup>26</sup> et celui de leurs collocatifs dans un corpus issu de Frantext (voir sitographie). Nous avons observé que *prendre un bain* (attesté à partir du 17<sup>e</sup> s.) avait progressivement éliminé ses concurrents *avoir bain* (attesté dès le 12<sup>e</sup> s.) et *se plonger aux bains* (attesté dès le 16<sup>e</sup> s.). L'hypothèse que nous avons avancée repose sur l'ensemble des traits sémantiques (TS) inhérents au verbe *prendre* (/mouvement/, /possession/, /centripète/, /inchoatif-punctuel/) qui lui ont permis finalement de s'imposer car ceux-ci sont « en cohérence avec les TS des bases qui auraient sélectionné le “bon” collocatif » (Cavalla & Sorba 2017, p. 199-200). De même pour les collocations construites autour du verbe *donner* (*donner + gage ; donner + réponse*), nous avons constaté que l'affinité sémantique entre base et collocatif, reposant ici sur le partage des TS communs /transmission/ et /interaction/ (Cavalla & Sorba 2018), expliquait très certainement le maintien de ces collocations au détriment d'autres concurrents (*offrir/fournir + gage ; fournir/livrer + réponse*).

Enfin, c'est également une analyse diachronique qui nous a permis d'appréhender le phénomène de renouveau d'une collocation avec le cas de la collocation *sentir une émotion* (attestée au 16<sup>e</sup> siècle) qui change progressivement son verbe support pour devenir *éprouver une émotion* dès le 18<sup>e</sup> siècle (Novakova, Sorba & Tutin 2018). L'étude de l'évolution de la combinatoire nous a également permis de relier le développement d'une complémentation syntaxique avec la bascule progressive du nom *sentiment* vers le statut de classifieur qu'il occupe désormais (*un sentiment de* + SN, *un sentiment que* + SV). Ces quatre exemples illustrent le fait que la collocation doit bien être saisie comme un phénomène dynamique non seulement en synchronie mais aussi en diachronie.

---

<sup>26</sup> Dans ces deux articles, nous considérons que les deux verbes *prendre* et *donner* ne sont pas des verbes supports complètement « vides de sens » (Gross 1998, p. 25) quand ils forment des collocations. Ils conservent en partie leurs traits sémantiques inhérents.

Au terme de ce premier chapitre, nous avons montré que le terme *phraséologie* nomme un ensemble de faits lexicaux à l'interface de la sémantique et de la syntaxe tout en désignant un principe constitutif de la langue. Cette conception s'appuie sur une approche lexicologique de la phraséologie qui a profondément influencé la modélisation de ses unités. Les différentes typologies présentées sont construites autour du phénomène scalaire qu'est le figement. Mes recherches n'ont pas eu pour objectif de discuter les typologies existantes que je pense complémentaires mais plutôt d'affiner la description d'une catégorie d'UP particulière, la collocation, à laquelle j'ai consacré une partie de mes travaux

Pour caractériser ces unités particulières, j'ai montré tout l'intérêt d'analyser la combinatoire lexico-syntaxique qui apporte des éléments pour la description de leur fonctionnement linguistique aussi bien en synchronie qu'en diachronie. Les apports principaux en synchronie développés dans ce chapitre concernent le traitement opératoire de la synonymie dans différentes langues et dans différents corpus. En diachronie, nous avons posé quelques jalons pour décrire le changement linguistique affectant les collocations et expliquer le succès de certaines et le déclin d'autres. Ces premières études en diachronie mériteront d'être complétées à l'avenir.

Toutefois, l'approche strictement lexicologique de la phraséologie est désormais considérée comme incomplète. Pour Legallois & François (2006), l'approche à l'interface de la syntaxe et de la sémantique ne suffit plus :

[S]ouligner les rapports entre syntaxe et sémantique est devenu une sorte de *topos* ; pourtant, les analyses qui réclament une appréhension d'ensemble de la forme et du contenu laissent toujours persister une indépendance entre les deux niveaux : elles ne font que mieux identifier l'interface. (Legallois & François 2006, p. 1)

Cette impasse est sévèrement soulignée par Frath & Gledhill (2005, p. 14) : « Faute d'un point de vue correct, la linguistique des unités phraséologiques est condamnée à errer entre les affirmations de la lexico-grammaire et les empiètements de la syntaxe. ». Le défi à relever est donc désormais celui d'une approche holistique des phénomènes linguistiques dans la construction du sens<sup>27</sup>. C'est pour cela qu'un changement d'échelle est nécessaire.

---

<sup>27</sup> Pour Legallois & François (2006, p. 1), ce sont les grammaires de construction qui sont à même de relever ce défi car celles-ci « conçoivent que les cadres syntaxiques sont dotés d'une signification [...] et] plaident pour un traitement phraséologique des formes. ». Dans l'approche d'obédience cognitive de Fillmore et al. (1988), la notion de *construction* est le seul moyen de rendre compte des connaissances phraséologiques du locuteur. Même si nous la mentionnons ici pour mémoire, celle-ci ne constitue pas le cadre théorique dans lequel j'ancre mes travaux.

## **Chapitre 2. Au-delà du syntagme : « Vers une extension du domaine de la phraséologie »**

Le changement d'échelle consiste à opérer un « saut qualitatif » pour reprendre les mots de Nølke (2002) qualifiant ainsi le processus de passage des énoncés au texte<sup>28</sup>. Dans ce deuxième chapitre, nous montrerons comment a été théorisé ce tournant phraséologique qui fait déborder la phraséologie du cadre lexicologique du syntagme (section 2.1). J'envisagerai ensuite les implications concrètes de ce que Blumenthal nomme la « portée textuelle » des lexies et ce que nos modélisations sur les lexies d'affect ont apporté à ce cadre (section 2.2).

### **2.1. Le tournant phraséologique**

Legallois & Gréa identifient le « tournant phraséologique » dès les années 1990. La phraséologie est devenue « la matière première de certaines conceptions linguistiques : elle propose un renouvellement de l'analyse syntaxique, sémantique, textuelle et psycholinguistique sur la base d'observations plus empiriques, grâce, notamment, au recours aux corpus de grandes dimensions. » (Legallois & Gréa 2006, p. 5). Comme nous allons le voir, le recours à de vastes corpus renouvelle non seulement l'approche de la phraséologie mais permet également « une extension du domaine de la phraséologie » pour reprendre le titre du volume coordonné par Legallois & Tutin (2013) :

La phraséologie intègre désormais des objets d'étude très variés, allant des collocations aux séquences discursives en passant par la parémiologie, ou encore, les schémas syntaxiques. Les approches proposées s'étendent maintenant au-delà des disciplines traditionnelles de la lexicologie, de la syntaxe et de la sémantique, et abordent largement la linguistique du discours, la psycholinguistique ou la linguistique informatique. Les objets de la phraséologie, autrefois perçus comme des anomalies ou des exceptions, deviennent maintenant des éléments centraux dans les modèles linguistiques où la notion de principe phraséologique du langage se développe.  
(p. 3)

Cette extension du domaine de la phraséologie est double car elle concerne à la fois ses objets d'étude et les disciplines impliquées pour en rendre compte. Dans cette section, nous présenterons les réflexions théoriques qui ont servi de cadre à nos recherches (2.1.1) et tout

---

<sup>28</sup> À la suite de Halliday & Hasan (1976, p. 273), je considère le *texte* comme une unité sémantique en contexte (« *unity of meaning in context* »). Le texte est le lieu de réalisation du *discours* qui englobe « l'intradiscours, ou le déroulement linéaire du discours, et l'interdiscours, le déjà-dit des autres discours inlassablement repris. » (Veniard 2013, p. 33). Sur ces questions, voir le chapitre 3.

particulièrement celles qui établissent un lien entre le palier local des unités phraséologiques et le palier global du texte (2.1.2).

### 2.1.1 La théorie du *lexical priming*

Selon Legallois & Gréa (2006), ce tournant phraséologique est à mettre en rapport avec la recherche contextualiste britannique et, tout particulièrement, avec les travaux de Hoey (2005) qui développe une vision radicale qu'il nomme *lexical priming* :

Every word is primed for use in discourse as a result of the cumulative effects of an individual's encounters with the word. If one of the effects of the initial priming is that regular word sequences are constructed, these are also in turn primed. More specifically:

- 1) Every word is primed to occur with particular other words; these are its collocates.
- 2) Every word is primed to occur with particular semantic sets; these are its semantic associations.
- 3) Every word is primed to occur in association with particular pragmatic functions; these are its pragmatic associations.
- 4) Every word is primed to occur in (or avoid) certain grammatical positions, and to occur (or avoid) certain grammatical functions; these are its colligations.
- 5) Co-hyponyms and synonyms differ with respect to their collocates<sup>29</sup>, semantic associations and colligations.
- 6) When a word is polysemous, the collocates, semantic associations and colligations of one sense of the word differ from those of its other senses.
- 7) Every word is primed for use in one or more grammatical roles; these are its grammatical categories.
- 8) Every word is primed to participate in, or avoid, particular types of cohesive relation in a discourse; these are its textual collocates.
- 9) Every word is primed to occur in particular semantic relations in the discourse: these are its textual semantic associations.
- 10) Every word is primed to occur in, or avoid, certain positions within the discourse; these are its textual colligations.

Very importantly, all these claims are in the first place constrained by domain and/or genre. (Hoey 2005, p. 13)

Dans cette liste, Hoey présente l'amorçage d'un mot (*lexical priming*), à savoir sa préférence à apparaître dans certaines configurations (ou à les éviter), comme un phénomène multifactoriel.

---

<sup>29</sup> Dans ce cadre, le terme *collocation* est à comprendre au sens de Firth (1951) : « the company words keep ».

Pour l'expliquer, sont convoquées les contraintes d'ordre lexical, sémantique, syntaxique, pragmatique, textuel, discursif. À chacune de ces contraintes correspondent des unités phraséologiques particulières (*collocate*, *semantic association*, *pragmatic association*, *colligation*, etc.). Les items 1 à 7 énumérés par Hoey modélisent les rapports de la lexie avec son environnement lexical et syntaxique. Il s'agit toujours d'une approche lexicologique de la phraséologie à l'interface de la syntaxe et de la sémantique. Nous avons d'ailleurs pu montrer, dans le premier chapitre, la validité de ces propositions lors de nos études et tout particulièrement celle de l'item 5 sur le traitement de la synonymie.

Dans ce cadre théorique, le traitement de la polysémie repose sur l'examen des différences de réalisations phraséologiques des acceptions d'une même lexie (item 6). Par exemple, en indo-aryen ancien, l'élaboration du profil combinatoire de la lexie *samudrá-* « flot, mer, océan » permet de dégager quelques faits saillants pour distinguer ses différentes acceptions car dans les hymnes religieux de la *Ṛksamhitā*, *samudrá-* désigne aussi l'oblation de soma<sup>30</sup> ou le récipient qui la contient lors du rituel (Sorba 2014a). En effet, cette acception est actualisée préférentiellement quand la lexie entre dans la structure actancielle d'un verbe de mouvement ou quand elle se trouve à proximité de termes techniques (*ávya-* « filtre », *kalása-* « cuve », *sávana-* « pressurage », etc.) où elle se trouve ainsi prise au sein d'une isotopie<sup>31</sup>. En l'occurrence, ce sont les éléments lexico-syntaxiques, l'isotopie mais aussi le genre textuel (hymne) qui permettent ici l'actualisation du sens technique de *samudrá-*.

Ce qui est inédit dans la modélisation de Hoey, c'est l'affirmation des affinités entre lexie et niveau textuel (items 8 à 10). Selon Legallois & Gréa (2006, p. 8), l'objectif des contextualistes

---

<sup>30</sup> Le terme *soma* désigne une plante non identifiée « dont le jus sert de liqueur oblatrice » (Renou 1956, p. 8-9). Les hymnes regroupés dans le recueil de la *Ṛksamhitā* sont adressés aux divinités du panthéon brahmanique et récités lors de la célébration du rituel.

<sup>31</sup> À la suite de Rastier (1987, p. 91-92), nous définissons l'*isotopie* comme « toute itération d'une unité linguistique » quel que soit le niveau où ces unités se situent. Le terme *itération* indique que l'isotopie est « le résultat d'un processus (conscient ou non, là n'est pas la question) d'encodage et de décodage. » (p. 93). Ce concept d'isotopie a été créé par Greimas en 1966 pour répondre à la nécessité de montrer « comment les textes entiers se trouvent situés à des niveaux sémantiques homogènes, comment le signifié global d'un ensemble signifiant [...] peut être interprété comme une réalité structurelle de la manifestation linguistique. » (Greimas 1986 [1966], p. 53). À l'instar de Rastier (2006), nous considérons que les isotopies constituent un « fond sémantique » sur lesquels se détachent des « formes textuelles ».

est désormais d'affirmer que « la colligation<sup>32</sup> est la règle ». Avec une remise en cause des théories syntaxiques compositionnelles, la colligation prend chez Hoey une portée non seulement propositionnelle mais aussi textuelle : « colligation may be interpreted as going beyond traditional grammatical relations and embracing such phenomena as the positioning of a word or word sequence within the sentence or paragraph and even its positioning within the text as a whole. » (Hoey 2005, p. 43). Hoey opère ainsi une extension du domaine d'observation de la phraséologie en formulant l'existence d'une affinité sélective entre les lexies et le palier supérieur du texte.

### 2.1.2 La contrainte dépasse le cadre du syntagme

Cette extension du domaine de la phraséologie opère donc quand la phraséologie est directement reliée au palier du texte. De manière similaire, Rastier (2006) estime que :

Le sens suppose une contextualisation maximale, aussi bien par l'étendue linguistique – le contexte, c'est tout le texte – que par la situation, définie par une histoire et une culture, bien au-delà du *hic et nunc* seul considéré par la pragmatique. (p. 99-100)

Rastier affirme l'interrelation entre les niveaux local et global dans la détermination du sens. La construction du sens d'une unité phraséologique ne peut donc trouver sa complétude qu'à l'échelle du contexte maximal à savoir le texte.

Pour sa part, Blumenthal (2002a) considère la cohérence textuelle comme « une forme de prévisibilité qui permet de formuler des hypothèses sur le déroulement du texte à partir de certains substantifs assumant des fonctions sémantiquement importantes. » (Blumenthal 2002a, p. 29). En analysant la combinatoire lexico-syntaxique en italien des substantifs désignant les sept péchés capitaux (*superbia, avarizia, lussuria, ira, gola, invidia, accidia-pigrizia*) dans des textes du 14<sup>e</sup> siècle, son objectif est de montrer que la contrainte dépasse le cotexte immédiat du mot, ce qui permettrait de « jeter un pont entre la sémantique du mot et la linguistique du texte, en passant par la syntaxe, en particulier par celle des prépositions. » (*ibid.*, p. 30). En l'occurrence, les unités phraséologiques sont les associations privilégiées de ces noms avec d'autres mots (prépositions, verbes, substantifs) : par ex. les associations privilégiées de *ira* avec les prépositions *con, in, et a* indiquent que le péché *ira* constitue « non pas un mobile,

---

<sup>32</sup> Hoey (2005), définit la *colligation* selon trois critères : « 1 the grammatical company a word or word sequence keeps (or avoids keeping) either within its own group or at a higher rank; 2 the grammatical functions preferred or avoided by the group in which the word or word sequence participates; 3 the place in a sequence that a word or word sequence prefers (or avoids). » (p. 43)

mais une manière d’agir [...et] renvoie à une cause événementielle dans le texte antérieur. » (*ibid.*, p. 35). Le mot et ses associations font partie ou participent alors d’un « scénario »<sup>33</sup> qui se réalise dans un texte en amont ou en aval de leur apparition : en somme, l’unité phraséologique *con / in / a ira* conditionne l’expression d’une « cause événementielle » en amont de son apparition dans le texte.

## 2.2 La « portée textuelle » d’une unité lexicale

Blumenthal (2014) propose ultérieurement d’introduire le concept de « portée textuelle » d’un mot pour décrire « l’impact qu’a concrètement, dans un texte donné, ou que peut avoir habituellement l’acception d’un mot (= « lexie structurante ») sur l’organisation de son voisinage textuel. » (p. 179). À titre d’exemple, il montre comment la lexie *dépit* est susceptible de « structurer sémantiquement » plusieurs propositions autour d’elle. En effet, ce ne sont pas les mêmes éléments du scénario de *dépit* (déception, chagrin, colère) qui sont impliqués en fonction de la combinatoire de la lexie (*dépit amoureux, de dépit, par dépit, en dépit de*) : par exemple, alors que *par dépit* réalise la totalité du scénario, *dépit amoureux* ne réalise que le début du scénario (déception, chagrin). J’envisagerai dans cette section ce concept de « portée textuelle » dans deux cas de figure : la capacité d’une lexie à actualiser une acception<sup>34</sup> d’une autre lexie (2.2.1) et la capacité à structurer son environnement textuel (2.2.2).

### 2.2.1 Actualiser une acception

La « portée textuelle » d’une lexie peut se mesurer à sa capacité à actualiser une acception dans un environnement spécifique. Par exemple, en latin, dans le cas de la lexie *altum*<sup>35</sup>, une dénomination occasionnelle de la mer (Sorba 2010b), c’est la présence contextuelle du nom du

---

<sup>33</sup> Le terme *scénario* est emprunté à Hausmann dans une étude où il expose les trois types de traitement à appliquer lors du travail lexicographique concernant l’exemple : « La collocation est le contexte codé minimal. L’isotopie est la répétition des traits sémantiques à d’autres endroits du texte. Le scénario est le cadre sémantique qui lie le mot à une tranche de vie. La loi de la collocation montre le mot dans une unité semi-phraséologique, réutilisable dans les textes (domaine de l’idiomatique). La loi de l’isotopie insiste sur la structure textuelle de la langue (domaine de la cohésion). La loi du scénario insiste sur les situations dans lesquelles se présente la vie humaine (domaine de la cohérence). » (2005, p. 283).

<sup>34</sup> Sur l’actualisation, voir *supra* chapitre 1 section 1.1.

<sup>35</sup> Ce neutre substantivé issu de l’adjectif *altus* s’est spécialisé dans le sens « haut » ou « profond ». Il ne revêt qu’occasionnellement le sens de « mer ».

« navire » *nauis* qui permet l’actualisation de l’acception « mer », comme on peut le voir dans ces deux exemples issus de la pièce *Rudens* de Plaute<sup>36</sup> :

3) *illorum nauis longe in altum apscesserat* (v.66)

« Leur navire est déjà loin, en pleine mer »

4) *nunc eam cum nauis scilicet abiisse pessus in altum* (v.395)

« Et maintenant ce [sac de cuir], évidemment, est allé avec le bateau au fond de la mer. »

En indo-aryen ancien, on peut observer le même phénomène d’actualisation de l’acception « mer, océan » dans le substantif *arṇavá-* « flot », une dénomination occasionnelle de la mer (Sorba 2008-2009). Dans la strophe suivante issue de la *Rksamhitā*<sup>37</sup>, c’est sa combinaison avec l’adjectif *samudrīya-*, dérivé de *samudrá-* « océan », et la présence du nom de la « rivière » (*nadí-*) qui actualise cette acception « mer, océan » pour *arṇavá-* :

5) *só arṇavó ná nádīyaḥ samudrīyaḥ práti gr̥bhṇāti vísritā várīmabhiḥ*. (*RS* 1.55.2a-b)

« Comme le flot océanique reçoit les rivières, celui-ci (= Indra) [reçoit] les [hommages] situés de toutes parts, par ses vastes dimensions. »

La structure comparative, introduite par *ná*, établit une analogie entre le dieu Indra (*só*) et l’océan (*arṇavó samudrīyaḥ*) : tous deux sont désignés comme des réceptacles recevant (*práti-GRAH-* « recevoir ») pour l’un, des hommages, pour l’autre, des rivières (*nádīyaḥ*). Cette comparaison entre le dieu Indra et l’océan, récurrente dans la *Rksamhitā*, constitue un scénario poétique bien identifié pour le poète : la comparaison implique la mention (explicite ou implicite) des éléments que sont le réceptacle (le dieu, l’océan) et ce qui s’y jette (les hommages, les rivières).

Un phénomène similaire se rencontre dans une strophe de l’*Atharvaveda* recension *śaunaka*<sup>38</sup> pour la lexie *salilá-*, une autre dénomination occasionnelle de la mer (Sorba 2010-2011) :

6) *vidhúm dadrāṇám salilásya pṛṣṭhé yívānaṃ sántam palitó jagāra /*

*devásya paśya kāvyaṃ mahitvādyā mamāra sá hyáḥ sám āna //* (*AVŚ* 9.10.9-10)

« Le vieillard grisonnant a englouti la [lune] solitaire, elle qui est jeune et qui court **sur le dos de la mer** ; regarde la sagesse inspirée du dieu dans toute sa grandeur : alors qu’aujourd’hui il est mort, hier celui-ci respirait encore ! »

<sup>36</sup> L’édition consultée est celle de Lindsay W.M. (1905). *T. Macci Plauti Comoediae tomus II*. Clarendon Press.

<sup>37</sup> L’édition consultée est celle de Nooten B. A. van & Holland G.B. (1994). *Rig Veda: A Metrically Restored Text with an Introduction and Notes*. Harvard University Press.

<sup>38</sup> L’édition consultée pour le livre 9 est celle de Devi Chand M.A. (1997). *The Atharvaveda. Sanskrit text with English translation with introductory remarks by M.C. Joshi*. Munshiram Manoharlal Publishers Pvt Ltd.

Cette strophe constitue la reprise presque à l'identique d'une strophe de la *Ṛksaṃhitā* (10.55.5), où le syntagme *salilāsya pr̥ṣṭhé* « sur le dos de la mer » se substitue à l'original *sámane bahūnām* (« dans une assemblée nombreuse »). Ces deux séquences au locatif indiquent le lieu sur ou dans lequel s'effectue le procès dénoté par le participe parfait moyen *dadrāṇām* (*DRĀ-* « courir ») dont l'agent est la lune, désignée par le terme énigmatique, et hapax, *vidhú-*<sup>39</sup>. La mention du nom du « dos » (*pr̥ṣṭhá-*) permet l'établissement d'un rapport d'analogie entre l'ellipse de la course lunaire et la voûte céleste, par le biais de leur propriété commune, la convexité. Ainsi, *salilá-* désigne « l'océan céleste », dans lequel se déroule immuablement la course des astres (*ibid.*, p.181). Cet emploi nouveau pour *salilá-* est néanmoins attesté dès la *ṚS* pour *samudrá-* lorsque celui-ci désigne l'« océan céleste » en tant que réservoir des eaux (voir par ex. *ṚS* 4.15.11a-b ; Sorba 2010-2011). Dans ce cas, c'est la conjonction de deux facteurs, un facteur local, la portée textuelle de la lexie *pr̥ṣṭhá-*, et un facteur global, le jeu d'intertextualité entre les recueils d'hymnes *Ṛksaṃhitā* et *Atharvaveda*, qui permet l'actualisation de l'acception « océan céleste » pour *salilá-*.

Dans ces quelques exemples, issus de mes premiers travaux portant sur des textes poétiques anciens, j'ai procédé à une relecture inédite des résultats à la lumière de ce concept de « portée textuelle » pour en tester la validité.

## 2.2.2 Structurer l'environnement textuel

La « portée textuelle » d'une unité lexicale s'analyse également dans sa capacité à structurer son environnement textuel. Les mots lexicaux peuvent jouer un rôle dans la structuration textuelle comme le rappelle Stubbs (2014) dans son étude sur les *Dubliners* de James Joyce :

[...] words in text occur in bursts and semantic clusters, and this uneven distribution of individual content words and words from particular semantic fields signals textual structure. (p. 245)

Celui-ci montre à titre d'exemple, comment, dans la nouvelle intitulée « Eveline », les mots nouveaux qui arrivent en « rafales » (*bursts*) à un moment donné du texte (ici à la fin de la

---

<sup>39</sup> Nous nous appuyons sur Whitney et Griffith pour la traduction du syntagme *vidhúm yúvānaṃ* « the young Moon » ; la lune est qualifiée de « jeune » (*yúvāna-*), dans la mesure où son cycle mensuel est perçu comme une unité temporelle, comprenant un début et une fin, contrairement à la permanence apparente de la course du soleil. Voir Whitney W. D. (2001 [1905]). *Atharva-Veda-Saṃhitā: Translated with a Critical and Exegetical Commentary*, revised and brought nearer to completion and edited by C. R. Lanman. Harvard Oriental Series vol. 7 et 8 / Motilal Banarsidass ; Griffith R.T.H. (1986 [1893]). *The Hymns of the Sāmaveda*. Munshiram Manoharlal (sa traduction de SVS n°325 reprend *ṚS* 10.55.5).

nouvelle) permettent d'introduire un nouveau sujet dans la structure narrative. La progression narrative est ainsi détectée grâce à la fréquence de la première apparition des mots.

De plus, l'apparition en « rafales » permet de structurer une séquence narrative en fournissant ce que Plantin nomme un « script émotionnel » (2011, p. 23). C'est le cas des lexies d'affect comme *jalousie* et *jaloux* qui apparaissent fréquemment, dans le corpus journalistique, en série, dans un dispositif énumératif déroulant les ingrédients du « script émotionnel » : l'attachement initial (*amour, amitié, passion*), la rivalité (*haine, soupçon, convoitise*), la souffrance (*frustration, rancœur, amertume*), la trahison (*trahison, mensonge*), la violence (*agressivité, dispute, meurtre*), comme on peut le voir dans les deux exemples ci-dessous :

- 7) Elle ne supporte pas sa **jalousie**, sa **violence** quand il a bu, ses **menaces**. (*Ouest-France* 28/10/2008)
- 8) La Loi du désir le voit se transformer en amant **possessif, jaloux** et **meurtrier**. (*Le Monde* 04/11/2007)

Tous les ingrédients de ce script émotionnel n'apparaissent pas simultanément mais la présence de ce script est bien conditionnée par un paramètre externe, à savoir les « attentes culturellement formatées des lecteurs » (Novakova & Sorba 2014b p. 640 ; 2014c, p. 167).

Dans la lignée de Blumenthal (2014), nous avons montré que l'environnement textuel des lexies d'affect semble être conditionné par leur sémantisme (Sorba & Novakova 2018). Par exemple, *jubilation* active une isotopie de polarité positive (ex. 9) alors que *dédain* active une isotopie de polarité négative (ex. 10). Les lexies formant chaque isotopie sont soulignées dans les exemples :

- 9) Le programme était alléchant et la cour d'assises de Paris avait pris, lundi 26 novembre, son air des grands jours avec une foule savourant d'avance la **jubilation** de pouvoir dire « J'y étais », tandis que quelques personnalités chanceuses munies de laisser-passer se faufilaient aux premières loges. (*Le Monde* 28/11/2007)
- 10) Et si le touriste venu profiter du soleil est plaintif, pleurnichard quand il chausse ses bottes en caoutchouc et regarde le ciel avec désolation, il est considéré avec **dédain**. (*Ouest-France* 24/07/2007)

L'isotopie de la polarité (positive ou négative), alliée très souvent à celle de l'intensité<sup>40</sup>, constitue ainsi « un fond sémantique » (Rastier, 2006, p. 104), caractéristique de

---

<sup>40</sup> À l'instar de Plantin (2011), nous considérons que la polarité et l'intensité sont « deux composantes essentielles de la séquence émotionnelle. » (p. 123). Nous avons pu effectivement montrer que l'isotopie de l'intensité était présente de manière significative dans l'environnement textuel des lexies d'affect (Novakova & Sorba 2018).

l'environnement textuel des lexies d'affect dans le corpus journalistique (Novakova & Sorba 2014a ; 2014c ; Sorba & Novakova 2018).

Les travaux que j'ai présentés ici ont permis de valider le caractère opératoire du concept de « portée textuelle » tel que posé par Blumenthal. Dans la section suivante (2.3), je détaillerai les apports du projet EMOLEX et de mes travaux à cette modélisation du lien entre unités phraséologiques et niveau textuel.

### **2.3 Modélisation du lien entre UP et texte : le cas des lexies d'affect**

La modélisation du lien entre unités phraséologiques (UP) et niveau discursif a été au cœur des travaux du projet EMOLEX *Le lexique des émotions dans 5 langues européennes*. Pour reprendre les mots de Plantin, il s'agit de montrer que « l'émotion, l'expression de l'engagement personnel dans le discours ne sont pas des phénomènes discursifs limités, locaux strictement assignables à un mot ou à un énoncé ; elles se diffusent sur tout un discours. » (2011, p. 75). Les études ont été menées sur de gros corpus journalistiques collectés pour cette occasion. Le corpus français qui a servi pour mes différents travaux est constitué de deux années de parution (2007-2008) des quotidiens *Le Monde*, *Le Figaro*, *Libération* et *Ouest-France* (soit 120 millions de mots). Dans cette section, j'examinerai les principaux résultats du projet modélisés dans une partie de mes travaux<sup>41</sup>, à savoir l'impact des lexies d'affect sur l'organisation discursive (2.3.1), et notamment, leur rôle dans le cas particulier de la séquence argumentative (2.3.2).

#### **2.3.1 L'impact des lexies d'affect sur l'organisation discursive**

En se fondant sur les postulats de Van Valin & LaPolla (1997) distinguant valence syntaxique et valence sémantique<sup>42</sup>, Novakova, Grossmann & Goossens (2013) ont montré, dans une étude

---

<sup>41</sup> Nous avons travaillé sur les lexies d'affect des champs de la surprise (Novakova & Sorba 2013a et b ; 2014a et c : *stupeur/stupéfier/stupéfait*), de la colère (Sorba & Goossens 2016 : *rage/rageur, fureur/furieux*), de la joie (Sorba & Novakova 2018 : *jubilation*), du mépris (Sorba & Novakova 2018 : *dédain*), du respect (Novakova & Sorba 2014a : *estime*), de la jalousie (Novakova & Sorba 2013a et b ; 2014a et c), de la tristesse (Novakova & Sorba 2014a : *déception* ; Novakova & Sorba 2014b : *nostalgique*). La modélisation issue de ces études de cas a été publiée dans Novakova & Sorba 2018.

<sup>42</sup> « The syntactic valence of a verb is the number of overt morphosyntactically coded arguments it takes. One can talk about the semantic valence of the verb as well, where the valence here refers to the number of semantic arguments that a particular verb can take. These two notions need not coincide. » (Van Valin & LaPolla 1997, p. 147).

s'appuyant sur un corpus littéraire issu de Frantext, que la saturation ou non de la valence des lexies des champs de *surprise* et de *respect* avait un impact sur l'organisation discursive de l'énoncé, sur sa façon de structurer ou de hiérarchiser l'information. L'hypothèse qui fonde ce travail est que le locuteur sélectionne une structure actancielle en fonction de ses visées discursives. Ainsi, par exemple, dans les emplois monovalents (un seul actant exprimé), le « centrage » discursif<sup>43</sup> se fait sur l'expérienceur (ex. 11) ou la cause de l'émotion (ex. 12) qui sont mis en avant :

11) Ni femme, ni bête, elle **étonne**. (C. Paysan, *Les feux de la Chandeleur*, 1966)

12) Après ce repas animé, le silence, l'immobilité **surprenaient**. (P. Sabatier, *Trois sucettes à la menthe*, 1972)

Nous avons constaté les mêmes résultats dans le corpus journalistique EMOLEX pour les lexies *stupeur* et *jalousie* (Novakova & Sorba 2014c). Nous avons même pu observer des emplois avalents (pas d'actant exprimé) qui ne semblaient pas attestés dans le corpus littéraire :

13) Nouveau déboire pour la compagnie aérienne australienne. Vendredi, l'un de ses Boeing 747-700 atterrit sans encombre à Singapour. Les techniciens procèdent à des contrôles de routine. **Stupeur** ! L'appareil a perdu, en vol, un panneau d'accès au moteur de 30 cm sur 30 cm. (*Ouest-France* 16/08/2008)

14) Allers-retours entre Rome et Paris. Passion. Et puis **jalousie**. Tromperies. Scènes à gifles. Retrouvailles. (*Libération* 08/08/2007)

Dans ce cas, le centrage discursif se fait sur l'affect lui-même qui est exprimé « à l'état pur » (Novakova & Sorba 2014c, p. 164-165). La combinatoire syntaxique de la lexie d'affect a bien un impact sur le déroulement du discours.

En outre, dans une autre étude contrastive portant sur les noms de surprise, Kern & Grutschus (2014) précisent que la lexie *surprise* est dotée d'une « fonction discursive, voire argumentative », contrairement à *étonnement*, qui permet au locuteur de « donner plus de poids à des points importants de son argumentation » (p. 192). De même, elles révèlent que la lexie allemande *Überraschung* « surprise » témoigne d'un « réel potentiel de structuration discursive » (p. 196) lorsqu'elle est employée en combinaison avec les adjectifs *eigentlich* « à

---

<sup>43</sup> En reprenant le concept de *mise en relief* théorisé par Weinrich (1973) et le terme *centrage* avancé par Blumenthal (2006), Fesenmeier montre le caractère opératoire de cette notion pour distinguer les quasi-synonymes français *se souvenir* / *se rappeler* et italiens *ricordare* / *ricordarsi (di) qc.* Le type de centrage discursif permet ainsi de « guide[r] l'attention ou les attentes du lecteur » (Fesenmeier 2010, p. 93) car il a « pour fonction de donner du relief à un texte en projetant au premier plan certains contenus et en repoussant d'autres dans l'ombre de l'arrière-plan. » (Weinrich 1973, p. 107).

proprement parler » et *wirklich* « réel », deux collocations hautement spécifiques de ce pivot : par exemple, le syntagme *eigentliche Überraschung* forme l'apogée d'un « crescendo informationnel » (p. 196), au sein duquel plusieurs informations s'enchaînent pour déboucher sur une information principale. La configuration lexico-syntaxique correspond ainsi à une visée discursive.

Comme nous allons le voir, un lien étroit a été également établi dans les travaux présentés dans la sous-section suivante (2.3.2), toujours dans le cadre du projet EMOLEX, entre le profil discursif des lexies d'émotion et un type particulier de séquence discursive<sup>44</sup>, la séquence argumentative.

### 2.3.2 Le rôle des lexies d'affect dans l'argumentation

En partant des trois idées que les émotions sont argumentables (Plantin 1997), que « la capacité à exprimer des émotions est une condition d'exercice de l'argumentation » (Plantin 2011, p. 113) et que l'argumentation est inséparable de la manière spécifique d'agencer les énoncés (Micheli 2010 ; Jacquin & Micheli 2012), nous avons exploré le corpus journalistique du projet EMOLEX pour élaborer le profil discursif de plusieurs lexies d'affect représentatives des différentes classes<sup>45</sup>. Ce corpus de presse d'opinion est particulièrement bien adapté à l'étude de l'argumentation de l'émotion et par l'émotion<sup>46</sup> dans la mesure où « la problématique des émotions est essentielle pour la rhétorique argumentative, préoccupée d'efficacité sociale de la parole et, par conséquent, tenue par le souci d'adéquation descriptive. » (Plantin 2011, p. 75).

En effet, l'expression des émotions participe de l'exercice de l'argumentation au sens où les lexies d'affect peuvent être étayées par un argument ou servir elles-mêmes d'arguments (Sorba

---

<sup>44</sup> À la suite de Adam (2011), nous postulons l'existence de « types relativement stables d'énoncés de base [...] disponibles pour d'innombrables combinaisons et transformations. » (p. 28). Ces « formes premières » ne sont pas génériques et « correspondent à des formes élémentaires de narration, de description, d'argumentation, d'explication et de dialogue. » (*ibid.*). Nous les nommons donc *séquences discursives*.

<sup>45</sup> Dans leur typologie des noms d'affect sur la base de leur combinatoire lexico-syntaxique, Tutin *et al.* (2006) distingue 6 classes : les noms d'affect interpersonnels (ex. *amitié, haine*), interpersonnels causés (ex. *respect, admiration, mépris*), ponctuels réactifs (ex. *surprise, désespoir, peur*), interpersonnels réactif (ex. *colère, honte, dégoût*), duratifs non contrôlés (ex. *ennui, bonheur, tristesse*) et duratifs contrôlés (ex. *peine, joie*).

<sup>46</sup> Je reprends cette terminologie à Plantin (1997, p. 82) qui distingue l'argumentation (fondée) « sur » l'émotion (« arguer d'une émotion pour justifier une action », l'argumentation par l'émotion) et l'argumentation « de » l'émotion (« justifier une émotion par l'existence d'un état de choses »).

& Novakova 2018) : celles-ci ne sont pas seulement « des *objets d'argumentation* (au sens où l'on peut les justifier ou les critiquer en formulant des raisons en leur (dé)faveur) » mais aussi des « *ressources pour l'argumentation* (au sens où l'on peut y faire appel pour accroître l'efficacité d'un raisonnement) » (Micheli, Hekmat & Rabatel 2013, p. 11).

Pour procéder à l'examen de la construction argumentative d'une émotion, il convient d'étudier « d'une part les énoncés qui attribuent une émotion à un sujet – il peut s'agir du locuteur, de l'allocataire ou d'un tiers – et, d'autre part, les énoncés ou, plus souvent, les enchaînements d'énoncés qui semblent venir à l'appui des premiers en leur donnant une justification possible. » (Micheli 2010, p. 106-107). L'élaboration du profil discursif des lexies permet ainsi de procéder à cet examen sur la base des phénomènes linguistiques observés en discours. Pour illustrer la première condition avancée par Micheli, nous prenons l'exemple des lexies d'affect des champs de la stupeur et de la jalousie. Les lexies *jalousie/jalouser* (affect interpersonnel, duratif) et *stupeur/stupéfier* (affect causé, réactif, ponctuel) apparaissent quasi exclusivement dans le discours rapporté (Novakova & Sorba 2013a ; 2013b). Dans le corpus journalistique, l'énonciateur premier (le journaliste) rapporte les propos d'énonciateurs seconds comprenant des lexies d'affect (qu'il s'agisse d'une figure majeure de l'art du 20<sup>e</sup> siècle, André Breton, à propos de la peintre hongroise Judit Reigl ou, plus anonymement, de l'avocat de la défense), ce dispositif participant à la co-construction du point de vue :

15) Reigl est plus sage, en ce qu'elle se contente de peindre, mais les personnages de Volupté incomparable n'ont pas besoin des artifices du réel pour être inquiétants : « Vous êtes en possession de moyens qui me **stupéfient** de la part d'une femme, lui écrit Breton, tout machisme bu, et je vous crois en mesure d'accomplir des choses immenses. » (*Le Monde* 30/06/2007)

16) Est évoquée « la **jalousie** malade », dixit la défense, de la concubine violentée. (*Ouest-France* 24/09/2008)

Le journaliste mentionne ainsi l'émotion sous une forme distanciée en déléguant la parole à un énonciateur second jugé digne de confiance ; ce dispositif polyphonique permet de présenter habilement le *devoir éprouver*<sup>47</sup>.

Concernant la seconde condition exprimée par Micheli à propos des enchaînements d'énoncés, nous avons testé, dans la lignée de Blumenthal (2002a) et de Hoey (2005),

---

<sup>47</sup> De manière traditionnelle, l'objectif du discours argumentatif est de fonder un *devoir croire* (défendre une thèse et y faire adhérer l'auditoire) ou un *devoir faire* (induire un passage à l'acte chez l'auditoire). Affirmer l'argumentabilité des émotions, c'est affirmer « la capacité de l'argumentation à fonder un *devoir éprouver* » (Micheli 2010, p. 110).

l'hypothèse des colligations textuelles dans ce même corpus : les lexies d'affect privilégient-elles ou évitent-elles certaines positions au sein d'unités discursives repérables comme la phrase ou le paragraphe ? Nous avons pu constater, par exemple, que le verbe *stupéfier* et le nom *stupeur* apparaissent très souvent à un moment clé de l'article : en position centrale pour *stupéfier* (Novakova & Sorba 2013a ; ex. 16) et en position initiale pour *stupeur* (Novakova & Sorba 2018 ; ex. 17 et 19) :

- 17) **À ma grande stupeur**, les gens étaient assez indifférents, rêvassant, bâillant aux corneilles, comme si la vie continuait. (*Le Monde*, 09/01/2007)
- 18) **(E1)** L'équipée de la première Occidentale kamikaze en Irak a été hier retracée par le tribunal de Bruxelles. Les cinq frères djihadistes de Muriel Degauque, connue comme la première femme occidentale à avoir épousé le destin de kamikaze en Irak, ont été condamnés, hier à Bruxelles, [...], pour avoir collaboré à la constitution d'un réseau terroriste. Le tout, a dénoncé le président du tribunal correctionnel, s'est opéré dans « un climat d'intolérance, d'exaltation, et de fanatisme religieux ». **(E2)** Le sort de cette femme belge de 38 ans **avait stupéfié** la Belgique. **(E3)** Née en 1967 dans une famille ouvrière de Charleroi (Wallonie), vivant de petits boulots, notamment dans une boulangerie, Muriel Degauque, avait basculé dans le fanatisme islamique à l'issue d'un séjour au Maroc effectué en compagnie de son petit ami marocain et mentor, Issam Goris. Elle portait alors gants et burka. En novembre 2005, elle se décide enfin à « faire ses noces ». (*Le Figaro* 11/01/2008)
- 19) Scandale autour de la nouvelle Miss Belgique. **Stupeur** et tremblement dans le camp flamand : **[Ca]** la Miss Belgique, fraîchement élue ce week-end, ne maîtrise pas le néerlandais, langue maternelle de 60 % des Belges. La réalité a éclaté en pleine cérémonie quand la demoiselle a avoué ne pas comprendre une question posée en néerlandais. **[Co]** Conséquence : huées des 4 000 spectateurs, puis gros titres de la presse flamande. À peine sacrée et déjà « massacrée », la jeune étudiante wallonne a promis de perfectionner son néerlandais. (*Ouest-France* 18/12/2007)

Ces deux extraits (18) et (19) peuvent être définies comme des « séquences émotionnelles » dont Tersis & Plantin (2017) énumèrent les composantes : « un expérient (ou expérienteur) ; un événement disruptif interrompant l'action en cours de l'expérient et ouvrant la séquence ; une émotion, déclarée ou inférable ; une description de l'état et du comportement de l'expérient sous le coup de l'émotion ; des stratégies de gestion de l'émotion. » (p. 108). En effet, ici le journaliste déroule son texte selon un ordre logique ou chronologique. En (18), il commence par exposer les circonstances à l'origine de l'émotion (épisode 1, E1) avant de déclarer l'émotion elle-même par le verbe *stupéfier* à un moment pivot (ou acmé), soit pour introduire

une transition, soit pour marquer une étape dans une gradation d'intensité (épisode 2, E2). Enfin, le cotexte droit large de la lexie est occupé par une séquence détaillant la cause de l'émotion ou confirmant celle-ci (épisode 3, E3). L'enchaînement des trois épisodes constitue le dispositif discursif (ou scénario) dans lequel l'expression de l'émotion (E2) apparaît entre une séquence d'amorçage (E1) et une séquence qui réalise le travail textuel d'étayage (S3) : le verbe *stupéfier* est ici suivi d'une séquence justifiant l'émotion ressentie par la communauté qui légitime la stupeur par le décalage incompréhensible entre le destin que laissait supposer le statut de femme ordinaire effectivement décrit et le destin sanglant avéré (Novakova & Sorba 2013a).

Dans l'exemple (19), le journaliste opte pour une présentation logique de type cause-conséquence (c'est le scénario le plus fréquent) : l'énoncé de la cause [Ca] puis des conséquences [Co] de l'affect suit l'apparition de la lexie d'affect elle-même. Le journaliste effectue un centrage sur un nom d'affect à l'aspect ponctuel dominant, ce qui accroche la curiosité du lecteur. Ce dernier attend de savoir quelles sont les causes de la stupeur, parmi un éventail des possibles très ouvert. Le nom fonctionne alors comme un embrayeur d'argumentation pour lancer le travail textuel d'étayage (Novakova & Sorba 2013b). Les lexies *stupeur/stupéfier* apparaissent donc souvent comme la manifestation linguistique du désaccord résultant du décalage entre les attentes culturelles topiques (le *devoir éprouver* légitime) et la réalité décrite par le journaliste. Cette justification représente le versant textuel de l'argumentation grâce auquel le sujet argumentant, le journaliste, entend assurer la solidité de sa position, l'une des conditions d'émergence de l'activité argumentative étant l'existence préalable d'un désaccord<sup>48</sup>.

À l'intérieur de l'ensemble des lexies d'affect, nous avons pu observer qu'une configuration particulière de l'environnement textuel permettait distinguer les lexies d'affects causés (par ex., *jubilation, stupeur, déception*) des lexies d'affects interpersonnels (par ex., *dédain, jalousie, estime*). En effet, les lexies d'affects interpersonnels produisent beaucoup plus de séries avec d'autres adjectifs ou noms d'affect (Novakova & Sorba 2014a) :

20) Je me suis dit que la femme que vous aimiez est heureuse car vous devez être avec elle  
**exigeant, insupportable, colérique, jaloux, invivable.** (*Libération* 14/06/2007)

---

<sup>48</sup> Cette thèse classique trouve ses origines dans la *Rhétorique* d'Aristote (1356b) ; l'argumentation est bien « un mode spécifique du traitement du désaccord » pour reprendre la formule de Jacquin & Micheli (2012, p. 601).

21) Les joueurs sont applaudis dans la rue, signent des autographes. « On les a vus les épaules se redresser à mesure qu'ils retrouvaient le **respect**, la **confiance** et l'**estime** de soi qu'ils avaient perdus. » (*Libération* 30/07/2007)

Cette propriété des interpersonnels est utilisée comme moyen de justification d'un état des choses, comme une *ressource pour l'argumentation*. La structure cumulative des émotions en série contribue à la mise en relief du message : elle aide au martèlement des opinions exprimées (Novakova & Sorba 2018).

À l'issue de ces différentes études, nous avons proposé un modèle fonctionnel pour l'analyse discursive des lexies d'affect (Novakova & Sorba 2018). À partir du tableau 2 qui récapitule nos observables et nos observations, nous avons construit notre modèle fonctionnel qui relie tous les éléments du profil discursif sur les deux niveaux phrastique et transphrastique (figure 3) :

<b>Paramètres liés au profil discursif</b>	<b>Affects interpersonnels</b>	<b>Affects causés</b>
Structures actanciennes	Plus fournies (2 ou 3 Asé <sup>49</sup> ) Complétude, insistance	Moins fournies (2, 1 ou 0 Asé) Économie
Colligations phrastiques	Préférence pour la fonction COD Commentaire descriptif	Préférence pour la position frontale, en début de phrase Effet de projecteur, mise en relief
Associations sémantiques	Moins d'intensité Plus d'emploi en séries avec d'autres lexies d'affect	Plus d'intensité Moins d'emploi en séries avec d'autres lexies d'affect
Isotopies textuelles	Moins d'intensité	Plus d'intensité
Colligations textuelles / scénarios discursifs	Position plutôt médiane Scénario plus complexe	Position initiale : effet d'accroche, travail d'étayage Scénario moins complexe
Rubriques	Politique et société	Culture et arts

Tableau 2 : Profils discursifs des lexies d'affects interpersonnels et causés (d'après Novakova & Sorba 2018, p. 167).

<sup>49</sup> Asé = actant sémantique.

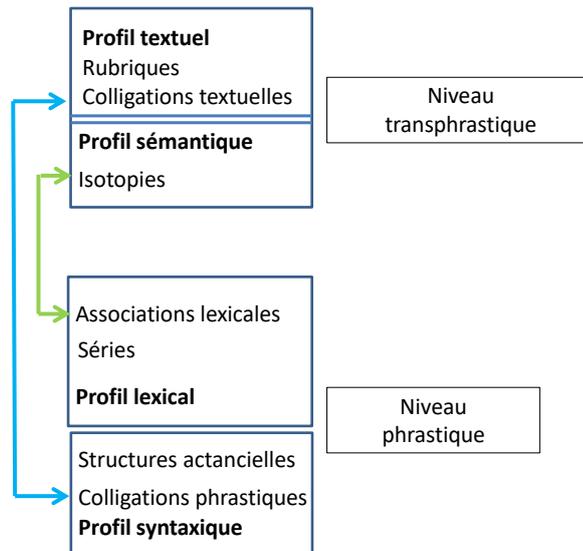


Figure 3 : Modèle fonctionnel pour l'analyse des lexies d'affect (Novakova & Sorba 2018, p. 168).

Dans la figure 3, pour chaque dimension du profil (lexicale, syntaxique, sémantique, textuelle), nous indiquons les observables (rubriques, colligations textuelles, isotopies, etc.) qui nous ont permis d'aboutir à cette modélisation. Par ailleurs, nos résultats révèlent la correspondance systématique entre les niveaux phrastique et transphrastique, visualisée par des flèches sur la figure 3. Sur le plan sémantique par exemple, comme on peut le voir dans le tableau 2, la dimension intensité, plus spécifique aux lexies d'affects causés, se retrouve à la fois dans les associations lexicales (niveau phrastique) et dans les isotopies (niveau transphrastique). Dans la lignée des approches fonctionnelles qui « étudient l'interaction entre les paramètres syntaxiques, sémantiques et discursifs du langage et leur interdépendance vis-à-vis de sa fonction de communication » (Novakova 2010, p. 25), nous ancrons nos analyses dans une démarche holistique qui place l'usage au cœur de ses descriptions de la langue<sup>50</sup>.

<sup>50</sup> Comme le rappelle François (2004), dans le fonctionnalisme linguistique contemporain : « Le fond du débat, c'est la question essentielle du sens donné à l'expressions "*expliquer un fait linguistique*" » (p. 101). Il convient ainsi de considérer que « la structure du langage ne peut être expliquée [...] qu'en termes de discours et de parole, ce qui contredit le souci des formalistes de se cantonner au niveau de l'unité-phrase et de ne tenir compte que de la fonction d'information au détriment de la fonction d'interaction. » (p. 104).

Dans ce deuxième chapitre, j'ai pu observer la mise en œuvre de ce que Combettes nomme la « transversalité » des niveaux de l'analyse linguistique (2012, p. 4). J'ai montré dans mes travaux comment se tissait l'affinité entre une lexie et le palier du texte, en particulier dans le cas des lexies d'affect et d'un type de séquence textuelle particulier, la séquence argumentative. Dans les recherches présentées ici, j'ai donc bien procédé à une « extension du domaine de la phraséologie » en développant une approche qui inclut la linguistique du discours, en illustrant le concept de « portée textuelle » et en montrant l'impact d'une classe de lexies sur la structuration de son environnement textuel.

Les travaux présentés dans cette première partie offrent de solides prémices pour enclencher une autre étape de ma recherche qui vise à affiner la modélisation du lien entre les différentes unités phraséologiques et le texte. L'extension des analyses phraséologiques au niveau discursif a profondément remodelé les contours de la phraséologie dans une dynamique de passage de la phrase au niveau supérieur du paragraphe, du texte mais aussi du genre textuel comme nous allons le voir à présent dans la deuxième partie.



## Partie 2. Pour un renouveau de l'approche générique

« Il n'y a pas de texte sans genre. »

(Derrida 1986, p. 264)

La question du *genre* remonte aux traditions antiques de notre monde occidental. Aristote est le premier auteur grec à aborder l'art poétique sous l'angle générique comme en témoigne le début de sa *Poétique* :

περὶ ποιητικῆς αὐτῆς τε καὶ τῶν εἰδῶν αὐτῆς, ἥν τινα δύναμιν ἕκαστον ἔχει, καὶ πῶς δεῖ συνίστασθαι τοὺς μύθους, εἰ μέλλει καλῶς ἔξιν ἢ ποιήσις, ἔτι δὲ ἐκ πόσων καὶ ποίων ἐστὶ μορίων, ὁμοίως δὲ καὶ περὶ τῶν ἄλλων ὅσα τῆς αὐτῆς ἐστὶ μεθόδου, λέγωμεν ... [1447a]<sup>51</sup>

« Parlons de l'art poétique lui-même et de ses espèces, du pouvoir que possède chacune d'elles, ainsi que de la manière dont il faut composer les histoires si l'on veut que l'œuvre soit réussie, et encore au sujet du nombre et de la nature de ses parties constitutives ainsi que de même tous les autres sujets relatifs à cette recherche. »

L'art poétique (ποιητικῆς) serait donc constitué de plusieurs espèces (τῶν εἰδῶν) qui possèdent chacune des propriétés distinctes et mettent en œuvre un ensemble de normes (déontique δεῖ) pour leur composition respective. Cette théorisation d'Aristote est une première réflexion pour rendre compte des différences entre les productions textuelles. Mais l'étude de la postérité d'Aristote révèle aussi bien l'abondance des réflexions menées sur la question générique dans son sillage que l'impasse dans laquelle elle se trouve :

Aucun des rares illustres successeurs d'Aristote n'a réussi à aller plus loin que l'auteur de la *Poétique*, chacun s'ingéniant au contraire à rendre les problèmes encore plus insolubles que ne les avait déjà rendus son prédécesseur. (Schaeffer 1989, p. 63)

Ce constat implacable est aussi celui de Bronckart (1996) :

Comme chacun le sait, depuis l'Antiquité grecque au moins, face à l'hétérogénéité évidente du « monde des textes », les propositions de catégorisation et d'organisation ont abondé, et à commencer par Aristote, nombre de grands penseurs occidentaux ont apporté leur pierre à l'édifice. Édifice qui demeure aujourd'hui encore cependant toujours instable, en ce sens qu'aucun classement n'a réussi à s'imposer, et qu'aucune véritable synthèse ne peut être opérée sur la diversité des propositions existantes. (Bronckart 1996, p. 2)

Néanmoins, nous défendons, pour notre part, la nécessité de ce que Rastier nomme une « théorie opératoire des genres » (Rastier 2011, p. 72). L'explosion numérique de l'écrit a entraîné la création de nombreuses banques de données textuelles et pose actuellement la question de

---

<sup>51</sup> Aristote. *Ars Poetica*. (1966). R. Kassel (éd.). Clarendon Press.

savoir comment classer les textes. La catégorie de genre, traditionnellement utilisée pour regrouper des textes en réduisant leur variabilité, est fort utile dans cette tâche.

Cette deuxième partie se présente comme un état des lieux raisonné des travaux sur le genre textuel et propose des pistes pour en renouveler l'approche. Dans cette optique, je propose d'examiner la question du genre en croisant les quatre approches qui en ont traité : la poétique, l'analyse du discours, la linguistique textuelle et la sémantique textuelle. Dans un premier temps, le chapitre 3 s'interrogera sur la manière de saisir le générique : où se réalise le genre ? comment peut-on le saisir ? Les réponses apportées à ces premières questions se sont souvent traduites, à l'échelle de l'histoire, par l'établissement de diverses typologies. C'est précisément le point traité dans le chapitre 4 qui questionnera l'entreprise typologique dans une perspective historique ainsi que les conditions de sa réussite.

## Chapitre 3. La question du genre : comment saisir le générique ?

Comme le souligne Poudat (2020), la notion de genre est problématique pour la description linguistique car « elle est paradoxalement de plus en plus mobilisée et de moins en moins définie » (p. 19). L'objectif de ce troisième chapitre est donc d'apporter des éléments définitoires sur la question du genre textuel pour situer le cadre de ma recherche. Dans un premier temps, je préciserai les contours du lieu où le générique – c'est-à-dire les propriétés constitutives du genre – peut se saisir. Cette mise au point théorique permettra de me situer dans la riche réflexion menée sur la nature de cet objet langagier (section 3.1). Dans un second temps, je réfléchirai à la conceptualisation possible de la notion de genre pour en proposer une définition (section 3.2).

### 3.1 Où saisir le générique ?

À la suite de Rastier (2005a), je considère que :

[...] l'objet empirique de la linguistique est fait de *textes* oraux ou écrits, non de mots ou de phrases – qui ne s'observent pas à l'état isolé, et, même quand on les isole, restent toujours relatifs à un genre et un discours. [...] Aucun texte n'est écrit seulement « dans une langue » : il est écrit dans un genre et au sein d'un discours, en tenant compte évidemment des contraintes d'une langue. (Rastier 2005a, p. 31 et 34)

C'est donc cet objet empirique, les textes<sup>52</sup>, qui matérialise le lieu où saisir le générique et que nous proposons de définir dans cette première section. Notre démarche s'inscrit pleinement dans un cadre linguistique même si la question des genres est un objet abordable depuis des points de vue disciplinaires très différents (Adam 2015) comme nous le verrons par la suite. Cette première sous-section (3.1.1) est donc consacrée à la définition de l'objet texte et aux critères qui permettent de fonder le jugement de textualité.

#### 3.1.1 Le texte, « un objet de haute complexité »

La formule d'Adam (2015), que nous avons reprise dans le titre de cette sous-section, résume bien un consensus actuel sur la nature complexe du texte. Cette complexité est d'abord d'ordre sémiotique comme le rappelle Schaeffer (1989, p. 132) pour qui les textes demeurent « des

---

<sup>52</sup> L'emploi du pluriel *textes* est ici conforme à la proposition d'Adam pour qui : « LE TEXTE comme objet abstrait est l'objet d'une théorie générale des agencements d'unités [...] au sein d'un tout de rang de complexité linguistique plus ou moins élevé » alors qu'il considère « LES TEXTES comme objets concrets, matériels, empiriques. » (Adam 1999, p. 40).

actes sémiotiquement complexes ». En effet, cette activité langagière est productrice de signes qui encodent de larges couches du monde représenté : « Les textes sont ces modalités de structuration de l'activité langagière par lesquelles les aspects illocutoires et locutoires sont intégrés, qui disent le monde en agissant dans le monde. » (Bronckart 1996, p. 9). Le texte est bien un signe, une « unité sémantique » (*a semantic unit*) pour reprendre l'expression déjà ancienne d'Halliday & Hasan (1976, p. 2).

En tant que tel, il met donc en œuvre un processus interprétatif voire plusieurs, si l'on suit la définition désormais classique de Charaudeau & Maingueneau (2002) :

Opposer *texte écrit* et *discours oral* [...] dissimule le fait qu'un texte est, la plupart du temps, *plurisémiotique*. Une recette de cuisine, un placard publicitaire ou un article de journal, un discours politique, un cours universitaire ou une conversation, ne comportent pas que des signes verbaux, ils sont également faits de gestes, d'intonations et d'images (photographies et photogrammes, dessins et infographies). (p. 570-571).

Ici, le point de vue sur le texte comme objet *plurisémiotique* se pose en rupture radicale avec les conceptions antérieures qui définissaient le texte comme « tout discours fixé par l'écriture », pour reprendre les mots de Ricoeur (1986, p. 137)<sup>53</sup>. Le texte est désormais conçu comme un ensemble de signes non exclusivement verbaux. Même si mes travaux s'intéressent à des textes écrits exclusivement composés de signes verbaux, je partage néanmoins pleinement cette conception contemporaine du texte et la mets en œuvre régulièrement dans ma pratique pédagogique.

En outre, la complexité sémiotique du texte se mesure aussi dans la nécessaire immersion intertextuelle mise en œuvre lors de son interprétation. C'est ce que rappelle Rabatel (2010) quand il discute le sens étymologique de la lexie *texte* :

La métaphore du texte comme *tissu* rassemble des éléments hétérogènes, mais convergents [...] [mais] ne saurait faire oublier que le texte n'est pas qu'une structure close (*textus*) : c'est aussi,

---

<sup>53</sup> Le rejet de cette conception du texte se retrouve déjà, par exemple, dans la définition de Weinrich (1989) : « Nous appelons TEXTE l'énoncé linéaire qui est compris entre deux interruptions remarquables de la communication et qui va des organes de la parole ou de l'écriture de l'émetteur aux organes de l'audition ou de la vue du récepteur. Ce disant, nous ne comptons pas comme interruption de la communication le simple changement de locuteur en cours d'échange. Un dialogue est aussi un texte au sens de la linguistique textuelle. » (p. 24). Voir aussi Coseriu (2007, p. 86 cité par Adam 2011, p. 19) qui définit le texte comme « une série d'*actes linguistiques* connexes que réalise un locuteur dans une situation concrète qui, naturellement, peut prendre une forme parlée ou écrite. ».

ainsi que le disait Quintilien, un espace ouvert (*textum*) vers l'amont des textes antérieurs ou l'avant des interprétations. (Rabatel 2010, p. 177)

Tout en rappelant la complexité inhérente à un objet constitué d'éléments hétérogènes, Rabatel indique les limites de la métaphore comparant le texte à un tissu. Comme un tissu, le texte est un produit fini mais contrairement à un tissu, un texte ne contient pas sa propre finitude en lui-même : il s'inscrit dans la lignée des productions textuelles antérieures et devient, à son tour, lignée de celles qui lui succèdent.

Ces liens génétiques de descendance et d'ascendance, qui contribuent à la complexité de cet objet langagier, s'observent dans l'hétérogénéité énonciative constitutive de tout texte<sup>54</sup> :

[...] un 'texte' n'est pas une entité énonciative homogène. Il se présente en général comme une succession, ou un emboîtement selon les cas, d'isotopies énonciatives, qui s'opposent les unes aux autres par la nature et/ou la modalité d'inscription de L dans l'énoncé. (Kerbrat-Orecchioni 1980, p. 162)

L'émetteur L est alors perçu comme un chef d'orchestre qui organise la polyphonie<sup>55</sup> au sein de son énoncé. Cette prise en charge de l'hétérogénéité énonciative est propre à chaque émetteur<sup>56</sup>, ce qui permet d'envisager le texte comme « le résultat toujours singulier d'un acte d'énonciation [et comme], par excellence, l'unité de l'interaction humaine. » (Adam, 1999, p.40).

Ce rapide panorama des facteurs faisant du texte « un objet de haute complexité » (plurisémiotisme, intertextualité, hétérogénéité énonciative) permet également de révéler l'intérêt de l'approche diachronique dans l'étude des textes<sup>57</sup> quand il s'agit de tracer des lignées

---

<sup>54</sup> Cette idée de l'hétérogénéité énonciative avait déjà été formulée par Foucault à propos des énoncés : « un énoncé a toujours des marges peuplées d'autres énoncés » (Foucault 1969, p. 128).

<sup>55</sup> Nous utilisons le terme *polyphonie* dans le sens le plus large qui couvre les deux usages rappelés par Carel (2011, § 1) : « Le terme "polyphonie" renvoie à des phénomènes que l'on peut classer en deux familles : ceux qui concernent l'allusion, par un unique énoncé, à plusieurs contenus ; et ceux qui concernent la présence de plusieurs instances énonçantes à l'intérieur de l'énonciation. ».

<sup>56</sup> Sur ce point voir aussi Bronckart (1996) : « les textes sont des séquences organisées de comportements verbaux, oraux ou écrits, qui sont ascriptibles à un agent singulier, dans un contexte d'action déterminé. [...] les productions textuelles sont le médium au travers duquel se construisent les représentations sociales et rationnelles qui permettent de situer et de juger chaque contribution singulière à l'activité. » (p. 8-9).

<sup>57</sup> L'approche diachronique est également mise en œuvre quand on considère que les textes ne sont pas des données, mais des « constructions issues de procédures médiatrices » (Adam 2015, p. 26) qui vont de la rature et de la réécriture par les auteurs ou les copistes jusqu'aux variations éditoriales, en passant par les traductions. Il existe bien un temps de l'écriture et de la vie des textes après leur écriture qui font de chacun d'entre eux un produit

de paroles entre des textes. C'est précisément l'un de mes premiers centres d'intérêt que j'ai pu mettre en œuvre pour examiner le sort de certaines unités phraséologiques homériques dans le corpus tragique (Sorba 2008b, 2015a). Sachant que la tragédie grecque se nourrit de l'épopée (Alaux 2006), j'ai analysé les variations des formules homériques comprenant des lexies du champ lexical de la mer (ἡ ἄλς, ἡ θάλασσα, τὸ πέλαγος, ὁ πόντος et leurs dérivés) pour montrer que le poète tragique Eschyle détourne le formulaire homérique pour innover tout en s'inscrivant dans le cadre d'un héritage. De même, j'ai mené une enquête en diachronie longue sur l'évolution de la lexie homérique ἡ ῥηγμῖν « rivage, brisant » qui a survécu de manière artificielle jusqu'au grec byzantin grâce aux pratiques d'un lexicographe du 12<sup>e</sup> siècle : « Utilisée à l'origine dans une phraséologie homérique très codifiée, ἡ ῥηγμῖν s'est retrouvée, au fil du temps, au sein d'un système de réminiscences complexes faisant écho au grand modèle, mais renvoyant aussi aux héritiers entre eux. » (Sorba 2014b, p. 260).

Analyser les facteurs à l'origine de la complexité des textes est une première étape dans la définition de cet objet empirique de la linguistique. Une deuxième étape est d'envisager les critères qui permettent de dire aux lecteurs que nous sommes, que ce qu'ils lisent peut être appelé « texte ».

### 3.1.2 Sur quels critères se fonde le jugement de textualité ?

C'est une expérience instructive et amusante que nous sommes sans doute nombreux à mener au début d'un cours de linguistique textuelle : demander aux étudiants d'écrire ce qui leur passe par la tête pendant 5 minutes et ensuite lire leur production à haute voix en leur demandant si l'on peut qualifier de *texte* cette production<sup>58</sup>. Cet exercice tout simple permet de faire prendre conscience aux étudiants de l'un des acquis de la linguistique textuelle :

A text [...] is not just a string of sentences. In other words, it is not simply a large grammatical unit, something of the same kind as a sentence but differing from it in size – a sort of supersentence. A text is best thought of not as a grammatical unit at all, but rather as a unit of a different kind: a semantic unit. The unity that it has is a unity of meaning in context, a texture that

---

soumis à variation. C'est tout l'objet des études de génétique textuelle, notamment celle de l'équipe de Lausanne avec laquelle nous travaillons actuellement sur le projet « Approche numérique des textes littéraires : vers une convergence méthodologique ? » (voir annexe IV) mais ce n'est pas l'approche que je développe ici.

<sup>58</sup> L'idée m'a été soufflée, il y a fort longtemps maintenant, par mon cher collègue Jean-Marc Colletta, Professeur de Sciences du langage (UGA, LiDiLEM) décédé prématurément en septembre 2021, à qui je rends hommage par cette mention.

expresses the fact that it relates as a whole to the environment in which it is placed. (Halliday & Hasan 1976, p. 293)

Cette conscientisation permet de voir le texte non pas comme un simple enchaînement de phrases mais comme une unité d'une autre espèce qui entretient un lien inextricable avec son environnement.

Pour Adam (2015), ce jugement de textualité se fonde « sur un triple sentiment : de connexité (liages micro-textuels des énoncés), de cohésion (sentiment de totalité locale et globale, des parties elles-mêmes et des parties en relation avec le tout textuel) et de cohérence (sentiment d'adéquation des énoncés à une situation socio-discursive et à un genre de discours). » (p. 45-46). Comme il le rappelle ailleurs (1993), la connexité s'envisage sur plusieurs plans : dans la structuration morphosyntaxique de chaque unité (proposition-phrase), dans le cadre de la tension textuelle qui assure la reprise-répétition (la continuité textuelle) tout en garantissant la progression ou encore dans les phénomènes de démarcation graphique locale et de marquage global du plan de texte (Adam 1993, § 33-35). La cohésion ou « dimension sémantique globale » (*ibid.*, § 24) est quant à elle représentée par la macro-structure sémantique soit le thème global d'un énoncé :

La notion sémantique de cohésion a pour but de répondre à des questions naïves : comment expliquer le fait que, quand on lit et comprend un énoncé, on éprouve ou non un sentiment d'unité ? Comment rendre compte sémantiquement du fait qu'une phrase ne soit pas un tas de mots et un texte une simple juxtaposition de phrases ? Distinguée de la connexité interne aux formants linguistiques d'une expression (des lettres/sons aux composants morpho-syntaxiques) et opérant aussi de phrase en phrase [...], distinguée également de la **cohérence** [...] et de la pertinence contextuelle, la cohésion sémantique est un fait de co-textualité que la notion d'isotopie permet de théoriser. (Adam 1993, § 31)

Cohésion, connexité et cohérence sont donc les trois critères qui permettent de fonder le jugement de textualité.

Le genre fait partie de l'un de ces trois critères définitoires de la textualité : il relève de la cohérence, une compétence linguistique<sup>59</sup> à part entière. Il peut être considéré comme un

---

<sup>59</sup> « La cohérence fait partie de la compétence linguistique de deux manières : d'une part, l'homme est capable de produire des textes, c'est-à-dire des suites cohérentes de phrases, d'autre part, il est en mesure de décider si une suite de phrases est cohérente ou non et si elle constitue un texte ou non. » (Lundquist, 1980, p. 20). Cette capacité humaine de percevoir et de mettre en oeuvre la cohérence a été théorisée par Charolles (1989) sous la forme de quatre métarègles : la *répétition* (pour qu'un texte soit cohérent, il faut qu'il comporte dans son développement linéaire des éléments à récurrence stricte), la *progression* (pour qu'un texte soit cohérent, il faut que son

élément de « l'environnement dans lequel est placé le texte » pour reprendre les mots de Halliday & Hasan ci-dessus – au service de la réalisation de la textualité en complémentarité des deux autres critères que sont la connexité et la cohésion. Le genre qui relève de la cohérence joue bien un rôle dans la définition du texte. Rastier (2003 ; 2005) le mentionne d'ailleurs comme l'un des éléments permettant de normer la mise en corrélation des deux plans qui constituent le texte, celui des formes sémantiques et celui des formes expressives : « un genre définit précisément un rapport normé entre signifiant et signifié au palier textuel : par exemple, dans le genre de l'article scientifique, le premier paragraphe, sur le plan du signifiant, correspond ordinairement à une introduction, sur le plan du signifié ; dans le genre de la nouvelle, il s'agit le plus souvent d'une description. » (Rastier 2005a, p. 40).

Cette première sous-section (3.1) a précisé l'ancrage théorique sur lequel j'appuie ma conception du texte et des critères qui font la textualité. Le critère « genre » est bien de ceux-là. C'est cette notion sur laquelle je propose désormais de réfléchir en vue d'en ébaucher une première définition.

### **3.2. Comment définir le genre ?**

La consultation des dictionnaires<sup>60</sup> à l'entrée « genre » permet de rappeler le caractère éminemment polysémique de cette lexie héritée du latin *genus, eris, n.* À titre d'illustration, je récapitule dans le tableau 3 ci-dessous les dix acceptions répertoriées par *Le Petit Robert électronique* :

---

développement s'accompagne d'un apport sémantique constamment renouvelé), la *non-contradiction* (pour qu'un texte soit cohérent, il faut que son développement n'introduise aucun élément sémantique contredisant un contenu posé ou présupposé par une occurrence antérieure ou déductible de celle-ci par inférence) et la *relation* (pour qu'une séquence ou un texte soient cohérents, il faut que les faits qu'ils dénotent dans le monde soient reliés).

<sup>60</sup> La notion de *genre textuel* est progressivement entrée dans les dictionnaires de linguistique. Par exemple, en 1994, cette entrée n'existe pas dans le *Dictionnaire de linguistique et des sciences du langage* (Dubois et al.) qui ne répertorie que l'acception grammaticale du terme.

I. Ensemble d'êtres ou d'objets présentant des caractères communs	1. (début XIV <sup>e</sup> ) Didact. Idée générale d'un groupe d'êtres ou d'objets présentant des caractères communs.
	2. (début XIII <sup>e</sup> ) <i>Le genre humain</i> : l'ensemble des êtres humains.
	3. Phylog. Subdivision de la classification (taxinomie) des êtres vivants, située au-dessous de la famille.
	4. (anglais <i>gender</i> ) Didact. Construction sociale de l'identité sexuelle.
	5. (1549) Catégorie d'œuvres, définie par la tradition (d'après le sujet, le ton, le style).
	6. espèce, sorte, type.
	7. (XVII <sup>e</sup> ) <i>genre de vie</i> : façon de vivre, ensemble des habitudes d'un individu ou d'un groupe d'individus.
II. Catégorie grammaticale (milieu XIII <sup>e</sup> ) Catégorie exprimant parfois l'appartenance au sexe masculin, au sexe féminin ou aux choses (neutre).	
III. Façons de se comporter	1. Façons de s'habiller, de se comporter.
	2. <i>Faire du genre, se donner un genre</i> ; fam. <i>faire genre</i> : affecter certaines manières.

Tableau 3 : les acceptions de la lexie « genre » dans *Le Petit Robert électronique*.

La notion de « genre » à l'œuvre dans les textes relève de la première acception (I) : le genre textuel est bien un ensemble d'objets, les textes, présentant des caractères communs. Cette acception I.1 est ancienne en français (début 14<sup>e</sup> siècle) et apparaît avec une visée taxonomique et didactique qui se décline ensuite en différentes typologies (êtres vivants, œuvres, habitudes, etc.) comme l'indiquent les différentes acceptions. La notion de genre est donc liée, en tout premier lieu, à celle de catégorisation, une compétence cognitive, comme nous allons le montrer dans la première sous-section (3.2.1).

### 3.2.1 Le genre : la mise en œuvre d'une compétence cognitive

Distinguer une recette de cuisine d'un poème ou d'un manuel de mathématiques est à la portée de n'importe quel locuteur à partir du moment où il a été exposé à ces productions textuelles :

Any native speaker of a language will in principle be able to make a distinction between a poem and a hand-book of mathematics, between an article in a newspaper and a questionnaire. This implies that he has the initial ability to differentiate the universe of texts and to recognize different types of texts. We shall claim [...] that this fundamental ability is part of linguistic competence. We shall argue at the same time that this competence must be a textual competence. » (Van Dijk 1972, p. 297-298).

Tout locuteur met ainsi en œuvre une compétence cognitive qui se manifeste par une intuition pour catégoriser les différentes productions textuelles : le classement générique est certainement l'un des tout premiers auquel procèdent les locuteurs d'une même communauté linguistique. En effet, comme le souligne Bronckart (1996), tout membre d'une communauté linguistique est confronté à un univers de textes « déjà là », un univers organisé en « genres » empiriques et historiques, c'est-à-dire en formes d'organisation concrètes. La confrontation des genres préexistants produit, chez les locuteurs, une manière de connaissance intuitive de leurs règles et propriétés spécifiques. C'est cette connaissance intuitive que Jauss (1970) a qualifié, en son temps, d'« horizon d'attente ».

Le genre fait donc l'objet d'une construction, c'est bien un « artefact », selon les mots de Kerbrat-Orecchioni (1980) :

Ce terme 'genre' dénote un 'artefact', un objet construit, par abstraction généralisante, à partir de ces objets empiriques que sont les textes, qui ne sont jamais que des représentants impurs de tel ou tel genre : tel texte se caractérise par un certain taux de poéticité, de polémique, etc. Tout genre se définit comme une constellation de propriétés scientifiques, que l'on peut appeler des 'typologèmes', et qui relèvent d'axes distinctifs hétérogènes (syntaxiques, sémantiques, rhétoriques, pragmatiques, extralinguistiques, etc. (p. 170)

Cette construction repose sur l'identification de « propriétés scientifiques » – à comprendre ici comme issus de la méthode expérimentale d'observation – sur différents plans linguistiques ou extralinguistiques. En tant que construction mentale, le genre est donc objectivable par l'analyste qui s'appuie sur l'observation de ses propriétés. L'analyste procède à une catégorisation, à partir des observations qu'il a réalisées sur plusieurs textes, pour réunir ceux-ci au sein d'une même catégorie dénommée « genre ». Ainsi, « un genre se réfère donc au texte dans l'« unité » qu'il constitue, pour définir sa « parenté », ou son « air de famille », avec d'autres unités-textes. » (Baroni 2009, § 21). La distribution de textes au sein d'une même catégorie repose sur la faculté de l'analyste de dégager des phénomènes réguliers au sein de ses observations linguistiques et extralinguistiques. C'est sur la base de ce principe que Baroni (2009) propose de définir le genre comme « la régularité au sein de l'interdiscours » (§ 10). L'une des implications de cette conception du genre est donc que, comme le notent Ablali, Bouhouhou & Tebbaa (2015), « le genre bat en brèche la clôture du texte [... car il] suppose une ouverture au monde au-delà de l'analyse de l'œuvre considérée. » (p. 11). C'est cette ouverture du texte au monde que nous allons désormais interroger.

### 3.2.2 Le genre : une partie constitutive du discours

La définition de Rastier (2005a, voir *supra*) selon laquelle un texte « est écrit dans un genre et au sein d'un discours » introduit une distinction entre deux lieux de réalisation d'un texte : le genre et le discours. L'objectif de cette sous-section est bien d'envisager le rapport entre genre et discours.

Le texte est un objet dont on peut décrire la structure et le contenu informationnel à priori, mais dont la signification et les effets ne peuvent être décrits qu'une fois le texte actualisé en discours dans une situation de communication effective. Malrieu & Rastier (2001) distinguent quatre niveaux hiérarchiques supérieurs au texte : les discours (juridique, littéraire, scientifique, politique, etc.), le champ générique<sup>61</sup> (théâtre, poésie, récit, etc.), le genre proprement dit (comédie, roman sérieux, nouvelle, conte, etc.) et éventuellement le sous-genre (roman par lettres, etc.). Au niveau inférieur de la classification, se trouvent les textes d'un même auteur.

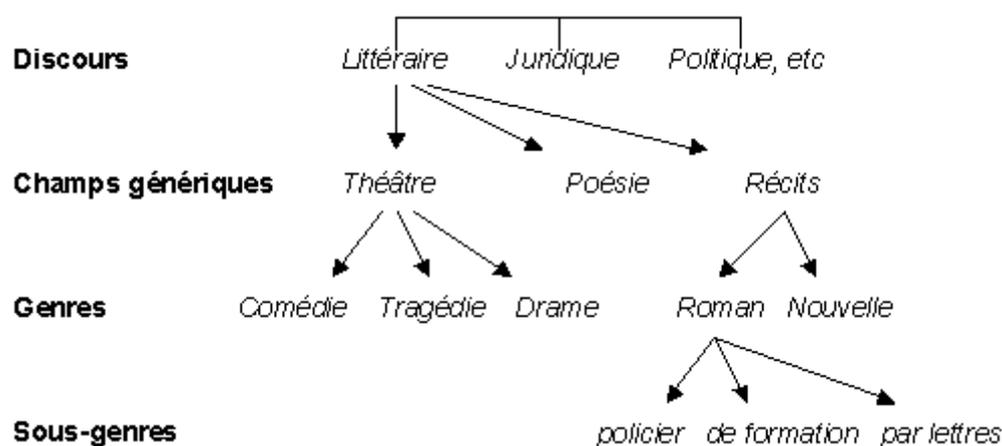


Figure 4 : Les quatre niveaux hiérarchiques supérieurs au texte (Malrieu & Rastier 2001, p. 548).

Cette représentation en niveaux hiérarchiques a le mérite d'identifier, selon Rastier (2005a), « les niveaux stratégiques qui permettent de passer de la généralité de la langue aux particularités des textes, car les relations sémantiques entre les textes s'établissent préférentiellement entre les textes du même genre, du même champ générique et du même discours. » (p. 35). Se trouve ici l'un des fondements de la sémantique textuelle défendue par Rastier : les relations sémantiques entretenues par les textes les uns avec les autres – en somme

---

<sup>61</sup> « Un champ générique est un groupe de genres qui contrastent voire rivalisent dans une pratique. » (Rastier 2005, p. 33).

ce que l'on appelle « intertextualité » – doivent être replacées au cœur des théories de la textualité pour permettre l'interprétation correcte d'un texte (Rastier 2011, p. 23).

Le rapport d'inclusion du genre au sein du discours a été théorisé, dans le champ de l'analyse du discours, par Maingueneau (2009) qui nomme *genre de discours* « des groupements de textes fondés sur les critères les plus variés (secteur d'activité sociale, thématique, règles de composition, présence de tels ou tels traits linguistiques, etc.) » (p. 68). Cette définition très générale permet de distinguer le secteur des productions verbales d'une société (le discours) du dispositif de communication particulier socio-historiquement défini mis en œuvre dans ces productions (le genre). Pour reprendre les exemples de Maingueneau, ce système d'inclusion permet de décliner le discours religieux en genres tels que le sermon, la messe, le chant, le missel, etc., ou le discours politique en genres tels que le débat télévisé, le tract, le programme électoral, etc. La notion de *genre de discours* ne fonctionne pleinement que pour les activités verbales qui ne relèvent pas de l'oral spontané, par essence instable, c'est-à-dire pour ce que Maingueneau nomme les *genres institués* (voir par ex., Maingueneau 2007a). Il situe ceux-ci sur une échelle qui va des *genres routiniers* aux *genres auctoriaux* : dans les premiers, les partenaires de l'activité verbale se conforment aux contraintes du genre (par ex., la consultation médicale, la dissertation) ; dans les seconds, l'auteur décide d'attribuer un genre à son texte (par ex., c'est le journaliste qui appelle son texte « chronique » et indique au lecteur comment il doit le lire mais cette liberté est toutefois limitée, la liberté la plus grande de l'auteur se trouvant dans le discours littéraire). À l'intérieur des genres institués, Maingueneau (2007b) distingue précisément quatre modes de généricité :

- (i) *Genres institués de mode (1)* : les textes n'y sont pas ou peu sujets à variation et caractérisés par des formules et des schèmes compositionnels préétablis (par ex., courrier commercial, annuaire téléphonique, actes notariés, etc.) ; les locuteurs doivent être pratiquement interchangeables ;
- (ii) *Genres institués de mode (2)* : les locuteurs produisent des textes individués mais soumis à de fortes contraintes (guide de voyages, faits divers, etc.) ; la scénographie est attendue, préférentielle mais quelques écarts sont possibles (un manuel de géographie sous forme d'un récit d'aventure présente ainsi une scénographie atypique) ;
- (iii) *Genres institués de mode (3)* : pas de scénographie préférentielle ni prédictible (par ex., chansons, annonces publicitaires, etc.) mais pas de vocation à remettre en cause la scène générique ;
- (iv) *Genres institués de mode (4)* ou « genres "proprement" auctoriaux » : les textes ne correspondent pas ici à des activités discursives bien balisées dans l'espace social ;

« l’auteur contribue à construire le contexte de son énonciation et prétend définir le sens de l’activité dans laquelle il engage son destinataire. » (*ibid.*, p. 61). Dans ce mode (4), certaines étiquettes formelles indiquent des modes d’organisation textuelle (les *hypergenres* « lettre, essai, journal, dialogue, etc. » permettent de formater le texte) ; d’autres sont destinées à fournir un cadrage interprétatif (surtout dans les œuvres postérieures au 18<sup>e</sup> siècle) ; d’autres enfin fournissent des indications généalogiques pour des œuvres élaborées par ressemblance plus ou moins grande avec une ou plusieurs œuvres prototypiques.

L’ensemble de mes travaux scientifiques s’attache à l’étude de textes relevant des genres institués et plus spécifiquement à ceux des genres auctoriaux (mode 4) : d’une part, les genres auctoriaux relevant du discours littéraire de différentes époques et dans différentes langues (tragédie, comédie, épopée, roman, hymne) ; d’autre part, ceux relevant du discours journalistique. Je propose d’aborder ici ceux de mes travaux qui se sont appuyés sur un corpus journalistique et de montrer comment la réflexion générique y a été abordée.

Le corpus journalistique est celui du projet EMOLEX, constitué des deux années de parution (2007-2008) des quotidiens *Le Monde*, *Libération*, *Ouest-France* et *Le Figaro*. La perspective générique y a été abordée sous l’angle des rubriques journalistiques. En effet, à l’instar de Herman & Lugin (1999), nous considérons la rubrique comme « un outil de classification et de hiérarchisation de l’information » (p. 72). À partir du modèle proposé par Malrieu & Rastier (figure 4), nous avons posé la rubrique comme un sous-genre textuel. La distinction entre les différentes rubriques a été faite sur une base thématique à partir de la proposition de Grosse (2001) pour qui l’identification d’une rubrique repose sur « l’intention dominante et manifeste (affichée), c’est-à-dire telle que *les lecteurs* doivent la comprendre » (p. 51). Nous avons donc organisé les niveaux hiérarchiques de la manière suivante :

*discours* : journalistique → *champ générique* : presse écrite → *genre* : presse écrite quotidienne → *sous-genre* : rubrique (politique, culture, société, économie, justice, sport, autres).

Dans nos études sur les lexies d’affect, nous avons pu montrer la préférence de certaines lexies pour certaines rubriques journalistiques, ce qui rejoint l’idée de Plantin (2012) selon laquelle existent des « situations culturellement formatées pour une émotion » (p. 637). Par exemple, la lexie *jubilation* est davantage utilisée dans la rubrique ‘culture’ comme argument pour inciter le lecteur à passer à l’acte (ex. 22 ; Sorba & Novakova 2018). Il s’agit d’une argumentation « sur » l’émotion (Plantin 1997, voir *supra*) où l’émotion justifie l’action : le

texte argumentatif utilise l'émotion comme un déclencheur de « devoir éprouver » puis du « devoir faire » (Micheli 2010, p. 110, voir *supra*). En effet, en (22), le nom *jubilation* qui appartient au sous-type des affect causés, est employé dans cette rubrique 'culture' pour inciter le lecteur à assister aux spectacles du festival Passages :

22) La 10<sup>e</sup> édition du festival Passages (des théâtres à l'est de l'Europe et ailleurs) qui s'achève samedi fait un tabac. La **jubilation** ambiante rappelle les temps du défunt festival mondial de Nancy, où Tadeusz Kantor et Pina Bausch se faisaient connaître. Les vingt spectacles invités ce début mai par Passages ont fait salle comble. (*Libération* 11/05/2007)

Un autre exemple est celui de la lexie *déception* qui apparaît massivement dans la rubrique 'politique' (40%, ex. 23, Novakova & Sorba 2014a) :

23) Même au parti socialiste, elle régresse de 9%. C'est donc un mouvement de **déception** assez largement répandu. Seules les prochaines semaines diront si ce revers est un simple trou d'air ou le début d'un éventuel dévissage. (*Le Figaro* 03/02/2007).

Nous avons expliqué cette tendance en nous appuyant sur l'analyse sociologique de Lipovetsky (2006) selon lequel le monde moderne a développé des facteurs secrétant structurellement de la déception vis-à-vis du politique.

Dans une autre étude (Sorba & Novakova 2020), nous avons observé que la lexie *repentir* apparaissait de manière préférentielle, dans la rubrique 'justice' (32%, ex. 25) alors que *repentance* privilégiait la rubrique 'politique' (64%, ex. 24) :

24) Sous le regard de l'ONU, qui a voté un texte défendant les droits des peuples premiers il y a moins d'un an, cette **repentance** est toutefois menée à pas comptés, l'Australie et le Canada ne souhaitant pas ouvrir le coûteux dossier des indemnisations et moins encore celui conduisant éventuellement vers l'indépendance de certains territoires. (*Le Figaro* 11/06/2008)

25) Cette ligne de défense fondée sur le **repentir** et l'émotion n'a donc pas suffi – les criminels « doivent avoir conscience des conséquences de leurs actes », a souligné le juge –, et c'est probablement la plus grande star féminine de l'histoire de l'athlétisme, aujourd'hui âgée de 32 ans, qui va se retrouver derrière les barreaux. (*Le Monde* 13/01/2008)

L'analyse des rubriques a ici permis d'apporter des éléments concrets pour distinguer deux synonymes dans leurs emplois.

Un dernier exemple me permettra d'illustrer comment la rubrique agit comme « une détermination du sens (un *actualisateur* du signe linguistique) » pour reprendre les mots de Bouquet sur le genre (2004, p. 4). Le verbe *stupéfier*, sémantiquement de polarité neutre, revêt

une polarité tantôt positive (ex. 26) tantôt négative (ex. 27) en fonction de la rubrique dans laquelle il apparaît (Novakova & Sorba 2013a) :

- 26) Si l'art rupestre et les figures de pierre ne peuvent qu'être évoqués par la photographie, l'ensemble des bois sculptés est d'une remarquable richesse. Il y a là une quinzaine de pièces dont l'étrangeté des déformations anatomiques et des expressions **stupéfié**. Pour en suggérer la splendeur, il suffit de citer la provenance de quelques-unes de ces œuvres : elles ont appartenu à André Breton et à Tristan Tzara. (*Le Monde* 16/12/2008)
- 27) En Angleterre et au Pays de Galles – l'Écosse possède un système à part –, la justice est très dure envers les mineurs délinquants. Cette sévérité renvoie à un drame qui, en 1993, avait choqué le pays : l'assassinat du petit James Bulger, 2 ans, par deux enfants âgés de 10 ans, Jon Venables et Robert Thompson. L'image, captée par une caméra de surveillance, de l'un des deux assassins tenant la main du garçonnet et l'emmenant vers sa mort **avait stupéfié** et longtemps hanté les Britanniques. (*Le Monde* 04/07/2007)

Tendanciellement, nous avons pu observer que la rubrique 'culture' actualisait la polarité positive dans le sémantisme de la lexie alors que c'était plutôt la polarité négative dans les rubriques 'société' ou 'justice'.

L'approche thématique des rubriques que nous avons mise en œuvre pour comprendre l'impact du global (le générique) sur le local (l'emploi d'une lexie) est une première proposition. Il serait tout à fait possible de structurer davantage les niveaux hiérarchiques du genre et du sous-genre en suivant la modélisation proposée par Grosse sur les genres journalistiques (2001) : *discours* : journalistique → *champ générique* : presse écrite → *genre* : information, opinion, divertissement, conseils, etc. → *sous-genre* : rubrique (reportage, éditorial, critique, recette, jardinage, etc.). Cette proposition qui diffère légèrement concernant les niveaux hiérarchiques mériterait d'être mise à l'épreuve des corpus pour tester l'hypothèse d'une influence du global sur le local.

Même si nous nous inscrivons pleinement dans cette perspective de l'inclusion du genre au sein du discours, notre choix terminologique est différent. Nous avons choisi de privilégier l'expression *genre textuel* afin de bien montrer tout l'enjeu de cette catégorie qui est d'articuler des contraintes linguistiques et situationnelles.

### 3.2.3 Le genre : une articulation des contraintes linguistiques et situationnelles

Charaudeau & Maingueneau (2002, p. 280) soulignent à juste titre la variété terminologique qui traduit cette complexité de la question des genres : « genres de discours », « genres de textes », « types de textes », « types de discours », « hypergenre », « genres et sous-genres

situationnels », autant de termes qui mettent en œuvre des éléments de définition différents privilégiant tantôt l’ancrage social du discours et sa nature communicationnelle, tantôt les régularités compositionnelles des textes et leurs caractéristiques formelles. Biber (1993b) a même introduit une équivalence entre « genre » et « registre » dans l’une de ses définitions :

I use the terms genre or register to refer to situationally defined text catégories (such as fiction, sports broadcast, psychology articles), and text type to refer to linguistically defined text categories. Both of these text classification systems are valid, but they have different bases. Although registers/genres are not defined on linguistic grounds, there are statistically important linguistic differences among these categories [...]. In contrast, text types are identified on the basis of shared linguistic co-occurrence patterns, so that the texts within each type are maximally similar in their linguistic characteristics, while the different types are maximally distinct from one another. [...] register are based on criteria external to the corpus, while text types are based on internal criteria. (p. 244-245)

Biber distingue ici les « types de textes » (*text type*) qui relèvent de l’analyse linguistique et les « registres » ou « genres » (*registers/genres*) qui correspondent à une catégorisation sociale. Il s’agit pour lui de deux systèmes de classement bien distincts. Les genres sont alors des catégories de textes repérées spontanément par les locuteurs confirmés (*mature*) d’une langue (voir aussi sur ce point Biber 1989) sur des critères externes aux textes.

C’est précisément contre cette tradition anglo-saxonne de séparation entre « type de texte » et « genre » qu’Adam a construit sa « reconception linguistique du concept du genre » (1997) dans le champ de la linguistique textuelle. Il commence par réfuter la pertinence du concept de « type de texte », en argumentant sur la trop grande complexité et hétérogénéité de cet objet qui ne peut être réduit à un seul « type ». Pour rendre compte des régularités linguistiques, il choisit de se situer au niveau des séquences considérées comme « des unités compositionnelles supérieures à la phrase-période, mais très inférieures – mis à part le cas relativement rare des textes très courts mono-séquentiels – à l’unité globale que l’on peut appeler *texte*. » (*ibid.*, p. 665). Ce changement de niveau d’observation lui permet de distinguer divers plans d’organisation de la textualité, possédant assez d’autonomie pour être disjoints, tout en se combinant entre eux pour former un texte. La présence massive d’un certain type de séquences (narratif, argumentatif, explicatif, descriptif ou dialogal) permet de créer « un effet de dominante » pour caractériser le texte. Néanmoins, l’ensemble des séquences est soumis lui-même à un « plan de texte » qui détermine la structure globale d’un texte. Et ce sont les genres qui « règlent globalement, et de façon plus ou moins contraignante, les différents plans de la structure des textes. Ainsi, le genre de la fable détermine clairement le mode dominant de

composition (narratif), l'organisation sémantique du double monde représenté (la fiction peut mettre en scène des acteurs simplement anthropomorphes et un monde qui ne fait qu'indirectement allusion au monde du lecteur), le choix d'une valeur illocutoire (dénoncer, instruire et / ou distraire). » (*ibid.*, p. 671).

À l'instar d'Adam (2011), je considère donc le genre comme un « cadre », responsable d'opérations descendantes de sélection et d'agencement de différentes unités linguistiques. C'est précisément cette idée que Rastier défend dans la perspective de la sémantique textuelle : « c'est le texte qui détermine le sens des mots – à partir certes de leur signification en langue, mais en l'élaborant, en l'enrichissant et/ou la restreignant par l'action des normes génériques et situationnelles. » (Rastier 2005b, p. 20). Ainsi redéfini comme un concept sémantique, le genre permet de concilier les traditionnelles approches globale et locale de la description du sens : le palier de la globalité exerce une détermination sur le palier local des composants d'un énoncé (tradition rhétorico-herméneutique) tandis que le palier local permet de déterminer le palier global de l'énoncé (tradition logico-grammaticale).

Un des objectifs de recherche que je souhaite déployer dans les années à venir est donc de montrer comment l'interaction de ces deux paliers permet d'analyser de manière différentielle les genres textuels, un genre n'existant que par opposition à un autre genre. C'est dans cette perspective que j'avais lancé une première réflexion sur la comparaison de la combinatoire lexico-syntaxique de plusieurs lexies d'affect dans un corpus journalistique (EmoBase) et dans un corpus romanesque (PhraseoBase) en encadrant les mémoires de Tesi di Laurea<sup>62</sup> d'Alessia Albanese (étude des lexies *égoïsme, indifférence, confiance, empathie, anxiété, déception, compassion, frénésie, orgueil, sincérité*) et Agnese Verdina (étude des lexies *dégoût, sérénité, haine, panique, venger, surprendre, angoissé*). Leurs premiers résultats ont fait émerger quelques phénomènes linguistiques remarquables qu'il serait intéressant d'approfondir. Par exemple, dans le cas de la lexie *haine* en construction binominale, les collocations dans le corpus romanesque sont plutôt des noms à l'aspect inchoatif (*bouffée / vague de haine*) ou ponctuel (*éclat de haine*) alors que dans le corpus journalistique, les collocations indiquent de manière

---

<sup>62</sup> Ces deux étudiantes étaient inscrites dans le cadre de la double licence LLCER / *Lingue straniera moderna* (Université Savoie Mont-Blanc et Università del Piemonte Orientale, 2018). Ce travail faisait suite à de premières observations menées sur les noms généraux *sentiment* et *émotion* (Novakova, Sorba & Tutin 2018) où nous avons remarqué des différences de fonctionnement de la combinatoire lexico-syntaxiques de ces deux lexies entre un corpus journalistique (EmoBase) et un corpus de romans (Frantext).

privilegiée une verbalisation communicative de l'affect (*discours / message / insulte / prêcheur de haine*) et sa manifestation physique (*cri / expression de haine*).

C'est également cette piste contrastive que nous avons commencé à explorer pour la caractérisation du roman médiéval (Denoyelle & Sorba 2019a). Dans un corpus constitué de deux tomes de romans en prose du 13<sup>e</sup> siècle (*Lancelot*, tome 1, éd. Micha ; *Tristan*, tome 3, éd. Curtis), nous avons sélectionné des unités phraséologiques construites autour des verbes *aidier* et *doner*. Nous avons ensuite observé les comportements de celles-ci dans un corpus de contraste comportant deux œuvres en prose narrative (la chronique de Robert de Clari *La Conquête de Constantinople* et la chantefable *Aucassin et Nicolette*) et trois œuvres en vers (*Tristan* de Béroul, *Yvain, le Chevalier au lion* de Chrétien de Troyes, *La Chanson de Roland*)<sup>63</sup>. Nous avons pu constater l'influence du facteur générique, par exemple, dans l'emploi du marqueur discursif *si m'aïst Diex* : même si le verbe *aidier* apparaît massivement dans les paroles des personnages aussi bien dans le genre des chroniques que dans la prose narrative, cependant, dans les chroniques, la formule rituelle *si m'aïst Diex* n'est pas utilisée dans les paroles rapportées des personnages car ceux-ci ne se trouvent pas dans une situation nécessitant le renforcement de leur énoncé. Nous continuons actuellement d'augmenter la taille des corpus du 13<sup>e</sup> siècle pour enrichir nos observations.

Dans cette section 3.2, j'ai montré que même si le genre peut être défini comme une compétence cognitive partagée par l'ensemble des locuteurs, il n'en demeure pas moins « un concept aujourd'hui aussi multiforme qu'heuristique » (Bouquet 2004, p. 4). C'est le maillon qui assure le rattachement d'un texte à un discours (Rastier 1989<sup>64</sup>) : le genre se situe bien à l'articulation des contraintes linguistiques et situationnelles. C'est pour rendre compte de cette définition que j'ai choisi d'utiliser l'expression *genre textuel*.

Les différentes réponses apportées à la définition du genre, que j'ai présentées dans ce troisième chapitre, ont conduit à l'élaboration de diverses typologies. S'interroger sur quelques-unes de ces différentes typologies présente l'intérêt d'affiner la définition de la notion de genre car tout processus de classement repose sur l'identification préalable des propriétés des objets à classer. C'est désormais tout l'enjeu du chapitre 4.

---

<sup>63</sup> Ces œuvres sont extraites du corpus SRCMF (*Syntactic Reference Corpus of Medieval French*, voir sitographie).

<sup>64</sup> « Un discours s'articule en divers genres, qui correspondent à autant de pratiques sociales différenciées à l'intérieur d'un même champ. Si bien qu'un genre est ce qui rattache un texte à un discours. » (Rastier 1989, p. 40).

## Chapitre 4. La typologie des genres : une entreprise vouée à l'échec ?

L'un des principaux écueils relevés par Bronckart (1996) dans l'entreprise de typologie des genres est la circularité méthodologique de la démarche<sup>65</sup> : pour élaborer une conceptualisation des genres, on doit d'abord disposer d'une connaissance de ce qu'ils sont. On est donc contraint de procéder à un classement intuitif au préalable sur lequel on mène ensuite une étude empirique pour dégager les caractéristiques des genres, probablement celles-là mêmes qui ont présidé au classement initial. Todorov le notait déjà quand il écrivait que « La définition des genres sera donc un va-et-vient continu entre la description des faits et la théorie en son abstraction. » (Todorov 1970, p. 26).

Cet obstacle méthodologique n'a pas empêché la réflexion sur les genres de fleurir, en particulier celle issue de la réflexion poétique aristotélicienne<sup>66</sup> – et de s'égarer parfois. Cette division traditionnelle en « genres littéraires » a été progressivement diffusée, depuis le 19<sup>e</sup> siècle, par l'institution scolaire<sup>67</sup>. C'est par cette typologie particulière que nous allons ouvrir ce chapitre 4 (section 4.1) avant de nous interroger, au-delà des textes littéraires, sur l'intérêt de classer les productions textuelles (section 4.2) puis sur les solutions à mettre en œuvre pour y parvenir de manière satisfaisante (section 4.3).

---

<sup>65</sup> Habert et *al.* (1997) partagent également ce constat de la circularité des démarches typologiques habituelles qui cherchent « soit à caractériser les modes de production des textes (typologies situationnelles), soit à identifier les fonctions visées par les textes (typologies fonctionnelles). [...] L'hypothèse partagée par ces différentes recherches est que chacun des types postulés se caractérise par l'association d'un certain nombre de caractéristiques linguistiques. » (p. 28).

<sup>66</sup> Depuis Aristote, deux traditions génériques co-existent : la tradition poétique et la tradition rhétorique. Je n'étudierai pas ici la typologie des genres issue de la tradition rhétorique (délibératif, judiciaire, épideictique) dans la mesure où elle repose exclusivement sur une approche situationnelle du genre (la parole dans l'assemblée, au tribunal, dans le cadre de la cité).

<sup>67</sup> Chantal Massol (UGA, Litt&Arts UMR 5316) mène actuellement une réflexion sur cette question de poétique historique concernant l'émergence de la notion de « genre littéraire » et elle en a livré les premiers éléments lors d'une conférence inaugurale à la Journée d'étude que nous avons organisée à Grenoble le 15 octobre 2021 avec l'équipe de Lausanne dans le cadre du projet « Approches numériques des textes littéraires : vers une convergence méthodologique ? ». Selon elle, la notion de « genres littéraires » a progressivement émergé au cours du 19<sup>e</sup> siècle en tant que catégorie esthétique propre à la littérature : les constructions discursives qui la dotent d'une existence sont scolaires (au sens large du métadiscours littéraire) allant des revues de critique littéraire au manuel d'histoire littéraire.

#### 4.1 Le cas de la typologie des genres littéraires

La tradition poétique est à l'origine de la typologie des genres littéraires que Baroni juge comme « une catégorie opératoire incontournable » (2003, p. 141) et dont Compagnon (2000) rappelle la valeur heuristique. C'est une question ancienne qui ne fait néanmoins pas vraiment consensus, comme le rappelle Schaeffer (2007) :

La question des genres littéraires est aussi ancienne que l'étude de la littérature et pourtant on ne peut pas dire qu'à ce jour il y ait consensus quant à la nature exacte de ce que l'on entend par « genre littéraire. » (p. 357)

À l'origine, c'est une typologie à deux items qui intéresse Aristote : l'épopée (ἡ ἐποποιία) et la tragédie (ἡ τῆς τραγωδίας ποιήσις) peuvent être distinguées selon leur modalité énonciative (dans la première, le poète raconte ; dans l'autre, il imite)<sup>68</sup>. À l'époque romantique, Hegel fournit une triade (épopée, lyrisme, drame) en s'appuyant sur les critères de subjectivité (le lyrisme est la poésie subjective car elle exprime l'état d'âme du poète), d'objectivité (l'épopée est la poésie objective car un narrateur extérieur raconte des événements du monde extérieur) et de la synthèse des deux (le drame est la poésie objective-subjective car elle traite à la fois de l'intériorité des personnages et des événements du monde extérieur, voir Hegel 1979). Pour Schaeffer, ces catégories « ont le double désavantage d'être à la fois floues et ambiguës », ce qui rend le système triadique « boiteux » (1989, p. 38-39). Ces deux exemples parmi d'autres montrent que l'entreprise classificatoire des genres littéraires est une préoccupation ancienne sans doute « incontournable », mais comme Schaeffer le rappelle, le résultat obtenu, « loin d'atteindre une supposée auto-organisation interne de la littérature, n'est jamais qu'un découpage parmi d'autres. » (*ibid.*, p. 75).

Afin de mieux comprendre cette notion de « genre littéraire » et l'entreprise typologique qui y est afférente, je propose tout d'abord un aperçu historique de la conceptualisation de cette notion (4.1.1). Ensuite, j'examinerai la transformation permanente qui opère au sein des genres

---

<sup>68</sup> (1448a) ἔτι δὲ τούτων τρίτη διαφορὰ τὸ ὡς ἕκαστα τούτων μιμήσαιο ἂν τις. καὶ γὰρ ἐν τοῖς αὐτοῖς καὶ τὰ αὐτὰ μιμεῖσθαι ἔστιν ὅτε μὲν ἀπαγγέλλοντα, ἢ ἕτερόν τι γιγνόμενον ὥσπερ Ὅμηρος ποιεῖ ἢ ὡς τὸν αὐτὸν καὶ μὴ μεταβάλλοντα, ἢ πάντας ὡς πράττοντας καὶ ἐνεργοῦντας ἢ τοὺς μιμουμένους† (Aristotle. *Ars Poetica*. R. Kassel (éd.). Oxford : Clarendon Press. 1966). « Il y a encore parmi ces arts une troisième différence : la manière dont on peut imiter chacun de ces objets. Il est possible, en effet, d'imiter par les mêmes moyens les mêmes objets soit en racontant (que l'on se fasse autre comme le fait Homère ou que l'on demeure le même sans changement), soit en faisant de tous les personnages des êtres en action, c'est-à-dire des acteurs de l'imitation. » (trad. B. Gernez, Les Belles Lettres, 1997).

littéraires (4.1.2) et la possibilité de les définir comme une « réalité interdiscursive globale » (4.1.3).

#### **4.1.1 Aperçu historique de la conceptualisation du genre littéraire**

La réflexion autour des genres littéraires est liée à la figure culturelle de l'écrivain comme le note Scholes (1986) :

Le procès d'écriture est générique en ce sens que tout écrivain conçoit sa tâche en fonction de sa propre culture littéraire. [...] Tout écrivain inscrit son travail dans une tradition donnée, et l'on peut mesurer parfaitement ses réalisations dans les termes mêmes de la tradition où ce travail s'inscrit. Le plumeur ou le tâcheron – qu'ils écrivent des westerns télévisés en 1960 ou des romans élisabéthains en 1590 – tiennent pour un fait qu'ils se situent dans telle ou telle tradition et débitent des œuvres (?) [*sic*] par application de schémas formulaires. L'artiste de génie, pour sa part, enrichit la tradition d'une contribution nouvelle parce qu'il prend conscience de possibilités qu'elle contenait mais qui étaient restées jusque-là inaperçues, ou parce qu'il découvre de nouvelles manières de combiner des traditions antérieures, ou de nouvelles manières d'adapter une tradition à la situation changeante du monde qui l'entoure. (p. 78)

Ainsi, depuis Aristote, la façon d'organiser « la constitution d'une classe textuelle est fondamentalement un processus discontinu lié à une causalité externe [...] le texte n'existe que grâce à une causalité non textuelle : il existe parce qu'il est produit par un être humain. » (Schaeffer 1989, p. 71). Écrire est bien un processus réalisé par un humain inscrit dans une culture et celui-ci incarne cette « causalité non textuelle » à l'origine du texte. Regrouper des œuvres au sein d'un genre littéraire revient alors à former une classe avec des objets qui possèdent des caractéristiques communes en vertu de causes externes à la classe textuelle, à savoir, notamment, les intentions humaines des écrivains (p. 71-72).

Dans la citation ci-dessus, Scholes rend également compte de la tension qui habite l'inscription d'un texte littéraire dans un genre : le genre littéraire est utilisé soit comme un modèle à suivre par l'écrivain soit comme un espace de liberté créatrice. C'est exactement cette réflexion que périodise Sorlin (2015) quand elle rappelle les conceptualisations successives du genre littéraire.

Tout d'abord, jusqu'à la fin du 18<sup>e</sup> siècle, le genre se conçoit comme un ensemble de règles que les écrivains sont censés appliquer fidèlement : les genres incarnent des catégories abstraites permettant de distinguer les œuvres en fonction de leur thématique interne. À ce stade, le genre littéraire peut être défini comme un « critère de jugement littéraire » (Schaeffer 1989, p. 33), l'œuvre s'évaluant à l'aune de critères de réussite préétablis. Le bon auteur est alors

celui qui joue parfaitement le jeu des règles du genre littéraire dans lequel il écrit son texte et la créativité est conçue comme la reproduction de modèles idéaux du passé, le génie artistique requérant avant tout le respect des règles.

De la fin du 18<sup>e</sup> siècle à la fin du 19<sup>e</sup> siècle, la perspective change, en lien avec les préoccupations classificatoires dans le domaine de la biologie et la naissance du romantisme. Comme le souligne Beebee (1994, p. 2), le genre a alors une finalité interne : on étudie les raisons d'être des genres ; on quitte le domaine du normatif pour celui de l'explicatif pour établir une classification phylogénétique des genres. La théorie s'attache à décrire empiriquement les genres littéraires afin d'établir des groupes ou des familles historiques et les genres littéraires sont désormais présentés comme mus par un processus d'engendrement interne naturel, en dehors de toute emprise humaine et de tout facteur social. C'est l'approche biologique de Brunetière : « Quiconque écrit est [...] sous le poids, si je puis ainsi dire, de tous ceux qui l'ont précédé, n'importe ou non qu'il les connaisse » (Brunetière 1889, p. 317). L'influence principale qui fait évoluer la littérature est celle des œuvres sur les œuvres. Dans ce contexte, la créativité de l'écrivain consiste à s'affranchir des règles.

Enfin, au début du 20<sup>e</sup> siècle, en réaction aux taxinomies descriptives, les approches structuralistes ont cherché dans le texte lui-même les traits narratifs, linguistiques et institutionnels constitutifs d'un genre littéraire. Le genre retrouve ainsi le pouvoir heuristique qu'il avait perdu et devient un principe organisateur de notre expérience de la réalité littéraire. Le *Dictionnaire de critique littéraire* le définit effectivement comme une « catégorie littéraire qui permet de regrouper un certain nombre de textes, selon des critères variables. La classification générique, puisqu'elle consiste à déterminer les relations que les œuvres entretiennent entre elles, a trait à l'intertextualité. Un genre n'existe et ne se définit que par rapport aux autres. » (Gardes-Tamine & Hubert 1993, s.v.).

Le pluriel employé jusqu'à présent indique que la typologie des genres littéraires doit comporter au moins deux items. C'est le nombre auquel s'arrête Hamburger dans son ouvrage fondateur *Die Logik der Dichtung* (1977) en distinguant *die fiktionale oder mimetische Gattung* (le genre fictionnel ou genre mimétique) et *die lyrische Gattung* (le genre lyrique). Les deux groupes sont opposés sur la base de l'expérience de fiction (*das Erlebnis der Fiktion oder der nicht-Wirklichkeit*) ou l'expérience de réalité (*das Erlebnis der Wirklichkeit*) procurée au sujet par les œuvres<sup>69</sup>. À l'opposé, Charaudeau & Maingueneau (2002, s.v.) propose une liste ouverte

---

<sup>69</sup> Cette opposition à deux termes repose, pour Hamburger, sur la fonction du *Je* : le *Je* fictif du genre fictionnel s'oppose au *Je* réel du genre lyrique : « La fiction épique se définit par le fait que 1) elle ne comporte pas de Je-

de « genres littéraires », issue de la tradition critique littéraire qui utilise des critères très hétérogènes pour procéder à ce classement :

- Critères de composition, de forme, de contenu : poésie, théâtre, roman, essai, subdivisés eux-mêmes (ode, sonnet, épique, élégiaque ; tragédie, comédie, etc.) ;
- Manières de concevoir la représentation de la réalité (école, période historique) : romantique, réaliste, naturaliste, surréaliste, etc.
- Structures textuelles et organisation énonciative : fantastique, autobiographie, roman historique, etc.

À l'issue de ce bref aperçu historique sur la conceptualisation du genre littéraire, on ne peut que constater que les propositions typologiques sont pour le moins variées. Pour Schaeffer, ce résultat s'explique par un problème dans la conceptualisation de la problématique générique qui se révèle « irréductible à une problématique de la classification. » (1989, p. 172). Pour lui, une théorie générique ne peut pas décomposer la littérature en classes de textes mutuellement exclusives ni considérer la littérature comme « une totalité qui elle-même serait un superorganisme dont les différents genres seraient les organes. » (*ibid.*, p. 63). C'est bien la métaphore biologique du tout et de ses parties qui est ici rejetée dans la conceptualisation du genre littéraire et l'entreprise de typologie afférente.

Pour essayer de surmonter ce problème, nous pouvons repartir du programme de Jauss (1970), s'inspirant de celui de Droysen (1967), qui propose d'aborder « les caractères généraux des genres littéraires non plus d'un point de vue normatif (*ante rem*) ou classificateur (*post rem*), mais historique (*in re*), c'est-à-dire dans une “continuité, où tout ce qui est antérieur s'élargit et se complète par ce qui suit” ». (Jauss 1970, p. 43). C'est alors concevoir les genres littéraires comme des entités en transformation permanente.

---

origine réel, 2) elle doit comporter des Je-Origines fictifs, c'est-à-dire des systèmes de référence cognitivement [...] sans rapport avec un Je réel. » (Hamburger 1986, p. 82) ; « *le sujet de l'expérience, et, avec lui, le sujet d'énonciation, le Je lyrique, ne peut être que réel.* » (p. 243). L'analyse des indices linguistiques (temps verbaux, déictiques spatio-temporels) appuie cette proposition théorique puisque Hamburger propose ici une « théorie linguistique de la littérature » (*die Sprachtheorie der Dichtung*) : « *La logique du littéraire, en tant que théorie linguistique de la littérature, a pour objet la relation de la littérature à l'ensemble du système de la langue.* » (*ibid.*, p. 22). La logique de la littérature ici décrite « examine l'existence et la nature des différences *fonctionnelles* entre le langage producteur de formes littéraires [...] et celui qu'on trouve à la base des activités de pensée et de communication. » (*ibid.*).

#### 4.1.2 Les genres littéraires en transformation permanente

Le programme de Jauss est de proposer, dans une double perspective diachronique et synchronique, « une histoire structurale des genres littéraires » (1970, p. 47). Pour classer les genres littéraires, il convient donc de spécifier « des groupes, dans la mesure où [un genre littéraire] réussit de façon autonome à constituer des textes, cette constitution devant être saisie aussi bien synchroniquement dans une structure d'éléments non interchangeables que diachroniquement dans une continuité qui se maintient. » (p. 45). Les critères pour délimiter un genre littéraire ne seraient ni formels ni thématiques exclusivement. Il s'agit pour Jauss de déterminer « un ensemble de caractéristiques et de procédés dont quelques-uns dominant et peuvent être décrits dans leur fonction, indice d'un système » (p. 46). Les genres littéraires sont ici conceptualisés dans une perspective structuraliste et leur histoire est conçue comme un phénomène dynamique soit un « processus temporel de l'établissement et de la modification continus d'un horizon d'attente » (p. 57). À la même époque, Todorov développe une réflexion similaire sur l'inscription du genre littéraire dans un système en mouvement :

[...] toute œuvre modifie l'ensemble des possibles, chaque nouvel exemple change l'espèce. [...Un texte] manifeste des propriétés qui lui sont communes avec l'ensemble des textes littéraires, ou avec un des sous-ensembles de la littérature [...] il n'est pas seulement le produit d'une combinatoire préexistante [...] ; il est aussi une transformation de cette combinatoire. (Todorov 1970, p. 10-11).

Ce sont les écrivains qui travaillent sans cesse le genre littéraire devenu ainsi « l'objet d'une série de processus d'accommodations et d'appropriations » (Baroni & Macé 2007, p. 15). L'une des conséquences de ce travail permanent du genre littéraire par les écrivains est la difficulté de maintenir un ensemble de propriétés communes stables sur lesquelles appuyer une typologie et donc une terminologie stable<sup>70</sup>.

C'est précisément cette conception historique du genre littéraire que je souhaite mettre à l'épreuve dans l'un des projets que je dirige sur le genre romanesque (IRGA Phraséo 13-18, voir annexe IV). Nous constituons actuellement un corpus de textes littéraires regroupés par la tradition sous l'étiquette 'romans de chevalerie' en partant des premiers exemplaires de ce

---

<sup>70</sup> C'est le constat déjà fait par Koelb (1975) à propos de la tragédie : « By cultural accident, the same word [tragedy] has been used to name many different sets of objects. [...] words, we must not forget, are social conventions, and such conventions change. "Tragedy" has meant many different things in its history, which should not surprise us. » (p. 258 et 261). Plus globalement, pour Schaeffer (1989), les noms génériques « ne se réfèrent que rarement à un ensemble de caractéristiques ou de propriétés textuelles récurrentes. » (p. 66).

« genre littéraire » en prose au 13<sup>e</sup> siècle pour aller jusqu’au 17<sup>e</sup> siècle<sup>71</sup>. Notre perspective étant linguistique et stylistique, nous avons partitionné le corpus en trois ensembles correspondant aux trois états de la langue française : ancien français (AF, 13<sup>e</sup> s.), moyen français (MF 1350-1550) et français préclassique (1550-1650)<sup>72</sup>.

Le genre romanesque a été choisi en raison de son caractère central en Occident : « Si l’importance d’une invention peut se mesurer au nombre de ceux qui en revendiquent la paternité, le roman est sans contredit une innovation de premier plan. » (Gingras 2017, p. 15). De la *lingua romana* qui désignait, au 9<sup>e</sup> siècle, « le latin abâtardi couramment parlé sur les territoires occidentaux » (Coulet 1991, p. 19), de laquelle naquirent les langues romanes, on glisse progressivement vers le terme *romanz* qui désigne, au 12<sup>e</sup> siècle, une langue vivante s’opposant au latin, puis le récit fait directement dans cette langue. Sous cette étiquette de ‘genre romanesque’ sont ainsi regroupées des œuvres, généralement en prose, constituées par un récit, dont les personnages fictifs sont donnés comme réels.

La perspective diachronique adoptée vise à renouveler la caractérisation de ce genre littéraire<sup>73</sup>. L’enjeu est ainsi de caractériser le genre non pas en termes de macrostructure, d’univers de référence ou de style individuel des auteurs, comme c’est souvent le cas dans les théories traditionnelles des genres littéraires, mais à travers des constructions lexico-syntaxiques (CLS) récurrentes, statistiquement spécifiques à ce genre. En effet, les recherches foisonnantes en phraséologie depuis une trentaine d’années permettent désormais d’aborder de manière plus fructueuse ce phénomène en diachronie. Buridant constatait, dès 1989, sa complexité et de ce fait, la part restreinte qu’il occupait dans les recherches d’alors. Une étude

---

<sup>71</sup> Les œuvres choisies sont des textes originaux ou des traductions de textes originaux. Nous les avons intégrées à notre corpus en totalité ou en partie seulement selon les principes de constitution que nous avons retenus pour assurer une bonne représentativité de la production littéraire de chaque période. Sur ce point, voir le chapitre 5.

<sup>72</sup> Nous suivons la répartition proposée par Combettes (2003).

<sup>73</sup> « Dans le cas de la littérature médiévale, la question du genre est complexe. En effet, les distinctions génériques des auteurs du Moyen Âge sont confuses, souvent incohérentes ou embryonnaires. La plus célèbre d’entre elles, le classement que fait Jean Bodel au début de *La Chanson des Saisnes* (v. 6-12) des trois matières dont sont faits les textes (matière de France, matière de Bretagne, matière de Rome), regroupe des objets (chansons de geste, romans antiques, romans arthuriens) dont les statuts sont très différents. Comment s’y retrouver entre *lais*, *dits*, *fables*, *contes*, ou *romans* et *histoires* qui, selon les manuscrits, désignent indifféremment les mêmes textes ? La plupart de nos catégorisations modernes aboutissent à des distinctions perméables peinant à classer une multitude de textes que nous qualifions d’*hybrides*, de *parodiques*, d’*intermédiaires* tant ils ont du mal à entrer dans des catégories construites *a posteriori*. » (Denoyelle & Sorba 2020, p. 2).

pilote (Denoyelle & Sorba 2020) a permis de montrer, par exemple, que les trois unités phraséologiques construites autour du verbe *doner* (*doner cop/conseil/congié*) jouaient un rôle structurant dans le roman de chevalerie du 13<sup>e</sup> siècle : *doner cop* forme un noyau narratif autour duquel s'organise le récit du combat singulier ; *doner conseil/congié* organise des séquences argumentatives révélatrices « des enjeux sociaux (permission d'agir) et moraux (conseils tactiques et religieux) » des cadres hiérarchiques de la société médiévale (*ibid.*, p. 13-14). L'idée serait désormais d'observer le comportement linguistique de ces trois unités phraséologiques dans le reste du corpus puisque notre recherche pose l'hypothèse d'une continuité sur le plan linguistique dans la fiction romanesque entre le 13<sup>e</sup> et le 17<sup>e</sup> siècle malgré la rupture thématique totale qui se produit au 16<sup>e</sup> siècle (Gingras 2015) et la variation de la langue entre ancien français (AF), moyen français (MF) et français préclassique (FPC). Pour ce faire, d'autres observables phraséologiques sont également envisagés :

- les CLS de l'oral représenté (ex. les marqueurs discursifs *sachiez que, il m'est avis*) ;
- les CLS de l'expression des affects (ex. *(se) mettre en + N affect, faire du + N affect*) ;
- les CLS relevant de la cognition (l'emploi des verbes de cognition désignant des processus intérieurs du type *penser, réfléchir, croire, sentir, espérer*, etc., est une nécessité pour le romancier dans la construction de sa fiction, voir Hamburger 1986, p.72-73) ;
- les CLS dénotant des interactions de groupes (la relation de l'individu à la communauté par rapport à laquelle il se situe, problématique fondamentale du roman depuis ses origines, se trouve dans les CLS exprimant les rencontres, les séparations, les accords ou les refus).

Nous proposons ainsi de montrer que l'entrée par les unités phraséologiques caractéristiques d'un genre littéraire est un critère pertinent pour saisir les propriétés du genre romanesque et modéliser son évolution en diachronie. Ainsi, une typologie des genres littéraires doit rendre compte de la transformation permanente qui atteint ceux-ci afin de les concevoir comme une « réalité interdiscursive globale ».

#### **4.1.3. Les genres littéraires comme « réalité interdiscursive globale »**

La formule de Letourneux (2016, § 13) choisie pour intituler cette section soutient l'idée que les genres littéraires ne sont finalement que des genres textuels parmi d'autres : « les

productions littéraires ne sont que du discours et n'existent pas hors de celui-ci. » (§ 2)<sup>74</sup>. En effet, pour reprendre les propos généralisants de Rastier, « tous les textes relèvent d'une pratique sociale, dont témoigne leur genre, mais aussi d'une action individuelle et d'un projet. » (2011, p. 27). Pour examiner les propriétés des genres littéraires, l'analyste doit donc abandonner les hiérarchies qui fondent le discours littéraire en se séparant de l'axiologie afférente à la notion d'œuvre, sous peine d'anachronisme<sup>75</sup>. L'objectif de Letourneux, qu'il indique clairement dans le titre de son article, est bien de définir le genre (littéraire) comme une « pratique historique » :

[...] il n'y a d'auteur se situant dans un genre que dans la mesure où il se fait relecteur d'un héritage. Cet héritage n'est pas une réalité transcendante, mais une construction située, celle d'une réception définie par la position du locuteur (auteur, lecteur, critique, éditeur, journaliste, publicitaire, etc.), par sa façon de dialoguer avec les discours sur le genre (et en particulier ceux avec lesquels il est appelé à interagir directement – les séries génériques, médiatiques, discursives, culturelles, etc. qu'il investit), par sa relation aux héritages, et par son expérience du genre (des œuvres et des définitions). (Letourneux 2016, § 11)

Le genre (littéraire) peut être considéré comme une « pratique historique » au sens où c'est une pratique située dans un cadre socio-historique, résultant d'une « négociation » entre différents acteurs (*ibid.*, § 13) : l'auteur qui s'empare d'un héritage et les instances de réception qui s'emparent de l'œuvre. Les genres littéraires sont bien « les supports d'opérations accomplies par les acteurs de la vie littéraire, et susceptibles de remplir une grande variété de fonctions : esthétique, herméneutique, cognitive, affective, politique. [...] Les genres littéraires représentent ainsi] la part de "généralité" et de "connu" qui assure la communicabilité du texte à son lecteur, ou son inscription dans l'histoire des formes. » (Baroni & Macé 2007, p. 8-9).

On pourrait tout à fait étendre cette description du phénomène au champ non littéraire dans la mesure où l'on passe son temps, dans la parole quotidienne, « à rapporter des textes à des séries architextuelles<sup>76</sup>, contribuant dans ses conversations courantes à imposer des définitions

---

<sup>74</sup> Sur ce point, voir aussi Adam & Heidmann (2007) pour qui les genres littéraires ne sont qu'un cas parmi d'autres de « système de genres » d'une société donnée.

<sup>75</sup> Letourneux (2016) : « Paul Veyne veut souligner que l'Histoire littéraire est dominée par la question de la valeur des œuvres [...] hiérarchies définies *a posteriori*, depuis le présent de la réception [...] l'Histoire littéraire axiologique est, fondamentalement, une Histoire anachronique. » (§ 5).

<sup>76</sup> La notion d'*architexte* a été développée par Genette qui la définit comme « l'ensemble des catégories générales, ou transcendantes – types de discours, modes d'énonciation, genres littéraires, etc. – dont relève chaque texte singulier. » (Genette 1982, p. 7).

intuitives des genres » (Letourneux 2016, § 13). Par exemple, le sentiment du « romanesque », « ne s'appuie pas sur une capacité à définir le genre, mais sur une certaine interface entre la littérature et la vie, un condensé d'affect, de valeurs et de mémoire qui nourrit la pratique du roman. » (Baroni & Macé 2007, p. 10). C'est parce que le discours collectif quotidien joue un rôle majeur dans l'identification des genres, leur nomination et leur institution que Letourneux propose de définir les genres comme « une réalité interdiscursive globale ».

Deux conséquences peuvent être tirées de ces observations : premièrement, si le genre est une « pratique historique », on peut donc en faire l'histoire. Cette conception permet d'étayer le projet d'une étude diachronique des genres textuels. Deuxièmement, si les genres se caractérisent comme « une réalité interdiscursive globale » résultant d'une « négociation », alors chacun d'entre eux se trouve pris dans un réseau de discours en interaction et en tension dont les échanges et les nécessaires contradictions qui en sont issues risquent de rendre l'unité d'ensemble problématique.

Ce détour par le cas des genres littéraires et par leurs typologies était nécessaire pour montrer que les genres littéraires sont des genres textuels parmi d'autres et que l'entreprise typologique, même si elle se révèle problématique, doit être considérée, dans une perspective dynamique, comme une « pratique historique » résultant d'une construction interdiscursive. Je vais maintenant interroger les raisons qui conduisent à classer les productions textuelles d'une société.

#### **4.2 Pourquoi classer les productions textuelles ?**

À cette question, Ryan (1981) répondait de la manière suivante : « The significance of generic categories thus resides in their cognitive and cultural value, and the purpose of genre theory is to lay out the implicit knowledge of the users of genres. » (p. 112). L'idée est bien de faire émerger le savoir empirique que les locuteurs ont déjà construit sur la base de leur intuition<sup>77</sup>. Ce savoir empirique correspond à ce que Schaeffer (1989) nomme les « logiques de la généricité ». À la suite d'Adam & Heidmann (2007), j'entends la *généricité* comme « la mise en relation d'un texte avec des catégories génériques ouvertes. [...] Un texte n'appartient pas en soi à un genre, mais il est mis, à la production comme à la réception-interprétation, en relation à un ou plusieurs genres. » (p. 62). Ces logiques sont plurielles car elles diffèrent selon que le

---

<sup>77</sup> Baroni (2003) rend compte de ce phénomène en définissant les genres littéraires comme des « constructions sémantiques produites par l'expérience textuelle des lecteurs. » (p. 141-142). Cette définition s'applique bien, à mon sens, à l'ensemble des genres textuels.

critère choisi est « l'exemplification d'une propriété, l'application d'une règle, l'existence d'une relation généalogique ou celle d'une relation analogique. » (Schaeffer 1989, p. 180). Je propose donc d'envisager, dans la présente section, trois réponses à cette question : on classe les productions textuelles en faisant émerger les logiques de la généricité pour accéder à une unité structurante (4.2.1), à la variation linguistique (4.2.2) et au changement linguistique (4.2.3).

#### 4.2.1 Classer pour accéder à une unité structurante

Le genre qui agit comme un « cadre », pour reprendre le mot d'Adam (2011), se présente comme une unité structurante :

[Les genres] permettent de socialiser la parole individuelle en la coulant dans des formes entérinées et répertoriées qui déterminent l'horizon d'attente. (Amossy 2006, p. 217).

[...] général mais non universel, objet d'accommodations permanentes tissant des liens entre des individus qui l'adaptent à leur tour, élément d'un système de relations et objet d'un archivage au long cours, le genre est une entité floue mais fortement structurante. (Baroni & Macé 2007, p. 16)

Cette structuration agit à un double niveau : celui de la production et celui de la réception qui réalisent toutes deux la mise en œuvre du genre comme pratique historique (« archivage au long cours », « formes entérinées et répertoriées ») et objet de négociations (« objet d'accommodations permanentes tissant des liens entre des individus »). C'est ce qui permet de considérer le genre, à l'instar de Bronckart (1996), comme un mécanisme d'apprentissage social parmi d'autres<sup>78</sup>, qui fait, à ce titre, l'objet de détournements, de modifications et de réorganisations.

Cette conception implique nécessairement de concevoir le genre comme « la régularité au sein de l'interdiscours » (Baroni 2009, p. 4). Cette régularité permet de situer les genres textuels comme éléments d'un système sur lequel les acteurs en présence s'appuient à la fois en production et en réception<sup>79</sup>. Côté production, l'auteur ou l'éditeur peuvent fournir des

---

<sup>78</sup> Sur ce point, voir aussi Vaugeois (2007) : « la généricité ne se prouve pas, elle se *reconnaît*. La généricité ne se décrète pas, elle a besoin d'une confirmation intersubjective [...]. Un genre est bien une réalité conventionnelle au sens juridique de ce qui est établi par accord. » (p. 217).

<sup>79</sup> Pour Baroni, le fait que le genre permette de produire et de recevoir des faits de langue est intrinsèquement lié à la mise en œuvre d'une typologie : « Il faudrait aussi se rappeler, sur un plan pragmatique, que le "genre" permet non seulement de produire et d'interpréter des faits de langue, mais qu'il permet aussi de les "classer", c'est-à-dire de les situer réciproquement au sein du "champ" de l'interdiscours : le genre n'est donc pas seulement une régularité isolée, mais une régularité qui s'intègre à un "champ" des genres et aux "tensions" spécifiques qui le

indications d'ordre générique sur la couverture d'un ouvrage (par ex. le titre « Poèmes barbares » du recueil de Leconte de Lisle ou l'intitulé de la série « Essais » dans la collection « Points » des Éditions du Seuil) ou dans les intitulés de rubriques journalistiques (par ex., les rubriques culturelles « La Boîte aux Images », « Le coin-coin des Variétés » et « Le Théâtre » de l'hebdomadaire *Le Canard enchaîné*). Ces indications permettent ainsi d'orienter la lecture<sup>80</sup> côté réception, pour aider ou au contraire induire en erreur, selon le projet initial de l'auteur ou du locuteur (par ex., la prétendue authenticité des *Lettres et billets galants* de Mme de Villedieu<sup>81</sup> ou encore le détournement du genre musical de la pastorale dans les *Machines agricoles* du compositeur Darius Milhaud<sup>82</sup>).

En outre, c'est bien ce caractère structurant du genre qui permet de réaliser des détournements, des adaptations, des modifications et c'est bien grâce à lui également que ces détournements, ces adaptations, ces modifications peuvent être pensées et perçues comme telles<sup>83</sup> :

On aura beau encanailler la littérature et les arts contemporains, les métisser, tenter des échappées en-dehors de la « zone de juridiction » des genres traditionnels, le genre reviendra sans doute toujours, comme une empreinte indélébile, un signe de reconnaissance, bref : une hypothèse de conception et de réception. (Dion, Fortier & Haghebaert 2001, p. 361)

---

traversent en fonction de la situation qu'il occupe. Un genre possède ainsi une dimension de "systématicité", il appartient à un "système des genres". » (2009, § 22).

<sup>80</sup> Ce serait même selon Baroni la fonction essentielle des genres littéraires : « l'acte qui consiste à rattacher une œuvre à un genre littéraire spécifique a pour fonction essentielle de soutenir et d'orienter l'acte de lecture. » (2003, p. 143).

<sup>81</sup> Les *Lettres et billets galants* sont un recueil de lettres d'amour, publié en 1668, par Marie-Catherine Desjardins, plus connue sous son nom de romancière Mme de Villedieu. L'un de mes étudiants de Master, M. Guillaume Ferachoglou-Mollard, a précisément travaillé sur ce corpus dans le cadre de son mémoire (*L'expression de la plainte dans les Lettres et billets galants de Mme de Villedieu*, 2018) et a pu conclure, dans son état de l'art, à une mise en scène élaborée pour faire passer pour d'authentiques lettres ce qui allait devenir par la suite un nouveau genre, le roman épistolaire.

<sup>82</sup> Les *Machines agricoles*, œuvre de Darius Milhaud composée en 1919, sont répertoriées comme « pastorales pour voix et ensemble de chambre » (opus 56). Le texte mis en voix détourne néanmoins fortement des attendus du genre car il s'agit de descriptifs de machines agricoles tirés d'un catalogue commercial (*La Moissonneuse, La Faucheuse, La Lieuse, La Faneuse, La Fouilleuse draineuse, La Déchaumeuse-semeuse-enfouisseuse*). Les machines agricoles font certes partie du paysage rural du début du 20<sup>e</sup> siècle mais sont bien loin des thèmes bucoliques et champêtres traditionnels des autres œuvres musicales rangées sous cette étiquette générique.

<sup>83</sup> Baroni formule cette idée sous la forme d'une tension au sein du genre « entre principe centrifuge de variation et principe centripète d'identité » (2009, § 7).

Néanmoins, cette « hypothèse de conception et de réception », le genre, reste propre à une culture comme le soulignait déjà Ryan (1981) : « genres have a social and cognitive reality [...] different cultures make indeed use of widely differing systems of genres. » (p. 113). Cette notion n'a donc pas de caractère universel. Classifier les productions textuelles permet finalement de rendre compte des caractéristiques discursives d'une culture.

#### 4.2.2 Classifier pour accéder à la variation linguistique

À l'intérieur d'une même culture, classer les productions textuelles ouvre également un accès à la variation linguistique, c'est-à-dire au phénomène par lequel une langue déterminée n'est jamais à une époque, dans un lieu et dans un groupe social donnés, identique à ce qu'elle est dans un autre lieu et dans un autre groupe social :

Very often a change in register will affect linguistic parameters such as frequency distributions and co-occurrence patterns [*sic*], and, conversely a change in the latter is likely to reflect a change in the former. (Tognini-Bonelli 2001, p. 61)

[...] l'étude des corpus montre que le lexique, la morphosyntaxe, la manière dont se posent les problèmes sémantiques de l'ambiguïté et de l'implicité, tout cela varie avec les genres, les champs génériques et les discours. [...] Même au palier du mot, les variations selon les discours et les genres sont fort importantes, tant au plan sémantique qu'au plan de l'expression. (Rastier 2005a, p. 35-36)

Les études menées depuis une vingtaine d'années dans le champ de la linguistique de corpus révèlent que l'appartenance d'un texte à un genre conditionne les variations lexicales, morphosyntaxiques et discursives qui s'y trouvent en comparaison avec d'autres genres (voir par ex. Malrieu & Rastier 2001 ; Stubbs & Barth 2003). Les analyses pionnières de Biber (1993a) dans ce domaine ont montré que les genres textuels (*registers*<sup>84</sup>) se différenciaient, de manière systématique et conséquente, sur les plans grammaticaux et lexicaux. Par exemple, sa description linguistique des emplois de trois types de propositions subordonnées en anglais (*relative clause, causative adverbial subordinate clause, that complement clause*) a montré de fortes variations dans leur fréquence entre les quatre genres textuels de son corpus (*press reports, official documents, conversations, prepared speeches*). Sur le plan lexical, son analyse des adjectifs relevant du champ de la 'certitude' (*certain, sure, definite*) révèle que la répartition

---

<sup>84</sup> Biber définit le terme *register* comme « the range of text categories » (1993a, p. 219) que je choisis de nommer « genre textuel ».

de ces trois lexies divergent entre un corpus littéraire (*fiction*) et un corpus académique (*social sciences*)<sup>85</sup>.

Dans deux de mes études portant sur le grec ancien (Sorba 2008b ; 2015a), j'ai pu émettre l'hypothèse de l'influence du genre sur la variation paradigmatique dans les formules nom-épithète comprenant un des noms de la mer. Dans ses tragédies, Eschyle innove par rapport au modèle épique homérique dans le choix des épithètes : ἀκύμων devient épithète de πόντος (*Agamemnon* 565, « la mer sans vague ») et ὑγρός celui de θάλασσα (*Suppliantes* 259, « la mer humide »). Il innove également dans l'emploi des adjectifs dérivés des noms de la mer. Par exemple, au sein des épopées homériques, l'adjectif ἄλιος est spécialisé dans l'expression de l'origine-filiation : il se rencontre dans les syntagmes désignant ὁ γέρον ἄλιος « le Vieux de la mer » et les différentes formules désignant ses filles, les Néréides<sup>86</sup>. C'est à partir de ces observations que Perpillou (1973, p. 248) a pu conclure que l'adjectif ἄλιος n'est pas utilisé « pour ceux qui vivent de la mer ou travaillent sur la mer, car ils lui sont extérieurs, et loin d'en être issus, ils y vont : ἄλιος exprime une dépendance à l'égard de la mer, origine, possession, partie d'un tout ». Dans les tragédies d'Eschyle, j'ai pu constater une diversification des emplois quand ἄλιος qualifie des noms non humains, τὸ κῦμα « la vague » (*Suppliantes* 14) et ὁ πρῶν « le cap » (*Perses* 131 ; 881), sans néanmoins remettre en cause l'analyse sémantique de Perpillou. L'influence du genre (et de sa métrique particulière) est l'une des hypothèses qui pourrait expliquer cette variation mais elle demeure fragile en raison de l'état fragmentaire du corpus. Pour être confortée, cette hypothèse mériterait d'être testée sur l'ensemble des textes disponibles pour l'état de langue du grec archaïque.

En bref, la variation linguistique est ce qui permet à un genre de se distinguer d'un autre et, de ce fait, motive, en partie, la nécessité d'une typologie des différentes productions textuelles.

---

<sup>85</sup> « Further, there are striking differences across written registers in the use of these words. In social science, *certain* is quite common, *sure* is relatively rare, and *definite* is common relative to its frequency in the whole written corpus. Fiction shows the opposite pattern: *certain* is relatively rare, *sure* is relatively common, and *definite* is quite rare. These patterns alone show that the semantic domain of certainty in English could not be adequately described without considering the patterns in complementary registers. » (Biber 1993a, p. 227).

<sup>86</sup> Pour les Néréides, voir *Iliade* 18.86 ; *Odyssée* 24.47, 24.55 (ἀθανάτησ' ἄλιησιν) ; *Iliade* 18.139 (ἀλίησι κασιγνήτησι) ; *Iliade* 18.432 (ἀλλάων ἀλιάων) ; *Iliade* 24.84 (ἄλαι θεαί). Pour les vieux de la mer, voir *Iliade* 1.538, 1.556, 18.141, 20.107, 24.562 ; *Odyssée* 4.349, 4.365, 4.384, 4.401, 4.542, 13.345, 24.58 (Nérée) ; *Odyssée* 17.140 (Protée) ; *Odyssée* 13.96 (Phorkys).

### 4.2.3 Classer pour accéder au changement linguistique

Le troisième et dernier argument pour justifier l'intérêt de classer les productions textuelles prend place dans une perspective diachronique. Il découle du premier constat que j'ai fait plus haut selon lequel classer les productions textuelles permet de rendre compte des caractéristiques discursives d'une culture. Ainsi, pour reprendre les mots de Maingueneau (2009, s.v.), le genre peut être considéré comme « un accès privilégié au changement social. Une société peut en effet être caractérisée par les genres de discours qu'elle rend possibles et qui la rendent possible ». Celui-ci prend comme exemple le développement de la démocratie politique sous la Révolution française qui a obligé les hommes politiques à inventer les genres de discours de la vie parlementaire, ou encore l'apparition d'internet qui est inséparable de nouveaux genres : chats, forum, courriels, ou encore la discussion Wikipédia (Poudat 2020).

En effet, depuis les années 1990, les recherches sur les changements des pratiques discursives s'intéressent aux phénomènes génériques qui sont appréhendés comme « des *institutions de paroles* articulant une identité énonciative avec un lieu social ou avec une communauté de locuteurs » (Chauradeau & Maingueneau 2002, s.v.). Cette approche se situe tout à fait dans la lignée des travaux de Jausse en plaçant le récepteur au centre de la démarche avec son « horizon d'attente ». Elle a permis d'engager une réflexion critique sur le caractère historique des typologies c'est-à-dire de décrire l'émergence de nouvelles catégories de genre et la mutation des anciennes<sup>87</sup>. C'est exactement dans cette perspective que je situe ma recherche actuelle.

Dans le cadre de mes récents travaux sur le genre épistolaire, réalisés en partie dans le cadre de la thèse d'Iris Fabry actuellement en cours, nous avons constitué un corpus de lettres familières pour moitié authentiques et pour moitié issues de manuels épistolographiques allant du 17<sup>e</sup> au 20<sup>e</sup> siècle. Notre première étude portant sur les formules de clôture dans un échantillon du corpus de lettres authentiques<sup>88</sup> (Fabry, Lignereux & Sorba 2021) nous a permis d'observer

---

<sup>87</sup> Par exemple, certains comme Baroni & Macé (2007, p. 12) se demandent, dans le cadre de leur réflexion sur les genres littéraires, en quoi la transformation de ces compétences génériques participe à l'histoire d'un genre sachant que celles-ci font partie des compétences présumées par l'oeuvre pour assurer son intelligibilité.

<sup>88</sup> Le corpus pour cette première étude est constitué des lettres authentiques suivantes : pour le 17<sup>e</sup> s., Mme de Maintenon / Mme de Dangeau, Roger de Bussy-Rabutin / Mme de Scudéry, Nicolas Boileau / Jean Racine ; pour le 18<sup>e</sup> s., Mme de Livry, Rose de Staal-Delaunay / la Duchesse du Maine, Julie de Lespinasse, Jean-Jacques Rousseau ; pour le 19<sup>e</sup> s., Marceline Desbordes-Valmore, Zulma Carraud / Honoré de Balzac, Stendhal / amis ;

des emplois différents des patrons lexico-syntaxiques selon les époques. En effet, dans les échanges épistolaires du 18<sup>e</sup> siècle, le patron [X interpellatif sans déterminant possessif] domine (ex. 28) tandis qu'au 19<sup>e</sup> siècle, ce sont plutôt les patrons [X nom propre avec déterminant possessif] (ex. 29) et [X relationnel avec déterminant possessif] (ex. 30) :

28) Je suis bien aise, **chère Présidente**, que vous preniez le parti de rester à Toulouse.  
(XVIII\_Livry\_amitié-femme)

29) J'ai tant de fièvre, **ma chère Caroline**, que je retourne à ma silencieuse amitié après vous avoir serré tendrement la main. (XIX\_DesbordesValmore\_amitié-femme)

30) Je suis forcé, **mon cher ami**, de m'arrêter ici, à cause de la poste. (XIX\_Stendhal\_amitié-homme)

L'analyse des formules de clôture dans cet échantillon a permis de constater une évolution dans les tendances d'emploi des patrons lexico-syntaxiques. C'est un indice parmi d'autres que nous commençons à exploiter pour montrer la mutation du genre épistolaire au cours de ces quatre siècles. Cette mutation est intrinsèquement liée à l'évolution des pratiques sociales auxquelles le genre de la lettre familière est associé. C'est bien le programme déjà ébauché par Rastier en 1989 :

L'origine des genres se trouve donc dans la différenciation des pratiques sociales. Et il ne suffit pas de dire, avec Todorov, que nos genres sont issus de ceux qui les précédaient ; il faudrait encore montrer comment les genres se forment, évoluent et tendent à disparaître avec les pratiques sociales auxquelles ils sont associés. (Rastier 1989, p. 40)

Finalement, pour répondre à la question de savoir pourquoi classer les productions textuelles, j'ai exploré trois pistes dans cette section 4.2. Tout d'abord, l'établissement d'une typologie des genres textuels permet d'accéder à une unité structurante dans un système afin d'accéder aux caractéristiques discursives d'une culture à une époque donnée. Dans cette perspective synchronique, une typologie des genres permet également d'observer les phénomènes de variation linguistique. Si l'on adopte une approche diachronique, l'utilité d'une typologie peut être d'accéder au changement linguistique afin de modéliser l'émergence de nouvelles classes et la mutation (voire la disparition) des anciennes<sup>89</sup>. Répondre à la question du « pourquoi » était un préalable avant de s'interroger sur la manière de classer les productions textuelles.

---

pour le 20<sup>e</sup> s., Colette, Tina Jolas / Carmen Meyer, Marguerite Audoux, Jacques-Émile Blanche / Maurice Denis, Jean Giraudoux / Louis Jouvet.

<sup>89</sup> Une autre piste serait d'exploiter les données que j'ai récoltées en indo-aryen ancien et en sanskrit classique (dans le cadre de mon DEA et de mon doctorat) pour observer le changement linguistique à l'œuvre dans certains phénomènes du lexique (le formulaire ou la polysémie, par exemple), dans le temps long, au sein de trois genres

### 4.3 Comment classer les productions textuelles ?

La méthode paraît simple comme le rappelle Ryan (1981, p. 118) : les catégories génériques constituent ce que Wittgenstein appellerait des ensembles regroupés par « air de famille » et à chaque ensemble serait associé un groupe de caractéristiques dont une partie d'entre elles doit être présente dans tous les membres de l'ensemble. Malrieu & Rastier (2001) excluent de la réflexion les grandes catégories de l'expression ainsi que les catégories sémantiques fonctionnels<sup>90</sup> :

Il n'y a pas de genres suprêmes (pas de genre de genres), puisque les critères de groupement des genres sont les discours et les pratiques qui leur correspondent. Aussi, de grandes catégories de l'expression, comme la prose ou l'oral, conduisent-ils à des regroupements oiseux : par exemple, l'oral, de la brève de comptoir au réquisitoire, n'a évidemment pas plus d'unité que la prose. De même, les catégories sémantiques de type fonctionnel (information, divertissement, etc.) regroupent des textes hétérogènes par leur genre et leur discours. (Malrieu & Rastier 2001, p. 548)

Il convient donc de rechercher d'autres types d'unités pour mener à bien cette entreprise typologique. Selon Bronckart (1996), les unités linguistiques qui sont empiriquement observables dans les textes sont le seul critère aisément objectivable pour classer ceux-ci en genres. C'est la piste qui a été explorée dans plusieurs typologies prenant en compte différents types d'unités linguistiques comme nous le verrons tout d'abord (4.3.1). Nous examinerons ensuite les propositions combinant plusieurs critères linguistiques et extralinguistiques (4.3.2) avant de conclure sur la nécessité de construire une typologie ouverte des genres textuels pour garantir le succès de cette entreprise (4.3.3).

---

poétiques distincts : les hymnes religieux (*veda*), les épopées (*Mahābhārata*, *Rāmāyaṇa*) et les contes en vers (*Kathāsaritsāgara*).

<sup>90</sup> Ce choix revient à exclure par exemple la typologie homogène de Werlich (1975) qui propose un point de vue fonctionnel sur les genres en les classant selon les fonctions de l'acte de communication (informer, convaincre, distraire, produire un effet esthétique). La typologie de Werlich revient à classer les genres traditionnels (poésie, essai, roman, manuel, etc.) au sein d'un nombre limité de méta-catégories (12) combinant les trois postures énonciatives (décrire, raconter, exposer) avec les quatre fonctions de l'acte de communication (informer, convaincre, distraire, produire un effet esthétique). Ce mode de classement regroupe des textes étiquetés de la même manière à deux endroits différents (la publicité peut relever de la description ou de la narration de même que la poésie, par exemple).

### 4.3.1 Classer les textes selon la spécificité de leurs unités linguistiques

Classer les textes selon la spécificité de leurs unités linguistiques revient d'abord à se demander quelles unités prendre en considération. Dans sa typologie pionnière en ce domaine, Biber (1988 ; 1989) a fait le choix d'observer la répartition statistique de traits grammaticaux (passif, pronom, subordination, etc.). Il rassemble des traits qui ont tendance à apparaître ensemble et repère les configurations de traits systématiquement évités par les mêmes rassemblements. Il aboutit à une typologie organisée autour de cinq dimensions considérées comme des prototypes de fonctionnements textuels (production impliquée ou informationnelle ; orientation narrative ou non ; référence dépendante ou non de la situation d'énonciation ; visée persuasive apparente ou non ; style impersonnel ou non). À partir de ces cinq dimensions, Biber a modélisé une classification des productions textuelles en huit catégories<sup>91</sup> :

- *intimate interpersonal interaction* (« interaction interpersonnelle intime », ex. conversation familière) ;
- *informational interaction* (« interaction informationnelle », ex. conversation professionnelle) ;
- *scientific exposition* (« exposé scientifique », ex. article scientifique) ;
- *learned exposition* (« exposé cultivé », ex. critique littéraire) ;
- *imaginative fiction* (« fiction narrative », ex. roman) ;
- *general narrative exposition* (« récit », ex. biographie) ;
- *situated reportage* (« reportage situé », ex. reportage sportif) ;
- *involved persuasion* (« argumentation impliquée », ex. discours politique).

L'appartenance d'un texte à un genre (ou ici *register*) se décide sur la base de ses caractéristiques les plus marquantes et récurrentes comme tendance principale mais Biber insiste pour que soit prise en compte également la variation afin de distinguer des sous-catégories (par ex., le roman de science-fiction par rapport au roman de littérature générale) :

[...] the linguistic characterization of a register should include both its central tendency and its range of variation. In fact, some registers are similar in their central tendencies but differ markedly in their ranges of variation (e.g. science fiction versus general fiction, and official documents versus academic prose, where the first register of each pair has a more restricted range of variation). (Biber 1993b, p. 252).

---

<sup>91</sup> Les traductions françaises des catégories de Biber sont celles proposées par Habert, Nazarenko & Salem (1997, p. 30).

Cette démarche novatrice a permis de mettre en œuvre une approche inductive pour construire « une typologie de textes, basée sur les corrélations effectives entre traits linguistiques. » (Habert, Nazarenko & Salem 1997, p. 30). Mais le risque induit par la méthodologie même est que cette typologie ne vaille que pour les textes du corpus et les traits linguistiques choisis pour les opposer<sup>92</sup>.

Dans la lignée des travaux de Biber qui visaient à « développer une typologie inductive des textes en les caractérisant par un ensemble de dimensions organisant des traits linguistiques » (Malrieu & Rastier 2001, p. 548), Malrieu et Rastier ont mené leur étude sur un corpus trois cents fois plus étendu que celui de Biber et avec des variables trois fois plus nombreuses que celles utilisées par ce dernier. Ainsi, concernant les genres narratifs, leurs résultats montrent que la nouvelle se distingue un peu plus du conte que du roman « sérieux » et que le roman policier se distingue nettement des autres genres narratifs (mémoires, roman « sérieux », récit de voyage, nouvelle, conte). L'observation des variables permet en somme de réunir un ensemble de traits linguistiques au service de la caractérisation d'un genre textuel :

On peut ainsi caractériser, minimalement, le roman policier : caractère bref et ramassé de la phrase, compatible avec des propositions subordonnées complétives ; importance du dialogue ; importance de l'action (par opposition aux analyses et commentaires) ; caractère objectiviste du policier, par contraste avec le roman psychologique. La technique narrative diffère du roman « sérieux », où dominant l'imparfait, le présent et le passé simple, l'imparfait jouant un rôle clé dans la création du monde subjectif de la remémoration, alors que le récit policier privilégie le passé composé qui domine à l'oral et dans les dialogues, et présente le compte rendu d'évènements « réels », non par un narrateur omniscient mais à travers le récit des personnages. (Malrieu & Rastier 2001, p. 563)

Les unités linguistiques observées ici sont des traits grammaticaux (emploi des temps verbaux) mais aussi des dispositifs énonciatifs (narrateur) et des types de séquences textuelles (dialogue).

Les séquences textuelles constituent d'ailleurs un type d'unité linguistique utilisée dans une perspective typologique. C'est même le fondement de la réflexion d'Adam qui rejette les

---

<sup>92</sup> « On ne sait toutefois pas si, en partant d'une autre typologie *a priori*, on ne rassemblerait pas sous un même chef des textes différents, ce qui aurait toutes chances de produire des agrégats de traits linguistiques distincts de ceux produits par la typologie précédente. La répartition des textes retenus sous les rubriques choisies est elle-même contestable. Il y a là une circularité d'autant plus gênante que l'existence de types textuels distincts paraît intuitivement fondée, même s'il s'avère délicat de l'étayer empiriquement. » (Habert, Nazarenko & Salem 1997, p. 28).

typologies de rang textuel pour privilégier une approche discursive du texte<sup>93</sup> en tant que combinaison de séquences textuelles<sup>94</sup>. Pour Adam, il en existe cinq types de base (narratif, argumentatif, explicatif, dialogal, descriptif) qui correspondent aux cinq types de « relations macrosémantiques acquises en même temps que la langue par imprégnation culturelle (par la lecture, l'écoute et la production) et transformées en schémas de reconnaissance et de structuration de l'information textuelle. » (Adam 2011, p. 44). Dans ce cadre, un texte appartient à l'une de ces cinq catégories sur la base de sa « dominante » : par exemple, le conte et la fable sont des genres à *dominante narrative* ; le plaidoyer et la harangue politique sont des genres à *dominante argumentative* ; les notices et contes étiologiques sont des genres à *dominante explicative* ; le portrait, la petite annonce et le guide touristique sont des genres à *dominante descriptive* ; l'épistolaire, le théâtre et l'interview sont des genres à *dominante dialogale*. Adam réfléchit à la manière de réduire l'hétérogénéité du monde des textes en combinant les séquences textuelles avec les plans de texte qui peuvent être *fixes* (par exemple, les cinq actes de la tragédie classique ; la *dispositio* oratoire d'un discours ; le cadre global de la lettre) ou *occasionnels* (dans ce cas, ce sont les signes de segmentation et de connexité qui orientent la compréhension du plan du texte). Les travaux d'Adam s'intéressent, comme ceux de Rastier, à l'effet de genericité c'est-à-dire à ce qui permet de rattacher un texte à un discours. Dans ce cadre, sa démarche est centrée sur la matérialité du texte tout en gardant néanmoins à l'esprit « l'incidence des pratiques sociales sur les codifications linguistiques » (Rastier 1989, p. 40).

Dans cette première sous-section, nous avons examiné la pertinence de classer les textes selon les unités linguistiques qu'ils contiennent et qui permettent de saisir et de rendre compte de la matérialité textuelle du genre. À un premier niveau, la démarche inductive de Biber repose sur l'identification de traits grammaticaux pour réunir les textes de son corpus en sous-ensembles mais elle présente le risque inhérent à toute observation de phénomènes dans un

---

<sup>93</sup> « Les typologies de rang textuel ne sont pas pertinentes, car l'unité empirique "texte" est beaucoup trop complexe et bien trop hétérogène pour présenter assez de régularités linguistiquement observables et codifiables qui permettent de définir des types. En revanche, on comprend que le besoin de classer les productions discursives humaines parte des interactions et des formations sociales dans le cadre desquelles elles se déroulent. » (Adam 2011, p. 35-36).

<sup>94</sup> Les séquences textuelles se définissent comme « des unités textuelles complexes, composées d'un nombre défini de paquets de propositions de base : les macropropositions. Ces macropropositions sont des unités liées aux autres macropropositions de même type et elles occupent des positions précises au sein du tout ordonné de la séquence. » (Adam 2011, p. 44).

corpus. À un deuxième niveau, Malrieu et Rastier (2001), dans leurs travaux sur les genres narratifs, ont croisé des unités linguistiques de niveaux différents (traits grammaticaux, postures énonciatives, types de séquences textuelles). À un troisième niveau, Adam présente une entreprise typologique précisément fondée sur la combinaison de deux unités linguistiques de rang différent : les séquences textuelles et les plans de textes, deux unités linguistiques profondément conditionnées par les pratiques sociales et culturelles<sup>95</sup>.

#### **4.3.2 Classer les textes en combinant plusieurs critères linguistiques et extralinguistiques**

D'autres typologies s'appuient sur la combinaison de critères linguistiques et extralinguistiques. C'est l'idée défendue par Baroni pour qui l'analyse des genres littéraires « ne saurait passer par la simple recherche de structures invariantes au sein d'un corpus d'oeuvres apparentées [...] il faut également prendre en compte le contexte pragmatique de la production et de la réception des œuvres pour pouvoir les approcher. » (Baroni 2003, p. 143). À côté des « structures invariantes » qui représentent les critères linguistiques, le « contexte pragmatique de la production et de la réception » tient lieu de critère extralinguistique.

Dans sa typologie, qu'il qualifie lui-même de typologie « intermédiaire », Bronckart distingue trois ensembles de critères :

[...] les critères psychologiques (qui ont trait au type d'action engagée et au type de processus cognitifs mobilisés), les critères langagiers (qui ont trait aux décisions que tout locuteur doit prendre pour accomplir un acte de production verbale), et les critères linguistiques (qui ont trait aux décisions à prendre pour réaliser concrètement un texte dans le cadre des contraintes d'une langue naturelle donnée). (Bronckart 1996, p. 7)

Les critères linguistiques sont ici entendus au sens restreint des contraintes inhérentes à la bonne mise en œuvre des énoncés dans une langue donnée. Les autres critères, psychologiques et

---

<sup>95</sup> C'est précisément ce cadre théorique que Qianyun Li met en œuvre actuellement dans sa thèse de doctorat *Les influences des conventions rhétoriques du français et du chinois sur les productions écrites d'apprenants sinophones du français*. Nous travaillons sur un corpus de lettres familières écrites par des francophones natifs, des sinophones natifs et des sinophones apprenant le français afin d'observer précisément les pratiques culturelles françaises et chinoises et leur transposition dans le cadre de l'apprentissage du français langue étrangère. L'entrée par le plan du texte et les séquences textuelles est utilisée car c'est précisément « la structuration normative de l'aire scripturale » qui permet d'identifier la lettre comme genre (Equoy Hutin 2007, p. 120). Le genre épistolaire a été précisément choisi dans cette optique puisque comme l'a noté Adam (1998, p. 41) : « en dépit d'une indéniable diversité générique, la forme épistolaire présente un certain nombre de constantes compositionnelles ».

langagiers, s'inscrivent dans le cadre de la psychologie interactionniste sociale choisi par Bronckart. L'action langagière (ou « action impliquant le langage ») se réalise dans un texte empirique considéré comme « le produit de l'interaction entre une action langagière et un genre historique déterminé ». (*ibid.*, p. 10). Le contexte dans lequel celle-ci se déroule peut être défini en faisant intervenir trois paramètres : la situation matérielle d'énonciation ; le cadre social dans lequel se déroule la production ; les représentations du monde mobilisées pour la production. Le processus de production est alors décomposé en une série de décisions et d'opérations psycholinguistiques de la part de « l'émetteur-énonciateur », nommées « opérations psychodiscursives générales », qui sont universelles car nécessaires à la réalisation de toute production textuelle quelle que soit la langue naturelle utilisée. Ainsi, soit l'émetteur intègre dans son texte des renvois explicites aux paramètres de l'énonciation (locuteur, interlocuteur, temps et lieu), paramètres alors *impliqués* dans le texte ; soit il élimine ces indications : le texte, *autonome* par rapport à sa situation de production ne requiert aucune connaissance de celle-ci pour son interprétation<sup>96</sup>. De plus, la représentation du monde fournie dans l'énoncé peut être présentée comme mise à distance de l'interaction sociale en cours (les faits sont racontés comme s'ils étaient passés), c'est la *disjonction*. Elle peut inversement se situer dans le monde de l'interaction sociale en cours, c'est la *conjonction* : le texte montre des états, des actions, des événements accessibles dans le monde des protagonistes de l'interaction. Le tableau 4 ci-dessous récapitule le principe de cette typologie des « architypes discursifs » :

		Rapport au monde	
		Conjonction	Disjonction
Rapport interactif à la situation	Implication	<i>Discours interactif</i>	<i>Récit</i>
	Autonomie	<i>Discours théorique</i>	<i>Narration</i>

Tableau 4 : La typologie des architypes discursifs de Bronckart (1996, p. 14)

Pour que ces architypes discursifs théoriques puissent se concrétiser en « types linguistiques ou types des discours » (*ibid.*, p. 15), c'est-à-dire attester de ces opérations à travers leur façonnage en langue naturelle dans des segments de textes concrets, ce sont « les opérations psychodiscursives spécifiques » qui sont mobilisées (l'émetteur-énonciateur peut choisir de réaliser une narration historique en français au passé simple/imparfait ou au présent, par exemple). Pour

---

<sup>96</sup> Cette distinction entre *implication* et *autonomie* dans le rapport interactif à la situation s'inspire de la distinction de Bakhtine (1984) entre « genres premiers » (dialogues, oraux, langue familière et quotidienne, etc.) et « genres seconds » (roman, théâtre, discours scientifique, etc.) qui simulent les formes des genres premiers.

Bronckart, ce sont les genres qui articulent et combinent de diverses manières ces types psychologiques (les « architypes discursifs ») et linguistiques (les « types linguistiques ou types de discours »). La production textuelle de l'émetteur-énonciateur se réalise par imitation, reproduction ou déformation d'un des genres dont il a une connaissance pratique et intuitive<sup>97</sup>. Finalement, le projet de Bronckart n'est pas de fournir une typologie des genres textuels mais de décrire les différents types psychologiques et linguistiques mis en œuvre dans un genre textuel.

D'autres recherches ont mis l'accent sur les unités linguistiques dans l'entreprise typologique. Ces dernières années, les recherches sur le discours scientifique, et tout particulièrement sur le lexique scientifique transversal et sa phraséologie, ont exploré la piste d'unités linguistiques caractérisant des genres textuels définis au préalable sur la base de critères extralinguistiques. À titre d'exemple, Jacques & Tutin (2018) rappellent les résultats de Fifielska (2015) et Fløttum, Dahl & Kinn (2006) qui ont « relevé un net suremploi (a) des constructions impersonnelles adjectivales comme *il est Adj de V (il est important/crucial de noter/mentionner)*, (b) des structures impersonnelles passives (*il est admis que, il a été montré que*), souvent associées à une forme d'effacement énonciatif de l'auteur fréquent dans ce genre. » (Jacques & Tutin 2018, p. 3). Il s'agit d'unités phraséologiques renvoyant à la démarche, au raisonnement et à l'écriture scientifiques (par ex., *faire une hypothèse, proposer une analyse, en arriver au terme de*) et apparaissant caractéristiques de ce genre textuel<sup>98</sup>. Dans ces études, les textes sont réunis au préalable sous une étiquette générique sur la base de critères externes (cadre socio-historique dans lequel ils ont été produits) et la recherche d'unités phraséologiques spécifiques permettrait de confirmer (ou pas) à posteriori ce classement réalisé à priori. On peut donc considérer que le genre est défini dans ce cas sur la base de critères linguistiques (les unités phraséologiques) et extralinguistiques (au sens du « contexte pragmatique de la production et de la réception » de Baroni).

---

<sup>97</sup> « Tout exemplaire de texte (ou texte empirique) peut dès lors être défini comme une entité linguistique correspondant d'une part à l'unité psychologique que constitue l'action langagière et élaborée d'autre part en interaction avec un des modèles de genres disponibles en un état de langue donné. » (Bronckart 1996, p. 12).

<sup>98</sup> Pour une présentation plus détaillée de ces travaux, voir par exemple Tutin (2014). Comme l'auteure le souligne, « [...] cette phraséologie joue un rôle clé dans le cadre de ce discours : elle permet tout d'abord à l'auteur/chercheur de renforcer son appartenance à une communauté de discours, avec l'emploi d'un langage "codé" (dans la littérature, nous écartons l'hypothèse...), et elle facilite, par son caractère préconstruit, l'écriture. » (p. 27). Des travaux très nombreux ont également été menés en langue anglaise sur ce genre textuel et les *lexical bundles* qu'il contient comme par exemple ceux de Biber, Conrad & Cortes (2004).

C'est dans cette même démarche que nous avons conduit nos recherches sur le genre romanesque au sein des projets PhraseoText et PhraseoRom<sup>99</sup>. Le corpus d'étude, constitué de romans dont la publication est postérieure à 1950, a été partitionné au préalable en six sous-genres romanesques sur la base de critères éditoriaux (la reconnaissance des auteurs par la critique, notamment lors de l'attribution de prix littéraires, et la collection dans laquelle les romans sont publiés) : romans policiers, historiques, sentimentaux, fantasy, de science-fiction et de littérature générale.

Sur la base de cette répartition établie par les stylisticiens participant à ces deux projets, nos travaux ont eu pour objectif de décrire les unités phraséologiques les plus spécifiques au sein de chaque sous-genre. Par exemple, nous (Kraif, Novakova & Sorba 2016) avons ainsi pu montrer que les constructions lexico-syntaxiques (CLS) mettant en œuvre le patron SN + *trop* + ADJ étaient surreprésentées dans les romans de science-fiction (*les soleils trop âgés, son ciel trop bleu*) comparés aux romans policiers tandis que les CLS construites autour du patron *lancer un regard* + ADJ + à étaient plus spécifiques au roman policier qu'à celui de science fiction. Un autre exemple est celui de la comparaison entre les romans historiques et les romans de littérature générale (Gonon & Sorba 2020) : on trouve dans le roman historique de nombreuses CLS traduisant une société hiérarchisée (*monter sur le trône, baiser les mains, être le chef, donner l'ordre, etc.*) et guerrière (*venir à bout, faire la guerre, se faire égorger, etc.*) en comparaison avec le roman de littérature générale. Dans le champ de l'interaction verbale que nous avons investigué plus particulièrement dans cette étude, nous avons pu observer, dans les romans de littérature générale, des CLS récurrentes témoignant d'une interlocution plus souvent tournée vers l'intériorité (*j'ai envie, j'ai peur, je ne sais pas, j'ai oublié, je me dis, etc.*).

Dans une autre étude consacrée aux expressions spatiales (Sorba 2020a), j'ai comparé les CLS exprimant une localisation spatiale dans les romans de littérature générale à celles de trois autres sous-genres (romans historiques, policiers et de science-fiction). J'ai pu observer un noyau commun de huit CLS statistiquement spécifiques aux romans de littérature générale par rapport aux trois autres sous-genres romanesques (*voir la mer, regarder la mer, regarder le ciel, être sur la plage, s'asseoir sur un|le banc, marcher dans la rue, entrer dans ma|la chambre, aller à l'école*). Tous ces faits linguistiques permettent d'esquisser les éléments spatiaux du script spécifique d'un roman de littérature générale.

---

<sup>99</sup> Pour la présentation du corpus et de la méthodologie du projet PhraseoRom, voir Diwersy, Gonon, Goossens, Kraif, Novakova, Sorba & Vidotto (2021). Les deux projets sont rapidement présentés dans l'annexe IV.

Ces études de cas ont pour objectif de faire émerger des unités linguistiques distinctives pour caractériser les genres ou sous-genres romanesques en complément (ou en discussion) des critères extralinguistiques que sont les critères éditoriaux. Classifier les textes en combinant des critères linguistiques et extralinguistiques est une piste envisagée par les recherches actuelles, qu'elles portent sur des textes littéraires ou non. C'est l'une des directions de recherche que je continuerai à explorer dans les années à venir<sup>100</sup>.

À l'issue de ce bref panorama des divers critères utilisés pour construire des typologies des genres textuels, je propose, comme dernier point de ce chapitre 4, de rendre compte de la généricité en construisant une typologie ouverte.

### 4.3.3 Construire une typologie ouverte pour rendre compte de la généricité

L'un des constats issus de la réflexion qui précède est que le genre est bien « le niveau fondamental pour la catégorisation des textes », pour reprendre les mots de Malrieu & Rastier (2001). Il convient donc de proposer une manière de rendre compte de la généricité des textes le plus justement possible.

Tout d'abord, un genre est identifiable dans un état de langue donné. Par conséquent, le « renouvellement constant des genres entraîne logiquement l'impossibilité d'en établir des typologies *a priori*. » (Charaudeau & Maingueneau 2002, p. 282). Pour réussir l'entreprise typologique, il faut accepter de concevoir l'inventaire des différents genres textuels comme une liste ouverte mettant en oeuvre une dynamique dont l'évolution est modelée par des contraintes externes ayant un impact sur les phénomènes linguistiques internes. C'est la réponse que nous pouvons apporter à l'interrogation de Ryan (1981, p. 119) : « The next problem to be taken into consideration, is whether a genre theory should take the form of a systematic taxonomy or of an informal catalogue. ». Choisir la seconde option, un catalogue informel ou une liste ouverte, permet de concilier la dimension historique du genre textuel avec la nécessité classificatoire.

C'est le choix qui a été fait, par exemple, par la *Grande Grammaire Historique du Français* (Marchello-Nizia et al. 2020) qui présente, dans son chapitre 9, une liste des genres attestés en langue française, dans les deux modalités de l'oral et de l'écrit, depuis le 9<sup>e</sup> siècle. Consacrer un chapitre aux genres textuels dans une grammaire historique témoigne de la nécessité de

---

<sup>100</sup> C'est dans cette perspective de contraster les sous-genres romanesques entre eux grâce à leurs CLS spécifiques que s'inscrit la recherche doctorale actuelle de Mohammed Belmokhtare. L'investigation porte plus particulièrement sur les CLS exprimant une émotion et comprenant un nom de partie de corps dans le but de contraster les six sous-genres romanesques.

prendre en compte le paramètre ‘genre’ pour analyser le changement linguistique. À titre d’exemple, est cité le cas de la correspondance privée dans la deuxième moitié du 17<sup>e</sup> siècle :

Dans la deuxième moitié du siècle, le principe de la correspondance privée se généralise, dans la classe élevée, sans intention systématique de publication. Les lettres les plus fameuses de cette fin de siècle, celles de Madame de Sévigné, ne commencent à être publiées qu’en 1734. Le développement de cette pratique spontanée contribue à modifier la syntaxe, avec notamment la multiplication des modalisateurs, la diminution de la subordination, l’extension de la juxtaposition, le développement du clivage, une plus grande souplesse dans la construction des compléments et des expansions. (Marchello-Nizia et al. 2020, p. 130).

Lier changement sociétal et changement linguistique permet de mieux comprendre l’évolution d’une langue. C’était déjà l’hypothèse sémantique de Meillet (1906)<sup>101</sup>. J’ai pu tester ce lien dans une étude diachronique des collocations construites autour du pivot nominal *feu* (Sorba 2018a). Dans un vaste corpus constitué des textes littéraires et journalistiques des bases Frantext et EmoBase, j’ai pu observer le développement exponentiel de deux collocations impliquant le verbe *donner* et la base nominale *feu* au début du 20<sup>e</sup> siècle : *donner du feu* avec l’acception « donner du feu pour allumer une cigarette » (ex. 31) et *donner le feu vert* avec l’acception « autoriser » (ex. 32) :

31) Soudain un manque se fait sentir. Une cigarette. Elle demande à ses amies si elles en ont une et les autres répondent « oui bien sûr ». Elles lui **donnent du feu**, elle fume. (B. Werber, *L’Ultime secret*, 2001)

32) Les autorités bancaires chinoises viennent de **donner leur feu vert** pour la création d’une banque postale. (*Le Figaro* 04/01/2007)

*Donner le feu vert* est une collocation récente (attestée à partir de 1965 dans Frantext), postérieure à l’apparition des premiers feux de circulation au début du 20<sup>e</sup> siècle d’abord aux États-Unis (vers 1914). Par ailleurs, l’explosion du nombre d’occurrences au 20<sup>e</sup> siècle de la collocation *donner du feu* au sens de « donner du feu pour allumer une cigarette »<sup>102</sup> peut

---

<sup>101</sup> « si le milieu dans lequel évolue le langage est un milieu social, si l’objet du langage est de permettre les relations sociales, si le langage n’est maintenu et conservé que par ces relations, si enfin les limites des langues coïncident avec celles des groupes sociaux, il est évident que les causes dont dépendent les faits linguistiques doivent être de nature sociale, et que seule, la considération des faits sociaux permettra de substituer en linguistique à l’examen des faits bruts la détermination des procès, c’est-à-dire à l’examen des choses, l’examen des actions, à la pure constatation de rapports entre phénomènes complexes, l’analyse de faits relativement simples considérés chacun dans leur développement particulier. » (Meillet 1906, p. 6-7).

<sup>102</sup> Dans le sous-corpus Frantext 20<sup>e</sup> siècle, près des trois-quarts des occurrences de la collocation *donner + feu* (29/68 occ. soit 42 %) sont employées dans l’acception « donner du feu pour allumer une cigarette ». La

s'expliquer dans le cadre du développement d'une société « nicotinisée ». En effet, comme le rappelle Nourrisson (1988), à la fin du 19<sup>e</sup> siècle, les progrès technologiques dans la fabrication des cigarettes permettent de faire doubler leur production (de 468 millions en 1874 à 900 millions en 1888) et « le tabac [cesse] d'être une mode pour devenir un besoin, depuis le riche fumeur qui brûle des cabanas à 25 centimes la pièce jusqu'au chiffonnier qui mâche sa cordelette ou ramasse les mégots. » (*ibid.*, p. 538 et 543). Au même titre que les développements technologiques et les habitudes addictives, les genres textuels sont des pratiques sociales qui reflètent et transmettent les évolutions d'une société.

Pour réaliser ce programme d'observation des genres et de leurs évolutions, on peut multiplier les études détaillées de genres particuliers comme le propose Schaeffer (1995). Pour choisir ces genres particuliers qui articulent formes linguistiques et fonctionnements sociaux, Chauraudeau & Maingueneau (2002, *s.v.*) proposent de privilégier « les catégories qui stabilisent des formes d'association entre des formes d'action (rôles discursifs, tâches cognitives), des contenus et des manières de dire (dispositifs d'énonciation, nouvelles dénominations, apparition de formules permettant de ritualiser les pratiques etc.). ». C'est le cas, par exemple, des travaux qui ont été menés récemment sur les sermons protestants du 16<sup>e</sup> au 18<sup>e</sup> siècles (le corpus *Sermo*, voir sitographie), sur les rapports d'éducateurs spécialisés (le projet *Écritures*, voir sitographie) ou sur les discours de vœux des présidents de la 5<sup>e</sup> République française (Leblanc 2016).

Néanmoins, afin de saisir correctement les propriétés des genres textuels, il convient de ne pas les traiter un par un mais de les étudier au milieu des autres pratiques discursives de la même époque :

[...] les genres se définissent non pas ontologiquement et isolément, mais par contrastes au sein d'un système codifié de genres. Un genre ne se comprend donc qu'à l'intérieur d'un ensemble de ressemblances et de différences entre genres et sous-genres définis par un groupe social à un moment donné de son histoire culturelle et discursive. (Adam & Heidmann 2007, p. 27)

La caractérisation d'un genre textuel se fait donc de manière doublement relative : elle est relative à une époque donnée mais aussi à l'ensemble des productions discursives de cette époque.

Par ailleurs, comme toutes les pratiques sociales, les genres sont des pratiques normées, cognitivement et socialement indispensables. Il est donc tout à fait possible d'y observer des

---

collocation *donner + feu* est bien attestée depuis le 16<sup>e</sup> siècle mais avec d'autres acceptions (« faire chauffer », « donner de la force », « tirer au moyen d'une arme à feu », voir Sorba 2018a, p. 274).

régularités puisque « [l]e caractère relativement stable et normé des genres est une des conditions de possibilité des interactions socio-discursives et du fonctionnement de la langue en discours » (Adam & Heidmann 2007, p. 27). Néanmoins, *régularité* ne veut pas dire *rigidité*. Les genres textuels sont bien des catégories dynamiques en variation : comme les langues, les genres sont « des conventions prises entre deux facteurs plus complémentaires que contradictoires : celui de répétition et celui de variation. » (*ibid.*, p. 29). L'une des conséquences de ce caractère dynamique est de pouvoir envisager la participation d'un même texte à plusieurs genres<sup>103</sup>. Il ne s'agit pas de suspendre toute visée typologique mais de penser la généricité comme le résultat d'une négociation, « d'un dialogue continu, souvent conflictuel, entre les instances énonciatives, éditoriale et lectoriale. » (*ibid.*, p. 24)<sup>104</sup>. Cette proposition permet de répondre au phénomène de l'hétérogénéité générique – qui a conduit parfois à abandonner l'entreprise typologique – tout en rendant compte de la complexité des faits de discours.

Dans ce chapitre 4, j'ai tenté de cerner les contours de la notion linguistique de genre textuel en tenant compte des apports de l'analyse du discours et de la sémantique textuelle sans écarter la tradition littéraire dont la longévité et les interrogations sont précieuses quand il s'agit d'examiner les corpus littéraires. Mon travail s'inscrit dans le renouvellement de la question des genres textuels. L'idée clé de ma démonstration est la nécessité de concevoir le genre textuel comme une entité dynamique pour deux raisons. Premièrement, le genre textuel est un phénomène historique, inscrit dans un système et soumise à la variation. C'est précisément la variation qui permet d'affiner la classification en distinguant les genres et les sous-genres les uns des autres. Deuxièmement, le genre textuel est le résultat d'une négociation permanente entre différentes instances qui le redéfinissent sans cesse. La description diachronique est donc particulièrement adaptée à cet objet car le genre soumis au changement en est le témoin premier.

---

<sup>103</sup> C'était déjà l'idée de Derrida (1986) : « Tout texte participe d'un ou de plusieurs genres » (p. 254). Pour sa part, Schaeffer (1989) souligne que la relation d'inclusion d'une classe à l'intérieur d'une autre ne fonctionne pas forcément pour des textes qui manifestent une « multiplicité générique » (p. 70).

<sup>104</sup> Comme le rappelle Baroni (2003), le lecteur se construit son propre cahier des charges d'un genre qui ne correspond pas forcément à celui construit par l'auteur ou par l'éditeur : « Le genre en lui-même, s'il se fonde sur des régularités structurales ou thématiques appartenant à un groupe d'oeuvres, ne peut exister que dans la mesure où il est constitué par un sujet capable de percevoir ces régularités. Le lecteur se constitue des "architextes" par abstraction de règles génériques tirées de l'ensemble des textes concrets qui font partie de son encyclopédie ; ce produit architextuel est donc partiellement subjectif, sujet à des corrections perpétuelles, et les attentes qu'il génère sont très souvent inconscientes. » (p. 143).

Tout l'intérêt d'une typologie réside dans la possibilité de faire émerger des savoirs empiriquement et intuitivement constitués pour rendre compte du genre textuel comme unité structurante puisque les genres sont des « pratiques-empiriques indispensables à la production-écriture comme à la réception » (Adam 1997, p. 978). Cette entreprise ne peut prendre que la forme d'une liste ouverte afin de rendre compte de la variation et du changement linguistiques. Pour ce faire, il convient de s'appuyer sur un ensemble de critères à la fois linguistiques et extralinguistiques. Ces critères réunis font des genres une catégorie « prototypique, c'est-à-dire définissables en termes tendanciels (plus ou moins) plutôt que par des critères stricts » (*ibid.*).

Tout l'enjeu de la troisième partie qui s'ouvre désormais consistera à contribuer à une « théorie opératoire des genres » (Rastier 2011, p. 72) en montrant comment la linguistique de corpus outillés permet de faire émerger de nouvelles unités linguistiques (les phraséologismes) pour servir l'entreprise de typologie générique au sein de corpus préalablement regroupés sur des critères extralinguistiques.



### **Partie 3. Ce que la phraséologie dit des genres textuels grâce à la linguistique de corpus outillés**

« Corpus linguistics does exist. »<sup>105</sup>

L'objectif de cette troisième partie est de montrer, dans le cadre de mes recherches actuelles, l'apport de la linguistique de corpus outillés à l'étude de la phraséologie dans le cadre de la caractérisation des genres textuels. C'est bien « un changement de perspectives des études linguistiques » (Jacques 2005, p. 21) que la linguistique de corpus a apporté depuis maintenant une trentaine d'années. En effet, comme le souligne Rastier, « les observables produits par la linguistique de corpus sont bien des phénomènes nouveaux » (2005a, p. 41). Ainsi, dès le début des années 1980, les observables phraséologiques ont suscité un regain d'intérêt en linguistique, certes, mais aussi en Traitement Automatique des Langues dont les techniques sont utilisées pour la constitution de corpus outillés (Habert *et al.* 1997, p. 56-57).

C'est pourquoi je propose dans un premier temps de situer les enjeux de la linguistique de corpus et de présenter, à titre d'exemples, trois corpus constitués dans le cadre des recherches actuellement en cours dans l'équipe (chapitre 5). J'examinerai ensuite l'outillage des corpus utilisés dans nos travaux pour explorer les corpus puis l'émergence de nouveaux observables phraséologiques qui aura été rendu possible (chapitre 6).

---

<sup>105</sup> Détournement volontaire d'une affirmation de Noam Chomsky recueillie lors d'un entretien avec Bas Aarts dont voici les deux premières répliques : « BA – What is your view of modern corpus linguistics ? NC – It doesn't exist. If you have nothing, or if you are stuck, or if you're worried about Gothic, then you have no choice. » (Aarts 2000, p. 5).

## Chapitre 5. Les enjeux de la linguistique de corpus

Lors de son émergence dans les années 1960, la linguistique de corpus a été traversée par une tension mettant aux prises les linguistes travaillant de manière empirique sur des corpus (*empirical linguists*) et ceux davantage intéressés par la notion abstraite de ‘langue’ (*theoretical linguists*) conçue dans le cadre de la Grammaire Universelle des générativistes (Aarts 2000, p. 5). Cette tension a pu être résolue en considérant la linguistique de corpus comme une méthodologie en linguistique et, à ce titre, il n’y avait aucune contradiction à être à la fois *corpus linguist* et *theoretical linguist*, à l’instar de Fillmore (*ibid.*, p. 5). Dans les années 1980, le terme *corpus linguistics* désignait alors « un ensemble de travaux britanniques, scandinaves et néerlandais sur des corpus informatisés de l’anglais. » (Léon 2008, p. 12). L’engouement pour les corpus à cette époque a pu d’ailleurs être considéré comme « la revanche de l’empirisme » (Habert et al. 1997, p. 8).

Néanmoins, les discussions se sont rapidement engagées au sein de la communauté scientifique pour considérer la linguistique de corpus non plus comme une simple méthodologie mais comme une discipline à part entière :

The problem of defining corpus linguistics and whether as a theory or a methodology has been debated from different standpoints. [...] corpus linguistics has become a new research enterprise and a new philosophical approach to linguistic enquiry. [...] it becomes a discipline in its own right and acquires its own domain. (Tognini-Bonelli 2001, p. 1 et 49)

La linguistique de corpus n’est pas simplement une méthode ; elle s’est révélée comme une discipline en soi avec sa propre philosophie. (Williams 2005, p. 13)

Afin de cerner les enjeux de la discipline nommée *linguistique de corpus*, je propose de montrer dans quelle conception de la linguistique de corpus se situent mes travaux (section 5.1) et comment j’envisage un corpus (section 5.2). Pour illustrer mes propos, je présenterai trois exemples de corpus construits dans le cadre de projets actuellement en cours (section 5.3).

### 5.1 Qu’est-ce que la linguistique de corpus ?

Cette section envisage la linguistique de corpus comme une discipline qui met en œuvre une démarche scientifique fondée sur l’observation (5.1.1). À ses débuts, deux approches différentes se sont constituées (*corpus-based* et *corpus-driven*) que certains opposent toujours alors qu’elles peuvent désormais être pensées en complémentarité (5.1.2).

### 5.1.1 Une discipline fondée sur l'observation

Selon Williams (2005), « La linguistique de corpus ne cherche pas nécessairement les formalismes mais plutôt à révéler les choix linguistiques opérés par des locuteurs dans des contextes réels [... c'est] une discipline d'observation [...] une discipline exploratoire, une discipline pratique ancrée dans les applications » (p. 13-16). En réintroduisant un locuteur réel et sa pratique de la langue, la linguistique de corpus permet aux linguistes de travailler en observant et en décrivant les usages authentiques de la langue. En reprenant les propos de Jacques (2005), nous soulignons les points forts de la linguistique de corpus qui met en lumière des fonctionnements linguistiques échappant à l'intuition (notamment les phénomènes discursifs) et permet donc non seulement d'ajuster lesdites intuitions quand celles-ci ne se vérifient pas, mais aussi de rendre compte de la variation :

[La linguistique de corpus est] le meilleur moyen d'atteindre la diversité des faits de langue authentiques [...et] LA langue, conçue comme l'ensemble des énoncés qui sont effectivement produits par tout ou partie de la communauté linguistique, dans le cadre de systèmes – au pluriel – qu'il convient de décrire en tant que tels. (Jacques 2005, p. 24 et 27)

C'est parce que la linguistique de corpus permet d'observer plus finement les phénomènes, et ainsi de remettre en question une partie des postulats de la linguistique qu'elle a pu progressivement accéder au rang de discipline (Habert et *al.* 1997, p. 9). Williams (2005) propose ainsi de la définir plus précisément comme une « linguistique appliquée qui cherche à comprendre les mécanismes de communication et à apporter des solutions à des questions pratiques. » (p. 13). La particularité de cette linguistique appliquée est formulée par Tognini-Bonelli (2001) : « unlike other applications that start by accepting certain facts as *given*, corpus linguistics is in position to define its own sets of rules and pieces of knowledge before they are applied » (p. 1). La linguistique de corpus a su ainsi se faire une place dans différents domaines comme l'enseignement des langues, la lexicographie, la traduction et la terminologie :

C'est une science appliquée, orientée vers la pratique : enseignement des langues, traduction, confection de grammaires et de dictionnaires, vulgarisation et enseignement de la linguistique, traitement automatique des langues, etc. (Léon 2008, p. 13)

Certains n'hésitent d'ailleurs pas à utiliser le syntagme au pluriel : *les linguistiques de corpus* pour souligner non seulement la diversité des champs d'application ainsi couverts mais renvoyer également aux origines de ces travaux où ce sont « les discontinuités qui l'emportent, ainsi que la diversité, voire l'éclatement, des horizons théoriques et des réalisations pratiques. » (Habert et *al.* 1997, p. 8). La section suivante propose d'aborder les deux approches des corpus

qui sont nées en même temps que la linguistique de corpus et d'examiner le lien qu'elles entretiennent l'une par rapport à l'autre.

### 5.1.2 Les deux approches des corpus : *corpus-based* et *corpus-driven*

Deux approches fondamentales et complémentaires des corpus se côtoient, l'une déductive (ou *corpus-based*) qui utilise le corpus pour confirmer ou infirmer une hypothèse, et l'autre inductive (ou *corpus-driven*) qui cherche à explorer les données du corpus sans à priori. Comme le rappelle Léon (2008), cette distinction entre deux courants est revendiquée essentiellement par « les tenants de la tendance Sinclair qui tiennent particulièrement à se démarquer de la tendance Leech » :

Ainsi Stubbs (1993) ou Tognini-Bonelli (2001) opposent les deux courants par leur méthodologie : alors que l'option « corpus-based » du courant Leech tient le corpus comme un simple réservoir d'exemples destiné à tester ou vérifier des positions théoriques existantes, l'option « corpus-driven », qu'ils préconisent, postule qu'aucune position théorique *a priori* ne préside aux observations sur corpus, la théorie étant induite du corpus. (Léon 2008, p. 12)

Comme l'illustre le tableau 5, ces deux courants ont leur origine dans la London School au sein de laquelle l'intérêt pour la linguistique descriptive (ou empirique), centrée sur l'étude de l'usage, est largement partagé.

London School		
Fondateurs	John Rupert Firth (1890-1960) et Daniel Jones (1881-1967)	
2 <sup>e</sup> génération	Michael A.K. Halliday (1925-2018)	Randolph Quirk (1920-2017)
3 <sup>e</sup> génération	John Sinclair (1933-2007) Chef de file du courant <i>corpus-driven</i>	Geoffrey Leech (1936-2014) Chef de file du courant <i>corpus-based</i>

Tableau 5 : Les principaux linguistes de la London School (d'après Léon 2008, p. 13).

La scission entre les deux courants intervient dès la deuxième génération et se poursuit à la troisième génération avec Sinclair s'inscrivant dans la continuité de Firth alors que Leech se veut en rupture radicale avec ce dernier. Même si Firth n'a pas laissé d'ouvrage majeur et n'a fait qu'esquisser certaines notions, son approche originale a marqué, de façon décisive, l'essor de la linguistique britannique dans les années 1960 (Léon 2008, p. 14). On lui doit notamment l'idée de non séparation entre lexique et grammaire reprise par Halliday et Sinclair dans leur conception probabiliste du langage. Par exemple, dans la lignée de Firth, la recherche des collocations doit s'effectuer dans des textes intégraux et authentiques et non à partir de corpus

constitués d'échantillons. Le sens doit être conçu comme potentiel, dépendant de choix faits dans l'ensemble du texte :

A literary text has meaning only by comparison with a particular language at a particular time. Any stretch of language has meaning only as a sample of an enormously large body of text; it represents the results of a complicated selection process, and each selection has meaning by virtue of all the other selections which might have been made, but have been rejected.... So a poem is a sample of a language; perhaps not a representative sample, but only carrying meaning because it can be referred to a description of a whole language. (Sinclair 1965, p. 76-77, cité par Léon 2008)

La description de ces deux approches *corpus-based* et *corpus-driven* et de leur implication dans la conception de la linguistique de corpus a été bien documentée par Tognini-Bonelli (2001) dont la préférence affirmée va néanmoins à la seconde approche :

[...] the term *corpus-based* is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study. (p. 65)

[...] in a *corpus-driven* approach, the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. (p. 84)

Tognini-Bonelli (2001) souligne que l'inconvénient majeur de la première approche est de ne pas laisser de place à la découverte d'autres observables pour permettre de faire des ajustements à la marge du modèle utilisé. Elle répertorie trois manières de faire du *corpus-based* en s'appuyant sur les travaux de chercheurs qui les mettent en œuvre :

- (i) *insulation* : isoler la théorie des données ou utiliser une théorie préexistante qui n'a pas été formulée à la lumière des données ; les données sont reléguées au second plan (ex. Aarts, Bod ; p. 68) ;
- (ii) *standardisation* : utiliser les principes de simplification et de standardisation pour rendre les données compatibles avec une théorie ou une description préexistante (ex. Leech, « First devise a theory, then test it out. » ; p. 71), ce qui suppose une intervention lors de l'enrichissement du corpus ;
- (iii) *instanciation* : construire les données dans un système de possibilités abstraites que les données ne viennent pas modifier (ex. Halliday ; p. 74).

Ces trois manières ont en commun que les effets du corpus sur la théorie se limitent à valider des paramètres déjà existants et interrogent sur le rôle à donner à la preuve apportée par le corpus. Les données du corpus correspondent rarement strictement à la théorie de départ car la variabilité inhérente de la langue est forcément présente dans le corpus. Dans sa version la plus extrême (*insulation*), on s'éloignerait ainsi trop de la tradition firthienne dans laquelle la preuve fournie par la langue doit être au centre des descriptions et des théories linguistiques.

La seconde approche, *corpus-driven*, est associée à une attitude résolument novatrice qui cherche de nouvelles hypothèses et de nouvelles pistes à explorer sans se contenter de valider des théories anciennes. C'est la voie tracée par Sinclair dont le positionnement cherche toujours à rendre compte de la donnée et à s'appuyer sur elle (« Reflect the evidence », Sinclair 1991, p. 4). Celui-ci est à l'origine des premiers travaux en lexicographie *corpus driven* avec le *Cobuild project* (1987) qui manifeste un changement qualitatif dans la description de la langue. Désormais, l'approche est holistique : « the general methodological path is clear: observation leads to hypothesis leads to generalisation leads to unification in theoretical statement. » (Tognini-Bonelli 2001, p. 85).

Dans mes travaux, j'ai pu expérimenter les deux approches. D'une part, l'approche *corpus-based* a été mise en œuvre lorsqu'il s'agissait d'étayer une hypothèse de départ par l'observation des données au sein d'un corpus. Par exemple, dans une étude en diachronie longue sur les deux noms généraux *émotion* et *sentiment*, nous (Novakova, Sorba & Tutin 2018) avons pu tester l'hypothèse selon laquelle ces deux lexies pouvaient être considérées comme des noms classificateurs mais que chacune avait un fonctionnement distinct : *émotion* est un nom général fonctionnant comme un hyperonyme alors que *sentiment* fonctionne plutôt comme un *shell noun* (ou nom coquille). L'approche *corpus-based* nous a permis de confirmer cette hypothèse de départ tout en fournissant en diachronie des éléments qui permettent d'expliquer comment *sentiment* est devenu un *shell noun*. En effet, alors qu'*émotion* perdait, à partir du 17<sup>e</sup> siècle, la possibilité d'une complémentation, *sentiment* l'a gardée et développée (*un sentiment de* + N affect / N psychologique). Dans ce cas, c'est le classifieur *sentiment* qui joue le rôle de « convertisseur » et permet aux N psychologiques de basculer dans le domaine affectif (*un sentiment de culpabilité, ~ impuissance, ~ impunité*). En outre, nous avons pu vérifier que *sentiment* était plus productif comme anaphore en particulier dans les contextes d'anaphores résomptives, ce qui est conforme au statut de *shell noun*, tout particulièrement, à partir du 19<sup>e</sup> siècle où s'amorce l'emploi plus fréquent du SN *ce sentiment*.

Récemment, dans notre corpus épistolaire en cours de constitution, nous (Fabry, Lignereux & Sorba 2022) avons testé l'hypothèse que la corrélation était un fait de langue caractéristique de la lettre de compliment<sup>106</sup>, une catégorie épistolographique particulière. Pour ce faire, nous

---

<sup>106</sup> Dans notre corpus, nous regroupons sous l'étiquette « lettres de compliment » les lettres de condoléances (CO) et celles de congratulation ou de félicitation (FE). Dans les *Lettres familières, galantes et autres sur toutes sortes de sujets avec leur instruction*, Milleran (1690) indique que « cette sorte de lettres est d'autant plus nécessaire entre amis, qu'elle contient une communication réciproque de déplaisirs, ou un témoignage de la part que nous

avons observé le fonctionnement de trois structures sémantico-logiques reposant sur une corrélation : les comparatifs de supériorité, d'infériorité et d'égalité (*plus/moins/autant que*) :

- 33) Je me sens **plus disposé à pleurer avec vous la perte que** vous avez faite que je ne me juge capable de vous en consoler. (Fenne 1684, CO)
- 34) Je n'ai pas **moins de joie de votre nouvelle alliance que** de votre nouvelle dignité. (Ortigue de Vaumorière 1690, FE)
- 35) Il est impossible de perdre la mémoire d'une personne qui a **autant de mérite que vous**, etc. (Furetière 1693, FE)

Nous avons pu ainsi conclure que la saillance de ce fait de style dans ce type de lettres permet d'ancrer l'interaction verbale dans un genre (la lettre familière) et un milieu (l'élite aristocratique et son idéal du tendre).

D'autre part, l'approche *corpus-driven* est mise en œuvre régulièrement dans les travaux utilisant la ressource Lexicoscope que je présenterai en détail au chapitre 6. À titre d'exemple, je mentionnerai l'une de nos premières études pilotes (Kraif, Novakova & Sorba 2016) qui a permis de faire émerger des corpus romanesques de l'EmoBase des unités phraséologiques très prévisibles (*scène de crime* pour le sous-corpus de romans policiers) ou peu prévisibles (*de cristal* pour le sous-corpus de romans de science fiction). Dans les deux cas, nous avons analysé ces unités phraséologiques comme des récurrences d'ordre thématique qui construisent l'univers de référence du sous-genre romanesque. Dans le cas du roman policier, la collocation lexicale *scène de crime* fait partie des mentions attendues, tout particulièrement dans le thriller depuis les années 1990, comme nous l'avons montré par ailleurs (Gonon et al. 2016, p. 3-4). En revanche, la collocation grammaticale *de cristal* est apparue, à notre surprise, comme spécifique des romans de science-fiction. L'analyse de la combinatoire de cette séquence a révélé que le SN de gauche appartenait à un fonds lexical banal ne présentant aucune difficulté de compréhension pour le lecteur (*barres, rideau, table, anneau, prison, caverne, montagne*, etc.) mais que la combinaison de celui-ci avec *de cristal* créait « un syntagme original provoquant une impression d'étrangeté » (Kraif, Novakova & Sorba 2016, p. 152) :

- 36) Dans les lacs transparents, des poissons aux longues **nageoires de cristal** nous sourient. (B. Werber, *Les Thanatonauts*, 1994)

---

prenons à tout ce qui les touche de quelque façon que ce soit, comme d'autres que nous-mêmes. » (p. 203). Le corpus de cette étude est constitué de manuels épistolographiques du 17<sup>e</sup> siècle (Puget de La Serre, Jacob, Labarre, Menudier, Fenne, Ortigue de Vaumorière, Milleran, Furetière, Richelet, Colomiès) et du 18<sup>e</sup> siècle (Milleran, Grimarest, Philippon de La Madelaine, Vitrac, Chaudon) pour un total de 112 899 mots et 453 lettres.

37) En guise de réponse, les Pégasins dégainèrent leurs **poignards de cristal** et vinrent encadrer l'échevin. (M. Gaborit, *Feals 02 – Le Fiel*, 2001)

On retrouve ici précisément l'une des caractéristiques lexicales du sous-genre romanesque qu'est la science fiction : la présence d'une « xéno-encyclopédie » :

La création d'un univers autre suppose un vocabulaire adéquat, à la fois compréhensible par le lecteur mais un peu différent du vocabulaire habituel, et qui fonctionne par des analogies que le lecteur est conduit à repérer s'il veut poursuivre la lecture. (Bozzetto 2007, p. 60-61)

Les *nageoires de cristal* et les *poignards de cristal* (ex. 36-37) contribuent à la création de cet univers autre, à la différence du *comptoir de cristal* ou du *lustre en cristal*, objets du quotidien, que l'on trouve dans le roman de littérature générale.

À l'usage, nous avons pu observer que les deux approches sont fort complémentaires : chercher, dans les corpus, des données pour étayer une hypothèse de recherche, souvent formulée sur la base d'une intuition (approche *corpus-based*), permet également de découvrir d'autres phénomènes linguistiques encore inaperçus dont il faut rendre compte fidèlement (approche *corpus-driven*). C'est la réelle plus-value de la linguistique de corpus : « corpus work offers no reason or motivation for selecting some evidence and ignoring the rest » (Tognini-Bonelli 2001, p. 49). Plutôt que de les opposer, il me paraît scientifiquement plus pertinent de prendre en compte les apports de ces deux approches.

Après avoir présenté dans cette première section (5.1) l'approche de la linguistique de corpus en tant que discipline scientifique fondée sur l'observation que je mets en œuvre dans mes travaux et avoir montré la complémentarité des deux démarches *corpus-based* et *corpus-driven*, je vais m'interroger sur ce qui est nécessaire à la mise en œuvre de cette discipline, à savoir le corpus. Comment le définir ? Comment le construire ? Telles seront les questions auxquelles je propose de répondre dans cette deuxième section (5.2) avant de présenter trois exemples de corpus différents construits récemment (5.3).

## **5.2 Les corpus : pourquoi et comment les constituer ?**

Cori & David (2008) notent la difficulté à définir une notion, le corpus, dont l'expression fait l'objet d'autant de qualificatifs :

[...] les différents qualificatifs associés au terme *corpus* montrent aussi l'impossibilité d'une définition unifiée : échantillonné / de référence / nu / annoté / comparable / parallèle / aligné / représentatif / de suivi, etc. (p. 113)

Pour reprendre les mots de Rastier (2005a), « un corpus n'est pas plus un sac de mots qu'un nébuleux intertexte » (p. 31). Aussi sa composition doit-elle être réfléchie selon des principes critiques. Ce sont ces principes que j'examinerai ici afin de définir un corpus comme un

ensemble de textes authentiques réunis en vue d'un objectif (5.2.1), un ensemble représentatif (5.2.2) et résultant d'un éventail de choix faits par les linguistes (5.2.3).

### 5.2.1 Le corpus comme ensemble de textes authentiques réunis en vue d'un objectif

L'étymologie latine du terme *corpus* renvoie à sa matérialité qui est celle des textes le constituant. Dans la lignée de la sémantique textuelle de Rastier (2005a), je considère, en effet, que « le texte est pour une linguistique évoluée l'unité *minimale*, et le corpus l'ensemble dans lequel cette unité prend son sens. » (p. 31). Cette première définition du corpus pose deux principes en répondant à deux questions : (i) quelle est l'unité de base du corpus ? Réponse : l'unité de base est le texte ; (ii) à quoi sert le corpus ? Réponse : le corpus sert à donner du sens à l'unité de base qui le constitue. Ainsi, Rastier souligne l'interrelation entre les deux entités, texte et corpus<sup>107</sup>, à la fois dans une perspective ontologique (le texte permet au corpus d'exister) et dans une perspective herméneutique (le corpus permet de donner du sens au texte).

Nous complétons cette première définition avec deux autres principes mis en avant par Sinclair (1991, p. 171) : *la provenance des textes* (« A corpus is a collection of naturally-occurring language text, [...] ») et *l'objectif visé* (« [...] chosen to characterize a state or a variety of a language »). Les unités qui constituent le corpus, les textes, doivent être ainsi des productions naturelles, ce qui garantit leur authenticité<sup>108</sup>. La mention de l'objectif visé par un corpus introduit la présence du linguiste dans une perspective fonctionnelle : un corpus est constitué pour être utilisé en vue de réaliser des analyses linguistiques (« to be used for linguistic analysis », Francis 1992, p. 17). Celles-ci sont menées afin de rendre compte des régularités d'un état de langue ou d'une variété de langue.

Cette perspective fonctionnelle nécessite la formulation d'un dernier principe, présent, par exemple, dans la définition que Tognini-Bonelli (2001) donne du corpus : « a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic

---

<sup>107</sup> Dans son projet de sémantique des textes en corpus, Rastier met l'accent sur deux complémentarités générales : « (i) celle des niveaux de langage ou plans de description (ponctuation, morphologie, syntaxe, sémantique), et (ii) celle des paliers d'organisation et de complexité : mot, phrase, texte, intertexte. » (2005a, p. 42)

<sup>108</sup> C'était déjà même un paramètre central pour Firth qui considère que l'attestation dans des textes authentiques est essentielle à l'approche du sens : « An approach to the meaning of words, pieces, and sentences by the statement of characteristic collocations ensures that the isolate word or piece as such is attested in established texts. » (Firth 1957 [1951], p. xi).

analysis. » (p. 2). Le linguiste qui constitue un corpus se heurte alors à la délicate question de la *représentativité* :

[...] a corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning language as a whole. [...] Representativeness refers to the extent to which a sample includes the full range of variability in a population. » (Biber 1993b, p. 243)

Un corpus représentatif est donc un corpus qui permet de rendre compte de la variation linguistique. Pour Jacques (2005), c'est même dans cette description de la variation que la linguistique de corpus prend tout son sens :

[...] dès lors que l'on pense la langue non comme UN système désincarné et abstrait mais comme un ensemble vivant, peut-être multiforme, où la description de la variation et de la multiplicité des usages peuvent [*sic*] être aussi fructueux pour la découverte des règles que les raisonnements sur les possibles et les impossibles. (Jacques 2005, p. 27)

Comment donc construire cette représentativité ?

### 5.2.2 La corpus comme ensemble représentatif

Selon Biber (1993b), la représentativité d'un corpus est obtenue grâce à la pratique de l'échantillonnage (*sample*). À la différence d'une *base textuelle*, c'est donc l'opération de choix raisonnés qui crée le corpus. C'est également l'idée de Habert (2000) qui définit un corpus d'un point de vue statistique comme un échantillon d'une population d'événements langagiers<sup>109</sup>. À ce titre, la construction d'un corpus est passible de deux types d'erreurs statistiques qui menacent la généralisation qu'on tendrait à faire à partir de lui :

- (i) l'incertitude (*random error*, Biber 1993b, p. 219-220) qui provient d'un échantillon trop petit pour représenter la population réelle ;
- (ii) la déformation (*bias error*, *ibid.*) qui se produit quand les caractéristiques de l'échantillon sont systématiquement différentes de celles de la population que cet échantillon a pour objectif de représenter.

La construction d'un corpus nécessite donc plusieurs phases de validation successives pour confirmer sa représentativité. Néanmoins, Sinclair (1991) rappelle la limite d'un corpus constitué sur la base d'échantillons puisque les observations linguistiques seront forcément plus

---

<sup>109</sup> Cette manière de considérer le corpus lui-même comme un échantillon est rappelée dans Nelson (2010, p. 56-57) qui l'attribue à Renouf : « However, it should also be remembered that the corpus itself is a sample and needs to be representative of a given aspects of language so that 'The first step towards achieving this aim is to define the whole of which the corpus is to be a sample' (Renouf 1987: 2) ».

complètes sur un corpus constitué de textes intégraux : « a corpus made up of whole documents is open to a wider range of linguistic studies than a collection of short samples. » (p. 19).

Habert (2000) précise d'ailleurs qu'améliorer la représentativité d'un corpus consiste à préciser la production et la réception de chacun de ses composants, en lien avec les motifs qui ont conduit à la création du corpus, mais aussi à pouvoir déterminer sur des bases objectivables les différents emplois du langage auxquels on s'intéresse. En fonction des critères retenus, le regroupement des textes en corpus représentatif peut donc se faire de deux manières :

- (i) à priori : en prenant en compte des critères externes comme les conditions de production, les buts visés, l'inscription dans des genres, etc. ;
- (ii) à postériori : en prenant en compte des critères internes c'est-à-dire la propension d'un groupe de textes à recourir à un ensemble de traits linguistiques et à en éviter d'autres.

Le linguiste créant un corpus choisira l'une ou l'autre manière selon son objectif. S'il souhaite confirmer ou infirmer une typologie préexistante et caractériser les emplois correspondant à chaque type présumé, ce sera la première (par ex., Bronckart et al. 1985). S'il souhaite s'affranchir des catégories préexistantes, ce sera la seconde (par ex., Biber 1988).

Le choix fait par les auteurs de la *Grande Grammaire Historique du Français (GGHF)*, Marchello-Nizia et al. 2020) tente, de manière originale, de concilier les deux approches dans la construction de leur « corpus à géométrie variable » (p. 41) qui a servi de base aux observations linguistiques regroupées dans l'ouvrage. Un double corpus a été élaboré pour chaque période (la période avant 1100 puis une période par siècle à partir du 12<sup>e</sup> siècle) : un sous-corpus « noyau » et un sous-corpus « complémentaire ». Le premier a été construit à priori sur la base de critères de composition stricts aussi bien sur le plan qualitatif (date, genre textuel, dialecte, forme) que sur le plan quantitatif (échantillon de 45 000 tokens maximum pour chaque texte mais seulement 20 000 tokens pour les textes répétitifs). Il s'agit donc d'un échantillon représentatif de 200 000 à 240 000 tokens par siècle (sauf pour la période avant 1100 qui ne dépasse pas les 10 000 tokens pour la quasi-totalité des textes disponibles). Le second, le sous-corpus complémentaire, a été constitué pour permettre de valider (ou non) une hypothèse ayant émergé du sous-corpus noyau mais que ce dernier ne permet pas de valider de manière satisfaisante en raison, par exemple, de la fréquence trop peu élevée des attestations ou de la dispersion pas assez grande. Pour certains phénomènes linguistiques, les auteurs ont même eu recours à d'autres corpus spécifiques (notamment pour l'observation des changements à partir du 17<sup>e</sup> siècle en intégrant des écrits de registre plus familier). Dans ce cas, le corpus pour l'observation d'un phénomène linguistique précis est finalisé à postériori.

Par ailleurs, pour dédramatiser le problème récurrent de la représentativité du corpus, Rastier (2005a) rappelle la nécessité vitale de concevoir une application en amont de la constitution de corpus :

[...] un corpus doit être aimé : s'il ne correspond pas à un besoin voire un désir intellectuel ou scientifique, il se périmé et devient obsolète. [...] Aucun corpus ne représente la langue : ni la langue fonctionnelle qui fait l'objet de la description linguistique, ni la langue historique qui comprend l'ensemble des documents disponibles dans une langue. En revanche, un corpus est adéquat ou non à une tâche en fonction de laquelle on peut déterminer les critères de sa représentativité et de son homogénéité. La linguistique de corpus peut ainsi être objective, mais non objectiviste, puisque tout corpus dépend étroitement du point de vue qui a présidé à sa constitution. (p. 32)

Ainsi, pour Rastier, penser une application lors de la construction de son corpus est bien une question de vie ou de mort pour le corpus, la tentative de construire un corpus représentatif de LA langue étant par nature vouée à l'échec. Cette prise en compte essentielle du « point de vue » du linguiste se traduit dans la définition que Rastier donne précisément du corpus comme issu d'un ensemble de choix faits par les linguistes.

### **5.2.3 Le corpus comme un ensemble de choix**

Rastier (2011) insiste en effet sur l'inévitable intervention du linguiste sur les textes lors de la constitution du corpus :

Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications. (p. 33-34)

Cette intervention du linguiste va au-delà de la seule sélection des textes de manière raisonnée : il fournit pour chaque unité du corpus les métadonnées correspondantes (« documentés ») et peut, le cas échéant, la découper en unités plus petites (phrase, token, etc.), et pourvoir celles-ci d'étiquettes (morphologique, syntaxique, etc.). Le corpus est un construit et les données ne sont pas des données au sens où elles résultent d'une sélection (Jacques 2005)<sup>110</sup>. Ces interventions enclenchent les prémisses d'une démarche herméneutique : l'action sur les unités

---

<sup>110</sup> Sur ce point, voir aussi Rastier (2005a) : « Enfin, la linguistique de corpus, dès lors qu'elle adopte un point de vue réflexif à l'égard de ses propres démarches, peut permettre de rompre avec l'objectivisme candide : elle ne pratique pas d'analyse automatique des données, dans la mesure où elles doivent d'abord être qualifiées comme données, puis interprétées avant et après traitement : les données sont ce qu'on se donne. » (Rastier 2005a, p. 41).

du corpus oriente l'interprétation qu'on va en faire. C'est pourquoi Rastier (2005a) insiste sur la nécessité d'élaborer une sémantique de l'interprétation pour qualifier les résultats obtenus, le détour interprétatif étant « une condition première de l'objectivation » (p. 39).

Pour ce faire, il faut extrapoler au corpus tout entier les considérations que Firth (1957 [1951]) réservait initialement au mot : le sens d'un mot est constitué par son usage, à savoir comme occurrence dans un texte dans un contexte situationnel spécifique<sup>111</sup>. C'est d'ailleurs ce que rappelle Mayaffre (2002) :

La linguistique de corpus ou la sémantique du discours ont définitivement démontré que pas plus qu'on ne pouvait comprendre un mot sans la phrase et la phrase sans le discours, on ne pouvait comprendre le discours sans l'interdiscours, le texte sans le co-texte (sans même parler ici du hors-texte), c'est-à-dire aussi et de manière plus générale, le corpus sans le hors corpus. (p. 5)

Avec ce saut qualitatif, on passe d'une « conception logique à une conception distributionnelle selon laquelle le sens d'un mot et plus largement d'une unité textuelle peut se décrire par les contextes dans lesquels il figure. » (Habert et *al.* 1997, p. 114). Mayaffre met en garde contre le recours à des sources externes (d'ailleurs souvent textuelles elles aussi, « le hors corpus ») pour interpréter les composants du corpus. Afin d'éviter que tous « les efforts scientifiques qui président à la constitution des corpus textuels comme à leurs traitements rigoureux [soient] anéantis au moment du bond interprétatif qui nous projette dans la lave d'un intertexte indéterminé, appréhendé intuitivement » (Mayaffre 2002, p. 5), il préconise de construire un « corpus réflexif » en intégrant ces sources externes au corpus. Le but est bien que « chaque texte puisse être replacé dans son réseau co-textuel utile et nécessaire pour sa compréhension et son analyse » (p. 8). Cette manière de faire reviendrait à percevoir autrement le corpus : il ne serait alors plus seulement « considéré comme un rassemblement inanimé de textes à disséquer sous la lumière crue de projecteurs extérieurs, mais comme un outil dynamique et interactif permettant de construire du sens et des connaissances grâce à ses ressources internes et à ses richesses intrinsèques. » (p. 8). Ce serait tout l'intérêt de construire des corpus complémentaires de contraste et d'exégèse pour interpréter les textes d'un corpus noyau.

Après avoir défini ce qu'est un corpus et les principes à mettre en œuvre pour sa constitution, je vais illustrer ces propos théoriques par trois exemples pratiques de corpus actuellement en construction dans mes différents projets consacrés à l'étude de la phraséologie en diachronie.

---

<sup>111</sup> Firth s'appuie lui-même sur Wittgenstein pour qui la signification d'un mot réside dans son usage : « One cannot guess how a word functions. One has to look at its use, and learn from that. » (Wittgenstein 1958 [1953], p. 109 § 340).

### 5.3. Trois exemples de corpus pour l'étude de la phraséologie en diachronie

Dans le cadre de mes recherches actuelles sur la phraséologie en diachronie, quatre corpus en langue française sont en cours de constitution : deux concernent le genre romanesque, un le genre épistolaire et un le genre scientifique. Cette section est consacrée à la présentation des objectifs et des principes qui ont présidé à la construction des trois premiers. Le quatrième corpus ne sera pas abordé ici car il n'a pas fait l'objet d'une réflexion du même ordre quant à sa construction. En effet, dans le cadre du projet GENESIS en partenariat avec l'Université de Lausanne, nous comptons étudier l'évolution de la terminologie au sein de la revue *GENESIS Manuscrits-Recherche-Invention*, consacrée à la critique génétique des textes, depuis sa création (1992-2022). C'est une étude en diachronie courte que nous prévoyons de réaliser à l'occasion des 30 ans de la revue en collaboration avec Rudolf Mahrer, le directeur actuel de la revue (Université de Lausanne). Je présenterai tout d'abord le corpus PhraseoRom\_19<sup>e</sup>\_fr qui a été pensé pour agrandir un corpus déjà existant (5.3.1). Ensuite, je détaillerai la création de deux corpus en diachronie longue, le corpus PhraseoRoChe, consacré aux romans d'Ancien Régime (5.2.2), et le corpus épistolaire PhraseoCorr (5.2.3).

#### 5.3.1 Agrandir un corpus romanesque : le corpus PhraseoRom\_19<sup>e</sup>\_fr (1830-1899)

La création du corpus PhraseoRom 19<sup>e</sup>\_fr a été pensée pour agrandir l'empan chronologique du corpus français PhraseoRom\_fr. Ce dernier est un corpus de romans français écrits entre 1950 et 2016, finalisé dans le cadre du projet ANR-DFG éponyme et réunissant 1 131 romans pour un total de 103 809 358 tokens (pour le détail, voir Diwersy et al. 2021). Afin de mener des études en diachronie sur les spécificités du genre romanesque, nous avons décidé de compléter ce premier corpus. L'objectif de la création de ce second corpus est donc de réaliser des études linguistiques pour caractériser une langue littéraire et son évolution au sein d'un genre textuel précis grâce à la détection d'unités phraséologiques caractéristiques. Je présente ici les principes que j'ai mis en œuvre dans le *design* de ce corpus.

Le *terminus ante quem* du corpus PhraseoRom\_19<sup>e</sup>\_fr a été fixé à 1830, un tournant dans l'histoire française et dans celle de la littérature (Agulhon 1980). Sur les sept décennies (1830-1899), les romans ont été sélectionnés sous la forme de textes intégraux comme dans le corpus PhraseoRom (1950-2016). Afin d'assurer la représentativité du genre romanesque au 19<sup>e</sup> siècle, les textes du corpus ont été sélectionnés selon quatre critères externes :

- (i) Le roman est une oeuvre majeure ou mineure : l'objectif est d'inclure, pour un tiers, les cinq auteurs phare du 19<sup>e</sup> siècle (Balzac, Flaubert, Hugo, Stendhal, Zola) ainsi que des œuvres mineures de la littérature populaire (jeunesse, anticipation, etc.) ;
- (ii) Le roman a une longueur de plus de cent pages : l'objectif est de ne pas intégrer des documents qui relèveraient plutôt du genre de la nouvelle ;
- (iii) Le sexe de l'auteur : l'objectif est d'atteindre environ 10 % de romans écrits par des femmes ;
- (iv) La périodisation : choisir dix romans par décennie (1830-1899) pour couvrir, de manière régulière, l'ensemble de la production romanesque du 19<sup>e</sup> siècle.

La première étape a été de faire l'inventaire des romans accessibles sous forme numérique dans les corpus déjà existants selon le principe rappelé par Nelson (2010) : « the best source of texts for corpus usage are others, pre-existing corpora. » (p. 61). Lors de l'élaboration de la liste des romans, nous avons donc consulté les bases textuelles Frantext et Wikisource ainsi que différentes ressources en ligne pour la mise en œuvre de ce cinquième critère<sup>112</sup>. Les 70 textes du corpus ont tous été sélectionnés sur la base de leur existence préalable sous format numérique, au moins sous la forme d'un texte déjà ocrisé. Le tableau 6 présente, par ordre chronologique, la liste des romans finalement retenus en croisant ces cinq critères :

<b>Auteur</b>	<b>Roman</b>	<b>Date</b>
Honoré de BALZAC	<i>La Peau de chagrin</i>	1830
Sophie GAY	<i>Le Moqueur Amoureux</i>	1830
STENDHAL	<i>Le Rouge et le Noir</i>	1830
Victor HUGO	<i>Notre-Dame de Paris</i>	1832
Honoré de BALZAC	<i>Le Colonel Chabert</i>	1832
George SAND	<i>Indiana</i>	1832
Théophile GAUTIER	<i>Mademoiselle de Maupin</i>	1835
Léon GOLZAN	<i>Le Notaire de Chantilly</i>	1836
SAINTINE	<i>Picciola</i>	1836
Frédéric SOULIE	<i>Les Mémoires du diable</i>	1837
Honoré de BALZAC	<i>Albert Savarus</i>	1842
Louis REYBAUD	<i>Jérôme Paturot à la recherche d'une position sociale</i>	1842
Eugène SUE	<i>Les Mystères de Paris</i>	1843
Paul FEVAL	<i>Les Mystères de Londres</i>	1844
Alexandre DUMAS	<i>Les Trois Mousquetaires</i>	1844
Delphine de GIRARDIN et al.	<i>La Croix de Berny</i>	1845
Honoré de BALZAC	<i>Une ténébreuse affaire</i>	1846

<sup>112</sup> Voir sitographie. Cette opération a été réalisée, sous ma supervision, par deux étudiants dans le cadre d'un stage de Licence (Léo Mayère, 2022) et de vacances de recherche (Ivan Romaniuk, 2021) au sein du laboratoire LiDiLEM. Elle s'inscrit dans le cadre du projet Approches numériques des textes littéraires : vers une convergence méthodologique ? que nous menons conjointement avec l'Université de Lausanne (voir annexe IV).

Théophile GAUTIER	<i>Les Roués innocents</i>	1847
Jules SANDEAU	<i>Mademoiselle de la Seiglière</i>	1848
George SAND	<i>La Petite Fadette</i>	1849
Henry MURGER	<i>Scènes de la vie de jeunesse</i>	1851
Jules d'Aureville BARBEY	<i>Une vieille maîtresse</i>	1851
Delphine de GIRARDIN	<i>Il ne faut pas jouer avec la douleur</i>	1853
Alfred de MUSSET	<i>La Mouche</i>	1854
Charlemagne I. DEFONTENAY	<i>Star ou Ψ de Cassiopée</i>	1854
Gustave FLAUBERT	<i>Madame Bovary</i>	1857
Théophile GAUTIER	<i>Le Roman de la momie</i>	1858
Victorine MONNIOT	<i>Le journal de Marguerite</i>	1858
Louise COLET	<i>Lui</i>	1859
Comtesse de SEGUR	<i>Les Malheurs de Sophie</i>	1859
Judith GAUTIER	<i>Le dragon impérial</i>	1860
Edmond ABOUT	<i>Le Nez d'un notaire</i>	1862
Victor HUGO	<i>Les Misérables</i>	1862
Gustave FLAUBERT	<i>Salammbô</i>	1863
Jules VERNE	<i>Voyage au centre de la Terre</i>	1864
Comtesse de SEGUR	<i>Un bon petit diable</i>	1865
Prosper MERIMEE	<i>La Chambre bleue</i>	1866
Emile ZOLA	<i>Thérèse Raquin</i>	1867
Aristide ROGER	<i>Aventures extraordinaires de Trinitus, Voyage sous les flots</i>	1868
Gustave FLAUBERT	<i>L'Éducation sentimentale</i>	1869
Emile ZOLA	<i>La Curée</i>	1872
Alphonse DAUDET	<i>Les aventures prodigieuses de Tartarin de Tarascon</i>	1872
Jules VERNE	<i>Le Tour du monde en quatre-vingts jours</i>	1873
André LEO	<i>La Commune de Malenpis : conte</i>	1874
Joseph Arthur de GOBINEAU	<i>Les Pléiades</i>	1874
Alphonse BROWN	<i>Voyage à dos de baleine : aventure du capitaine Bob Kincardy</i>	1876
Joris-Karl HUYSMANS	<i>Marthe, histoire d'une fille</i>	1876
Emile ZOLA	<i>L'Assommoir</i>	1877
Emile ZOLA	<i>Une page d'amour</i>	1878
André THEURIET	<i>La Maison des deux Barbeaux</i>	1879
Emile ZOLA	<i>Nana</i>	1880
Joris-Karl HUYSMANS	<i>Sac au dos</i>	1880
Emile ZOLA	<i>Pot-bouille</i>	1882
Emile ZOLA	<i>Au bonheur des dames</i>	1883
Guy de MAUPASSANT	<i>Une vie</i>	1883
Joris-Karl HUYSMANS	<i>À rebours</i>	1884
Léon BLOY	<i>Le Désespéré</i>	1886
Emile ZOLA	<i>L'Œuvre</i>	1886
Joris-Karl HUYSMANS	<i>Un dilemme</i>	1887
Georges COURTELINE	<i>Le Train de 8 h 47</i>	1888
Emile ZOLA	<i>La Bête humaine</i>	1890
MAUPASSANT, Guy de	<i>Notre Cœur</i>	1890
Joris-Karl HUYSMANS	<i>Là-bas</i>	1891
Louis BOUSSENARD	<i>Les français au Pôle nord</i>	1892
Paul d'IVOI/Henri CHABRILLAT	<i>Les Cinq Sous de Lavarède</i>	1894
Joris-Karl HUYSMANS	<i>En route</i>	1895
André GIDE	<i>Les Nourritures terrestres</i>	1897
Anatole FRANCE	<i>L'Orme du mail</i>	1897
René BOYLESVE	<i>Le parfum des îles Borromées</i>	1898
Alphonse ALLAIS	<i>L'Affaire Blaireau</i>	1899

Tableau 6 : Liste des 70 romans du corpus PhraseoRom\_19<sup>e</sup>s\_fr (par ordre chronologique).

La collecte des textes est terminée et le traitement numérique se poursuit. La première phase de traitement concerne uniquement les 19 textes disponibles sous forme ocrisée seule sans aucune validation. Dans ce cas, un travail de relecture intégrale est réalisé par un membre de l'équipe pour vérifier la conformité du texte par rapport à l'original. En l'état actuel de la constitution du corpus, nous disposons de 51 textes entièrement traités pour un total de 7 019 957 tokens intégrés dans le Lexicoscope<sup>113</sup>.

Une rapide interrogation sur les deux corpus PhraseoRom\_fr et PhraseoRom\_19<sup>e</sup>s\_fr permet d'entrevoir quelques pistes pour l'étude de la phraséologie en diachronie. Par exemple, une requête avec le mot *chagrin* montre que le collocatif lexical le plus fréquent dans le roman contemporain est le nom *amour* (LLR 510,61) alors que celui-ci n'apparaît pas dans la liste des collocatifs les plus spécifiques de *chagrin* du corpus romanesque du 19<sup>e</sup> siècle. Dans ce dernier, les collocatifs lexicaux spécifiques de la lexie *chagrin* sont plutôt les adjectifs (*grand, noir, gros, réel, profond, etc.*). Une première enquête thématique pourrait ainsi s'organiser autour des lexies d'émotion, un champ lexical bien étudié par les membres de l'équipe. En interrogeant tout le vocabulaire du corpus, on s'aperçoit néanmoins que les dix noms les plus spécifiques<sup>114</sup> sont relativement identiques comme le montre le tableau 7 :

PhraseoRom_19 <sup>e</sup> s_fr (LLR)	PhraseoRom_fr (LLR)
<i>homme</i> (11 961)	<i>homme</i> (140 734)
<i>femme</i> (9 659)	<i>œil</i> (111 337)
<i>Madame</i> (9 355)	<i>main</i> (107 313)
<i>jour</i> (8 344)	<i>femme</i> (91 291)
<i>œil</i> (7 908)	<i>jour</i> (90 533)
<i>M.</i> (7 883)	<i>temps</i> (86 180)
<i>Monsieur</i> (7 187)	<i>tête</i> (84 320)
<i>main</i> (7 124)	<i>vie</i> (70 498)
<i>heure</i> (7 075)	<i>heure</i> (66 130)
<i>chose</i> (5 992)	<i>père</i> (58 525)

Tableau 7 : Comparaison des dix noms les plus spécifiques dans les deux corpus.

En effet, on retrouve dans ces deux listes les cinq noms *homme, femme, œil, main, jour*. En revanche, les termes d'adresse *Madame, Monsieur* et *M.* sont plus spécifiques du corpus

<sup>113</sup> Dans une deuxième phase, tous les textes sont encodés dans un fichier xml avec un balisage minimal des chapitres suivant les principes de la TEI. Le corpus devrait atteindre la taille de dix millions de tokens. Il sera librement interrogeable sur le Lexicoscope une fois complet.

<sup>114</sup> La spécificité est exprimée par la mesure d'association *log likelihood ratio* (ou LLR), voir *infra* section 6.1.2.

PhraseoRom\_19<sup>e</sup>s\_fr. La présence des noms de partie du corps *main* et *œil* est intrigante et méritera une investigation poussée une fois le corpus finalisé<sup>115</sup>.

### 5.3.2 Créer un corpus romanesque pour des états de langue anciens : le corpus PhraseoRoChe (1200-1650)

Le deuxième exemple que je documente dans cette section concerne la création du corpus PhraseoRoChe (« Phraséologie dans les Romans de Chevalerie »). Celui-ci s'inscrit dans la continuité d'un projet que je mène depuis 2018 en partenariat avec le laboratoire Litt&Arts (UMR 5316) : Phraséologie et genres textuels : le cas du roman médiéval. Dans le cadre de ce premier projet, nous avons commencé à constituer un corpus des romans en prose du 13<sup>e</sup> siècle dans l'objectif d'étudier le rôle structurant de la phraséologie dans la caractérisation de ce genre textuel émergent (Denoyelle & Sorba 2020). Pour assurer sa représentativité, nous avons privilégié le critère d'exhaustivité car les documents sur la période sont finalement peu nombreux. À ce jour, nous avons sélectionné les sept romans suivants : *Tristan en prose*, *Lancelot en prose*, *Artus de Bretagne*, *Merlin en prose*, *Premiers faits du roi Artur*, *La Queste del saint Graal*, *La Mort Artu*. Comme aucune version numérique n'était alors disponible – hormis pour les deux extraits de *La Queste del saint Graal* (SRCMF) et de *La Mort Artur* (Frantext) –, les textes ont été ocrés puis vérifiés par l'équipe<sup>116</sup>. En fonction de la taille de chaque œuvre et des forces de l'équipe, nous avons sélectionné soit le texte intégral (*Merlin en prose* ; *Artus de Bretagne*, *Premiers faits du roi Artur*) soit un échantillon des romans (*Lancelot en prose* ; *Tristan en prose*) comme présenté dans le tableau 8 :

---

<sup>115</sup> L'objectif sera ensuite de construire la dernière partie du corpus romanesque couvrant la période 1900-1949. La réflexion sur sa construction sera prochainement lancée.

<sup>116</sup> Nous avons utilisé deux logiciels d'OCR : AbbyyServer mis à disposition par la TGIR HumNum et un logiciel gratuit en ligne : <https://www.onlineocr.net/fr/>. Les vérifications sont encore en cours pour les *Premiers faits du roi Artur*.

<b>Titre du roman</b>	<b>Nb token</b>	<b>Échantillon et édition</b>
<i>Artus de Bretagne</i>	215 797	Intégral – édition Ferlampin-Acher (2017)
<i>La Quête del Saint Graal</i>	43 658	Échantillon – corpus SRCMF (2013)
<i>Lancelot</i>	263 968	Échantillon (t.1-2) – édition Micha (1978)
<i>Merlin</i>	67 987	Intégral – édition Micha (1979)
<i>Tristan en prose</i>	380 328	Échantillon (t.1-2-3) – édition Curtis (1953-1985)
<i>Tristan en prose</i>	84 111	Échantillon (t.1) – édition Ménard (2000)
<i>La Mort Artu</i>	25 020	Échantillon – édition Hult (2009)
<i>Premiers faits du roi Artur</i>	256 366	Intégral – édition Freire-Nunes (2001)
Total	1 337 235	

Tableau 8 : Présentation du corpus des romans en prose du 13<sup>e</sup> siècle.

Les romans arthuriens réunis dans ce corpus couvrent, de manière équilibrée, les deux grands monuments de la littérature médiévale : le cycle de Lancelot ou du Graal (65 % des tokens), « la semence du roman du XIII<sup>e</sup> siècle » (Poiron & Walter 2001, p. x), et l’histoire de Tristan (35 % des tokens), « l’une des plus belles histoires d’amour et de mort dont l’humanité se soit enchantée » (Coulet 1991, p. 35). Dans la continuité de ce premier travail sur les romans en prose du 13<sup>e</sup> siècle, l’équipe a décidé de continuer à améliorer ses outils de traitement du français médiéval et de développer un corpus romanesque en diachronie longue, consacré au roman d’Ancien Régime, toujours au service des études phraséologiques et stylistiques en lien avec la caractérisation d’un genre textuel. C’est ainsi qu’a vu le jour le projet Phraséo 13-18 : Phraséologie et stylistique dans le roman d’Ancien Régime, que je dirige actuellement.

La première étape, actuellement en cours, consiste à construire un corpus de romans de chevalerie en prose couvrant la période 1200-1650 : le corpus PhraseoRoChe. Le choix du roman de chevalerie a été fait, après un inventaire de la matière romanesque sur la période concernée, pour deux raisons. La première, d’ordre pratique, se présente toujours aux *corpus designers* : la faisabilité dans le temps et avec les moyens impartis (« Any attempt at corpus creation is therefore a compromise between the hoped for and the achievable. », Nelson 2010, p. 60). C’est la raison pour laquelle nous avons choisi d’exclure les romans uniquement disponibles sous forme de manuscrits et de nous focaliser sur un seul sous-genre romanesque<sup>117</sup>.

<sup>117</sup> Durant la période d’Ancien Régime, le roman de chevalerie est un sous-genre parmi d’autres (romans pastoral, sentimental, picaresque, épistolaire, populaire, etc.).

La seconde raison découle de la première et se formule sous la forme de la question suivante : comment construire un corpus qui permette de tester valablement notre hypothèse de recherche qui est de montrer le rôle structurant de la phraséologie dans la caractérisation du genre romanesque ? Le choix de l'homogénéité thématique a donc été fait : les romans de chevalerie ainsi que leurs adaptations et leurs traductions. Le choix de la période 1200-1650 s'est également imposé pour deux raisons structurelles : la première, d'ordre générique, le roman en prose débutant au 13<sup>e</sup> siècle et l'écriture de romans de chevalerie s'arrêtant vers 1630 ; la seconde, d'ordre linguistique, la période du français préclassique se terminant vers 1650<sup>118</sup>.

Une fois la liste exhaustive des textes correspondant à ces deux critères, thématique et chronologique, établie, nous avons finalisé le choix des textes dans celle-ci selon plusieurs paramètres pour viser une représentativité et un équilibre satisfaisants. D'une part, nous avons mis en oeuvre une réflexion quantitative. Concernant la taille du corpus PhraseoRoChe, nous prévoyons d'arriver à deux millions de tokens afin de permettre des explorations variées dans le Lexicoscope. Le poids respectif de chaque période linguistique a donc été fixé en fonction de ce plafond de deux millions et de la production de ces romans de chevalerie durant chacune d'entre elles. Incontestablement, la période du moyen français (MF) étant la plus longue et la plus productive, les œuvres de cette époque occupent la moitié du corpus (soit 19 textes). Les deux autres périodes, ramassées chacune sur un seul siècle, occupent plus ou moins 25 % du corpus. La place du corpus en ancien français (AF) a été fixée à 23 % du total en raison du nombre limité de romans de chevalerie en prose au 13<sup>e</sup> siècle (soit 9 textes).

D'autre part, nous avons appliqué des critères d'ordre qualitatif. La période de français préclassique (FPC) apporte de nouveaux matériaux avec les premières traductions de romans de chevalerie italiens et espagnols (soit 12 textes pour 27 % du corpus). Pour privilégier une plus grande dispersion, nous avons choisi d'échantillonner toutes les œuvres du corpus à hauteur d'environ 50 000 tokens. La sélection des échantillons d'environ 50 000 tokens tient compte du découpage initial du texte en partie ou en chapitre afin de garder un tout cohérent. Le choix a ainsi été fait de sélectionner l'échantillon d'un seul tenant car la diversité des séquences textuelles (descriptive, narrative, dialoguée, etc.) présentes dans un échantillon de cette taille

---

<sup>118</sup> On a ainsi choisi d'écarter temporairement les romans dits de la Bibliothèque bleue même si ceux-ci sélectionnent des passages des romans de chevalerie du Moyen-Âge car ce corpus n'est stabilisé qu'au 18<sup>e</sup> siècle (Andries 2015).

est satisfaisante et représentative de l'ensemble de l'économie de l'œuvre<sup>119</sup>. Il s'agit également de ne pas surreprésenter une œuvre dans le corpus (par exemple, pour la période de l'ancien français pour laquelle peu de textes sont finalement disponibles), ce qui biaiserait quelque peu les résultats de recherche obtenus. Le tableau 9 ci-dessous présente le corpus PhraseoRoChe :

Période	Textes choisis
AF	Les 8 textes du tableau 8 + <i>Perceval en prose</i> (édition Cerquiglini 1981)
<b>Taille visée : 450 000 tokens (9 textes)</b>	
MF	<i>Le roman du noble et vaillant chevalier Tristan</i> (édition Winn 2020)
	<i>Perceforest</i> (édition Roussineau 2007)
	<i>Jehan de Saintré</i> (édition Blanchard & Quereuil 1995)
	<i>Le roman de l'Estoire dou Graal</i> (édition Nize 1927)
	<i>Gerard d'Euphrate</i> (édition Cooper 2012)
	<i>Philandre</i> (édition Mounier 2015)
	<i>Perceval le gallois</i> (édition Colombo Timelli 2021)
	<i>Mise en prose de Robert le Diable</i> (édition M. G. Ricci -M. Colombi Timelli, non publiée)
	<i>Mise en prose du roman de Troie</i> (édition V. Mezzana – F. Vigneron, non publiée)
	<i>Mélusine</i> (BFM)
	<i>Le Jouvencel</i> (BFM)
	<i>Roman de Berinus</i> (BFM)
	<i>Roland l'amoureux</i> (1 <sup>re</sup> traduction <i>Orlando innamorato</i> de Boiardo par V. Jacques, édition 1549-1550)
	<i>Amadis de Gaule</i> (1 <sup>re</sup> traduction <i>Amadís de Gaula</i> de Garcí Ordóñez de Montalvo par N. de Herberay des Essarts, édition Bideaux 2011)
	<i>Roland furieux</i> (1 <sup>re</sup> traduction <i>Orlando furiosio</i> de L'Arioste, édition 1544)
	<i>Palmerin d'Olive</i> (1 <sup>re</sup> traduction <i>Palmerín de Oliva</i> par J. Maugin, édition 1546)
	<i>Guillaume de Palerne</i> (édition 1552)
	<i>Richart sans paour</i> (édition 1529-1530)
	<i>Cronique et histoire faicte et composée par le reverend père en dieu Turpin</i> (édition 1527)
<b>Taille visée : 1 000 000 tokens (19 textes)</b>	
FPC	<i>L'Histoire palladienne</i> (1 <sup>re</sup> traduction de <i>Florendo de Inghlaterra</i> par C. Colet, édition 1555)
	<i>Amadis de Gaule XXI</i> (1 <sup>re</sup> traduction par G. Chappuys, édition 1577)

<sup>119</sup> Marchello-Nizia et al. (2020) ont fait un choix différent pour le corpus de la *GGHF* : les 45 000 tokens de chaque échantillon sont pris à trois endroits différents du texte (début, milieu et fin).

<i>Primaleon de Grece IV</i> (1 <sup>re</sup> traduction de <i>Primaleón de Grecia</i> de F. Vázquez, édition 1577)
<i>La Suite de Roland le Furieux</i> , (1 <sup>re</sup> traduction de plusieurs suites d'Orlando furioso de L'Arioste par F. De Rosset, édition 1615)
<i>Amadis de Gaule XXII</i> (1 <sup>re</sup> traduction anonyme, édition 1615)
<i>Le Renaud amoureux, histoire précédente de Roland l'amoureux et furieux. Imité de l'Italien du Sr Torquato Tasso, par le Sr de la Ronce et dédié à Monseigneur le Duc de Nevers</i> (édition 1620)
<i>L'Admirable Histoire du Chevalier du Soleil</i> (1 <sup>re</sup> traduction <i>Caballero de sol</i> de Pedro Hernández de Villaumbrales par F. De Rosset & L. Doüet, édition 1620)
<i>Le Romant des romans</i> (G.S. Du Verdier, édition 1626)
<i>Amadis I</i> (P. de Marcassus, édition 1629)
<i>Le Roman des Dames</i> (G.S Du Verdier, édition 1630)
<i>Les Admirables Faits d'armes d'Alcestes servant l'infidèle Lydie</i> (N. Des Escuteaux, édition 1613)
<i>Le Nouveau Tristan</i> (J. Maugin, édition 1554)
<b>Taille visée : 550 000 tokens (12 textes)</b>

Tableau 9 : Le corpus PhraseoRoChe.

Concernant les périodes de MF et de FPC, le travail s'oriente désormais vers la fabrication des échantillons numériques. En effet, la majorité des textes n'existant pas, à notre connaissance, sous une forme numérique autre qu'une image, un travail d'océrisation<sup>120</sup> et de vérification a été lancé (sauf pour les trois textes issus de la BFM). Ce travail d'océrisation est réalisé à partir d'éditions scientifiques modernes quand elles existent en moyen français (13 textes) ou le plus souvent à partir d'éditions anciennes (tous les textes de FPC et 6 textes en MF). Dans ce dernier cas, un guide récapitulant le protocole de transcription conforme aux habitudes des éditions imprimées des 15<sup>e</sup> et 16<sup>e</sup> siècles a été construit par l'équipe à destination des forces d'appoint (stagiaires et vacataires) en charge de transcrire ces textes.

<sup>120</sup> Le logiciel d'OCR utilisé ici est *trankribus* <<https://readcoop.eu/transkribus/download/>>.

La construction ambitieuse de ce corpus répond aussi à notre souhait de mettre à la disposition de la communauté scientifique des textes en langue ancienne non encore disponibles en format numérique. Ce sera l'une des plus-values de notre projet car comme le rappelle Prévost (2015) :

[...] c'est toujours, *in fine*, à travers les textes qui nous sont parvenus que nous percevons la langue ancienne. Les textes constituant par ailleurs l'objet d'étude, il convient, afin d'échapper à une possible circularité, d'envisager toujours des données nouvelles, de prendre en compte des textes encore peu sollicités : cela permet de confirmer, d'infirmer, de corriger ou de compléter des propriétés précédemment mises au jour ou d'en dégager de nouvelles. (p. 25).

Nous faisons ainsi l'hypothèse que la fouille automatique de ces textes apportera de nouvelles données linguistiques et stylistiques au service de la caractérisation du genre romanesque tout en permettant d'affiner les descriptions de phénomènes déjà connus.

### **5.3.3 Créer un corpus épistolaire : le corpus PhraseoCorr (1640-1960)**

Le troisième et dernier corpus diachronique présenté dans ce chapitre 5 concerne « l'épistolaire ». Comme le note Haroche-Bouzinac (1995), un certain « malaise lié à la nature même de la lettre » (p. 9) s'observe dans les réticences des théoriciens à le catégoriser comme un genre littéraire. Même si son écriture répond à un cahier des charges minimal, la lettre échappe parfois à son destinataire et à son auteur lorsqu'elle est publiée alors que ce projet n'avait pas été initialement conçu. À la suite d'Ablali (2013, p. 225), nous considérons qu'il n'y a pas de « discours épistolaire » tel que posé par Adam (1998) mais bien un « genre épistolaire » qui s'inscrit au sein d'une variété de discours (littéraire, médical, diplomatique, etc.). C'est un genre ancien, que l'on pense aux courriers diplomatiques des rois hittites du 2<sup>e</sup> millénaire avant n.è. (voir CTH, sitographie) ou, plus près de nous, à la correspondance privée entre le romain Cicéron et son ami Atticus au 1<sup>er</sup> siècle avant n.è.<sup>121</sup>. Ce genre se ramifie en divers « sous-genres » pour reprendre la hiérarchie des niveaux proposée par Malrieu & Rastier (2001, voir *supra* 3.2.2).

---

<sup>121</sup> Cécile Jullion vient de commencer une thèse de doctorat (octobre 2021) précisément sur ce corpus. Je co-encadre son travail dont l'objectif est de proposer une typologie des usages du bilinguisme gréco-latin dans la *Correspondance* de Cicéron.

Les textes regroupés au sein de ce genre se nomment des « lettres », définies comme une conversation avec un absent dès l'époque classique<sup>122</sup>. À la fois objet de la vie quotidienne – lettre authentique ou missive réellement expédiée – et objet de fiction dans le roman épistolaire, la lettre se caractérise par « l'instabilité de ses formes et la souplesse de son utilisation » (Haroche-Bouzinac 1995, p. 3). Néanmoins, selon Adam (1998, p. 41), « la forme épistolaire présente un certain nombre de constantes compositionnelles ». Parmi celles-ci, se trouvent des éléments structurels comme la formule de clôture ou les termes d'adresse<sup>123</sup> qui participent à ce que Equoy Hutin (2006, p. 120) nomme « la structuration normative de l'aire scripturale » et qui permettent la « mise en présence 'réelle', des partenaires de l'interaction » (Landowski 1988, p. 22). À l'époque classique, trois étapes sont distinguées par rapport à la fonction qu'elles occupent dans l'économie de la lettre : prendre contact avec le destinataire, présenter et développer l'objet du message, prendre congé (Haroche-Bouzinac 1995, p. 19). L'objectif de la construction de ce corpus est de dépasser ces constantes compositionnelles afin de caractériser le genre épistolaire aux moyens d'unités phraséologiques spécifiques.

Le projet Phraséologie de la correspondance a lancé la construction du corpus PhraseoCorr en 2018, en collaboration avec le laboratoire Litt&Arts (UMR 5316), à partir des premiers manuels d'art épistolaire des 17<sup>e</sup> et 18<sup>e</sup> siècles. Elle se poursuit actuellement dans le cadre de la thèse de doctorat d'Iris Fabry débutée en octobre 2020 et cible un sous-genre spécifique, la lettre familière, et une période chronologique couvrant quatre siècles, du 17<sup>e</sup> au 20<sup>e</sup> siècle. L'intérêt de l'enquête diachronique, comme le souligne Maingueneau, est aussi de faire ressortir que « [l]es genres de discours ne sont pas des catégories intemporelles mais des réalités historiques inséparables des sociétés dans lesquelles ils émergent » (Maingueneau 1991, p. 178).

---

<sup>122</sup> Voir par exemple, Ortigue de la Vaumorière (1690) : « nous envoyons à une personne absente pour lui faire savoir ce que nous lui dirions si nous étions en état de lui parler ». (p. 7 dans *Lettres sur toutes sortes de sujets avec des avis sur la manière de les écrire et des réponses sur chaque espèce des lettres*).

<sup>123</sup> Les termes d'adresse sont des syntagmes nominaux qui « entretiennent *de facto* des relations de coréférence avec le sujet du verbe à la deuxième personne » (Lagorgette 2003, § 52). Ils occupent ainsi très largement des fonctions pragmatiques dans la mesure où ils fournissent des indications sur les types d'interaction entre les correspondants dans le domaine des relations sociales. Noailly (2005) les classe en 4 catégories : ils peuvent prendre la forme d'interpellatifs (*Excellence, Sire, Papa, Maman, Monsieur, Madame*, etc.), de noms propres, de noms relationnels (*collègue, compagnon, ami, mari, femme, neveu, cousine*, etc.) et de mots affectueux ou *kinship terms* (*trésor, amour, lapin, poulette, jolie, grande*, etc.).

C'est en 1640 que paraît *Le Secrétaire à la Mode* de Puget de la Serre, nouvelle version entièrement remaniée de son *Secrétaire à la Cour* (1925) :

[...] Le titre développé du *Secrétaire à la Mode*, « ou methode facile d'escrire selon le temps diverses Lettres de Compliment, Amoureuses et Morales », n'hésite pas à afficher une intention didactique que *Le Secrétaire de la Cour* s'appliquait à masquer, comme si le manuel, porté par son succès, pouvait désormais paraître à visage découvert. (Chupeau 1987, p. 121)

Sur le plan linguistique, la langue utilisée dans ce manuel proposant des modèles épistolaires adaptés aux besoins de la correspondance mondaine, est toujours celle du milieu du 17<sup>e</sup> siècle et la typologie a peu changé (lettres de compliment, d'excuse, de prière, de remerciement, de congratulation, de congé, d'avis, de recommandation, de nouvelles, de protestation d'amitié, de consolation, amoureuses et diverses<sup>124</sup>). La nouveauté par rapport au précédent secrétaire intervient plutôt dans la « recherche d'un nouveau protocole épistolaire tourné vers plus de simplicité et de franchise » qui se veut davantage en phase avec l'aspiration nouvelle à la sincérité et au naturel de la correspondance mondaine (Chupeau 1987, p. 122). Le *terminus a quo* de notre corpus est donc 1640. La borne opposée, le *terminus ante quem*, a été posée en 1960 pour des raisons technologiques : l'acheminement postal des lettres privées est alors de plus en plus concurrencé par le développement des télécommunications. Pendant les années 1950, ces technologiques sont en plein développement pour prendre leur essor dans les années 1960<sup>125</sup>.

Le sous-genre choisi pour ce corpus est celui de la lettre familière. Elle est ainsi qualifiée par sa destination : c'est la lettre écrite aux proches parents ou aux amis. Cette proximité du lien entre les deux correspondants autorise une plus grande liberté d'écriture que celle des lettres commerciales ou protocolaires qui sont fortement cadrées par les lois de l'échange hiérarchique (Haroche-Bouzinac 1995, p. 24). De plus, la lettre familière s'inscrit le plus souvent dans un ensemble plus vaste, la correspondance ou échange épistolaire suivi entre deux épistoliers. Nous avons choisi de retenir des correspondances et non pas des lettres isolées dans la mesure où la correspondance forme un ensemble suivi cohérent sur le plan des pratiques linguistiques.

---

<sup>124</sup> Puget de la Serre définit ces différents types de lettres dans la première partie de l'« Instruction à écrire des lettres » qui ouvre *Le Secrétaire à la Mode* (voir Lignereux 2016).

<sup>125</sup> Pour le Comité d'Histoire de la Poste, l'année 1968 marque la fin de l'âge d'or du courrier en France. À titre d'exemple, en 1965, chaque Français a en moyenne expédié 93 lettres ou cartes postales, 21 imprimés et 5 paquets quand il a reçu 120 lettres ou cartes postales et 6 paquets. Voir le document de synthèse produit par Le Comité d'Histoire de la Poste <[https://www.laposte.fr/chp/mediasPdf/cdh\\_richez\\_saison4.pdf](https://www.laposte.fr/chp/mediasPdf/cdh_richez_saison4.pdf)> (s.d.).

Afin d'obtenir une bonne représentativité de ce sous-genre, nous avons décidé de construire le corpus avec 50 % de lettres authentiques et 50 % de lettres modèles issues des manuels épistolographiques. Les documents ont été sélectionnés à différents moments (première moitié, milieu et seconde moitié du siècle) afin de représenter au mieux la langue et les pratiques de chaque siècle<sup>126</sup>. La sélection des lettres authentiques repose également sur une prise en compte équilibrée du sexe des correspondants (un homme écrit à un homme *HH* ; une femme à une femme *FF* ; un homme à une femme ou vice-versa *HF*) et de la variété des liens d'amitié ou de parenté qui les unissent (enfant/parent, cousin/cousine, frère/sœur, oncle, neveu/nièce, époux/épouse). De plus, nous avons décidé de diversifier les profils des auteurs de lettres (écrivains, scientifiques, femmes du monde, artistes, de Paris ou de Province, simples soldats, politiques).

Une fois cette grille de critères qualitatifs élaborée, le choix des titres définitifs a été finalisé dans le cadre de la thèse de doctorat d'Iris Fabry actuellement en cours à l'Université Grenoble Alpes. C'est pourquoi nous avons sélectionné quelques correspondances locales (Stendhal, Barnave, Marquis d'Adhémar, la Comtesse de Laric) afin de valoriser le patrimoine régional. La disponibilité des textes sous format numérique a également été un critère de sélection (Correspondance de Sévigné sur la base Frantext ; Correspondance de la famille d'Argenson éditée par l'Université de Poitiers ; Correspondances de Flaubert éditée par l'Université de Rouen ; Correspondance de la Marquise de Livry en cours d'édition à l'EHESS ; Correspondance de G. Monge dans le projet EMAN ; *Corpus 14* de l'Université de Montpellier 3, etc.)<sup>127</sup>. Les deux tableaux 10 et 11 présentent la liste des documents retenus pour le corpus PhraseoCorr dans chacune des deux catégories (manuels et lettres authentiques) :

---

<sup>126</sup> Nous avons conservé la répartition traditionnelle par siècle, ce qui n'empêche pas quelques enjambements au tournant d'un siècle pour certaines correspondances suivies. Dans ce cas, nous avons classé la correspondance dans le siècle où les lettres étaient les plus nombreuses. Pour Mme de Maintenon, même si ses lettres sont datées de 1704 à sa mort en 1719, nous les avons intégrées au 17<sup>e</sup> siècle dans la mesure où c'est un personnage de ce siècle (elle est née en 1635).

<sup>127</sup> Pour les textes restant à transcrire, un travail collaboratif sur la plateforme de transcription TACT (voir sitographie) a été organisé et coordonné par Iris Fabry. Il est toujours actif pour les documents des 17<sup>e</sup> et 18<sup>e</sup> siècles.

<b>Période</b>	<b>Manuels d'art épistolaire</b>
<b>17<sup>e</sup> s.</b>	J. Puget de la Serre, <i>Le Secrétaire à la Mode</i> (1646)
	B. Piélat, <i>Le Secrétaire inconnu</i> (1671)
	F. de Fenne, <i>Le Secrétaire à la mode réformée</i> (1684)
	P. Ortigue de Vaumorière, <i>Lettres sur toutes sortes de sujets</i> (1690)
	R. Milleran, <i>Lettres familières, galantes</i> (1692)
<b>18<sup>e</sup> s.</b>	C.H.L.G. Grimarest, <i>Recueil de lettres sur divers sujets</i> (1725)
	L. Philippon de la Madelaine, <i>Modèles de lettres sur différents sujets</i> (1761)
	Abbé de Vitrac, <i>Traité élémentaire du genre épistolaire</i> (1780)
	L.M.D. Chaudon, <i>Nouveau Manuel épistolaire</i> (1785)
<b>19<sup>e</sup> s.</b>	J. Carolus, <i>Le secrétaire de poche</i> (1818)
	E. Deplasan, <i>Le secrétaire français contenant les principes du style</i> (1856)
	H. Bescherelle, <i>L'art de la correspondance</i> (1865)
	Z. Carraud, <i>Lettres de famille ou modèles de style épistolaire</i> (1879)
<b>20<sup>e</sup> s.</b>	Durand et Meslins, <i>Le secrétaire pratique</i> (1913)
	Liselotte, <i>Le Guide de la Correspondance</i> (1929)
	L. Chaffurin, <i>Le parfait secrétaire</i> (1932)
	E. de Breuil, <i>Le nouveau secrétaire</i> (1957)
<b>Taille visée : 500 000 tokens par siècle</b>	

Tableau 10 : Le corpus PhraseoCorr (1) : liste des manuels d'art épistolaire sélectionnés.

<b>Période</b>	<b>Amitiés</b>	<b>Famille</b>
<b>17<sup>es</sup>.</b>	<i>HH</i> : N. Boileau & J. Racine (1687-1698)	<i>Fratrerie</i> : Marquise & Marquis de la Moussaye (1640-1663)
	<i>HH</i> : Pellisson Fontanier & Abbé Boisot (1674-1693)	<i>Fratrerie</i> : P. Bayle & J. Bayle (1662-1684)
	<i>HH</i> : V. Conrart & Félibien (1644-1650)	<i>Fratrerie</i> : N. Catinat (1690-1701)
	<i>HF</i> : Mme de La Fayette & Ménage (1640-1692)	<i>Parent-enfant</i> : Mme de Sévigné à sa fille (1671-1696)
	<i>HF</i> : Marquise de Sablé & ses amis (1660-1675)	<i>Cousins</i> : Fénelon & Marquise de Laval (1681-1695)
	<i>HF</i> : R. de Bussy-Rabutin & Mme de Scudéry (1670-1691)	<i>Variés</i> : Famille Du Guet (1683-1720)
	<i>FF</i> : Mme de Villars & Mme de Coulanges (1679-1681)	
	<i>FF</i> : Comtesse de Maure & Marquise de Sablé (1640-1622)	
	<i>FF</i> : Mme de Maintenon & Mme de Caylus (1704-1719)	
<b>18<sup>es</sup>.</b>	<i>HH</i> : J.-J. Rousseau & ses amis (1754-1770)	<i>Époux</i> : Comtesse de Laric (1774)
	<i>HH</i> : Barnave & ses amis (1792)	<i>Époux</i> : G. Monge (1762-1799)
	<i>HH</i> : Bernardin de Saint-Pierre & Hennin (1779-1787)	<i>Époux</i> : femmes d'Argenson (1727-1782)
	<i>HH</i> : Marquis d'Adhémar & ses amis (1746-1773)	<i>Fratrerie</i> : S. de Murat & Comtesse de Laric (1760-1772)
	<i>HF</i> : Julie de Lespinasse & Condorcet (1773-1776)	<i>Fratrerie</i> : Mme du Deffand (1724-1780)
	<i>HF</i> : Mme d'Épinay & ses amis (1726-1783)	<i>Fratrerie</i> : R.-T. Lindet (1789-1799)
	<i>HF</i> : Mme du Deffand & Abbé Barthélémy (1768-1771)	<i>Cousins</i> : B. Constant (1786-1799)
	<i>HF</i> : A. Lecouvreur & G. Monval (1710-1729)	<i>Parent-enfant</i> : A. Morand & son fils (1787-1801)
	<i>FF</i> : Mme de Meinières à Mme de Lénoncourt (1770-1774)	<i>Parent-enfant</i> : anthologie (1780-1799)
	<i>FF</i> : Marquise de Livry à la Présidente du Bourg (1763-1792)	
	<i>FF</i> : Mme Roland aux Demoiselles Cagnet (1767-1789)	
	<i>FF</i> : Mme de Staal à Mme la marquise Du Deffand (1747)	
<b>19<sup>es</sup>.</b>	<i>HH</i> : M. Guizot & ses amis (1810-1874)	<i>Époux</i> : É. & C. Reclus (1859-1869)
	<i>HH</i> : A. Comte & ses amis (1814-1857)	<i>Époux</i> : femmes d'Argenson (1811-1824)
	<i>HH</i> : Stendhal & ses amis (1800-1830)	<i>Fratrerie</i> : H. de Balzac & sa sœur (1819-1850)
	<i>HH</i> : É. Renan & M. Berthelot (1847-1892)	<i>Fratrerie</i> : L.F. Veuillot & ses frère et sœur (1865-1874)
	<i>HF</i> : G. Sand & G. Flaubert (1863-1880)	<i>Variés</i> : A. Morand & sa famille (1800-1823)

	<i>HF</i> : M. Desbordes Valmore & ses amis (1809-1857)	<i>Variés</i> : anthologie (1800-1889)
	<i>HF</i> : Z. Carraud & H. de Balzac (1829-1850)	<i>Parent-enfant</i> : M. Guizot & ses enfants (1810-1874)
	<i>HF</i> : Rachel & G. d'Heylli (1840-1854)	<i>Parent-enfant</i> : Mme Thiers & sa mère (1841-1865)
	<i>FF</i> : É. Guérin & L. de Bayne (1835-1847)	<i>Oncle-nièce</i> : Lamartine & ses nièces (1842-1854)
	<i>FF</i> : F. Tristan & ses amies (1821-1844)	
	<i>FF</i> : M. Desbordes Valmore & ses amies (1809-1857)	
	<i>FF</i> : C. de Gaix (1817-1847)	
<b>20<sup>e</sup> s.</b>	<i>HH</i> : J.-É. Blanche & M. Denis (1901-1939)	<i>Époux</i> : B. Vian & sa femme (1941-1955)
	<i>HH</i> : J. Giraudoux & L. Jouvét (1927-1942)	<i>Époux</i> : G. Hyvernaud & sa femme (1939-1940)
	<i>HH</i> : R. Martin du Gard & ses amis (1896-1925)	<i>Fratricie</i> : S. Weil & son frère (1932-1943)
	<i>HH</i> : M. Ravel & ses amis (1905-1930)	<i>Fratricie</i> : F. Ponge (1899-1919)
	<i>HF</i> : M. de Régnier & E. Jaloux (1907-1939)	<i>Oncle-nièce</i> : F. Poulenc & B. Manceaux (1914-1960)
	<i>HF</i> : M. Audoux & ses amis (1904-1936)	<i>Variés</i> : corpus 14 (1914-1918)
	<i>HF</i> : A. Monnier & H. Michaux (1939-1955)	<i>Variés</i> : anthologie (1914-1919)
	<i>FF</i> : M. Yourcenar & ses amies (1924-1960)	<i>Parent-enfant</i> : F. Ponge (1899-1919)
	<i>FF</i> : T. Jolas et C. Meyer (1956-1960)	<i>Parent-enfant</i> : M. Jouhandeau & son fils (1908-1936)
	<i>FF</i> : Colette & L.-A. Moreau / H. Jourdan-Morhange (1929-1954)	<i>Parent-enfant</i> : M. Curie & ses filles (1905-1934)
	<i>FF</i> : F. Dolto & ses amies (1938-1960)	
	<b>Taille visée : 500 000 tokens par siècle</b>	

Tableau 11 : Le corpus PhraseoCorr (2) : liste des correspondances authentiques sélectionnées.

Parmi ces lettres authentiques, nous disposons de correspondances suivies parfois sur une seule année (Barnave), le plus souvent sur plusieurs (jusqu'à 40 ans pour celle de la Marquise de Livry avec la Présidente du Bourg). Nous prévoyons de rendre librement interrogeable ce corpus diachronique ambitieux par le Lexicoscope à la fin de la thèse d'Iris Fabry.

Pour conclure sur ces trois exemples de corpus, je soulignerai à la fois les points communs qui interviennent dans la constitution de tous les corpus écrits et la nécessaire réflexion à mener sur les particularités de chacun d'entre eux. En effet, construire un corpus représentatif dépend étroitement des phénomènes que l'on souhaite y observer. Dans le cas de la caractérisation d'un genre textuel par les unités phraséologiques spécifiques qu'il renferme, le choix des échantillons repose essentiellement sur des critères externes comme la thématique (un sous-genre précis), le profil des auteurs (auteurs majeurs ou mineurs, sexe, classe sociale, etc.) mais

aussi, comme pour tout corpus écrit, sur des critères pratiques comme la disponibilité des documents en format numérique ou le support de publication (manuscrit ou édition imprimée). Dans le cas de corpus diachroniques, ceux-ci doivent en outre se mettre au service de l'enjeu premier de la linguistique diachronique qui est, comme le rappelle Prévost (2015, p. 28), la mise au jour et l'analyse de l'émergence, de la disparition et de la transformation de constructions. Pour rendre compte de ces différents types de changement, les corpus diachroniques permettent l'établissement de fréquences, quel que soit l'état de langue considéré et offre la précieuse possibilité de décrire finement une langue qui n'a plus de locuteurs.

Dans ce chapitre 5, j'ai présenté la linguistique de corpus comme une discipline dont l'objectif premier est de mettre en évidence les choix linguistiques des locuteurs et les phénomènes discursifs qui caractérisent des textes dans le cadre d'un genre et d'une période chronologique donnés. Comme je ne situe pas mes travaux scientifiques dans une approche sémantique consacrée exclusivement au sens des unités lexicales, j'ai fait le choix de construire mes corpus sur la base d'échantillons sans écarter systématiquement le texte intégral. Je peux ainsi observer leurs usages au sein d'un genre ou d'un sous-genre textuel pour le caractériser. Ce travail permet de faire émerger du corpus le sens des textes qui le composent dans la lignée des recherches en sémantique textuelle. C'est d'ailleurs l'intuition de Tognini-Bonelli (2001) quand elle établit un parallèle entre la linguistique de corpus et la linguistique du texte : « corpus linguistics starts from the same premises as text-linguistics in that text is assumed to be the main vehicle for the creation of meaning. » (p. 2).

La construction des trois corpus présentés comme une mise en pratique des principes théoriques exposés précédemment permet de souligner que la linguistique de corpus est bien orientée vers une application. Cette application, c'est l'outillage de nos corpus en vue de leur interrogation automatisée qui en permet la mise en oeuvre. Le sixième et dernier chapitre de ce mémoire propose de faire un rapide point sur cet outillage et tout particulièrement sur les nouveaux observables phraséologiques qu'il a permis de faire émerger des corpus.

## Chapitre 6. L'émergence de nouveaux observables phraséologiques dans les corpus outillés

Quand elle définit la linguistique de corpus, Tognini-Bonelli (2001) rappelle que celle-ci est intrinsèquement liée à l'usage des « nouvelles technologies » : « a *empirical approach* to the description of *language use*; it operates within the framework of a *contextual and functional theory of meaning*; it makes use of the *new technologies*. » (p. 2). Mais quelle place leur accorder exactement ? Les uns, dans la lignée de Leech (« the new master is the computer » 1992, p. 105), pensent, dans une perspective philosophique et de manière très enthousiaste, que les nouvelles technologies de l'informatique ont permis d'ouvrir de nouvelles manières de penser la langue<sup>128</sup> ; d'autres, plus circonspects, comme Rastier (2005a, p. 41), en restent à l'idée que « Pour la linguistique de corpus, l'informatique n'est qu'un instrument, non un modèle théorique ». Pour ma part, je situe mes travaux plutôt dans cette seconde approche pratique qui pense l'outil informatique comme un facilitateur et un accélérateur de la recherche : il permet de saisir plus de données et plus vite et ainsi de faire émerger de nouveaux observables linguistiques et d'élaborer de nouveaux modèles théoriques pour en rendre compte.

L'outillage est conçu comme un apport d'informations qui transforme le texte brut en texte enrichi : « no longer the 'raw' (or pure) corpus which was originally input to the computer but a version in which linguistic information, of particulare kinds, is extensively provided. » (Leech 1991, p. 19). Grâce à cet enrichissement, le linguiste peut obtenir, par exemple, des indications en termes de fréquence et établir des relations statistiques entre ensembles de faits linguistiques (Jacques 2005, p. 25). Néanmoins, Tognini-Bonelli (2001, p. 73, à la suite de Sinclair) pointe l'appauvrissement auquel peut paradoxalement mener cet enrichissement (« loss of information ») : « Sinclair points out that the replacement of a text by a string of tags is a reduction of information; for example, words which are different but are allocated the same word class lose this distinctiveness in favour of the recognition that they belong to the same word class. ». La manière d'outiller un corpus doit donc faire l'objet d'une réflexion en lien avec l'utilisation qui en est attendue.

Dans ce dernier chapitre, mon objectif est d'expliquer pourquoi et comment j'ai exploré des corpus outillés dans mes travaux de recherche (6.1) avant de montrer l'apport de l'outil

---

<sup>128</sup> Leech (1992) évoque alors « a new philosophical approach to the subject [...in which the computer is doing] much more than acting as a research tool; it was going to open up new ways of thinking about language » (p. 105-106).

Lexicoscope dans la réflexion sur la caractérisation des genres textuels par les unités phraséologiques contenues dans les textes (6.2).

### **6.1 Explorer des corpus outillés**

Comme l'affirment Poudat & Landragin (2017), les corpus sont outillés pour permettre leur exploration. L'outillage est donc conçu comme une approche méthodologique du corpus :

De même qu'il existe une méthodologie pour constituer un corpus [...], il existe une méthodologie pour explorer un corpus. Cette méthodologie a pour objet de décrire comment appréhender les données textuelles regroupées dans le corpus. Elle inclut la recherche de mots, d'expressions, avec des possibilités beaucoup plus nombreuses que celles offertes par un logiciel de traitement de texte. Elle intègre également la méthodologie de l'annotation – on explore un corpus en l'annotant – et celles de la linguistique de corpus et des statistiques textuelles, qui fournissent des indicateurs (numériques, graphiques, multidimensionnels) pour aider l'analyste, linguiste ou non, à mieux comprendre les données constituant son corpus de travail, à mieux caractériser celui-ci, à mieux dégager les spécificités. » (p. 9)

C'est l'outillage qui rend les corpus textuels prêts à être analysés par le linguiste<sup>129</sup>. Sa plus-value est de faire ressortir des faits saillants dont le linguiste pourra s'emparer pour tester la validité de ses hypothèses ou en formuler de nouvelles. L'outillage n'est donc pas une fin en soi comme le rappelle Aarts (2000, p. 7) qui alerte sur le danger du 'number crunching' : « studies that are exclusively concerned with the presentation of frequency data as a goal in itself without couching those data within the framework of meaningful questions about the structure or usage of the language being studied. [...] It is the type of work that invariably elicits a 'so what' response. ». On outille donc un corpus mais pour quoi faire exactement ?

Dans cette section, j'ai choisi d'envisager les deux extrêmes en matière d'outillage de corpus : la fonctionnalité de concordancier qui en est, en quelque sorte, le degré zéro (6.1.1) et l'interface de fouille textuelle extrêmement sophistiquée qu'est le Lexicoscope avec ses corpus annotés en dépendances syntaxiques (6.1.2).

---

<sup>129</sup> On retrouve cette même idée dans la définition que Tognini-Bonelli (2001, p. 55) donne du corpus : « a corpus is taken to be a computerised collection of authentic texts, amenable to automatic or semi-automatic processing or analysis. ».

### 6.1.1. Le degré zéro de l'outillage : le concordancier

Mes premiers travaux de recherche sur les langues anciennes ont abondamment utilisé les concordanciers. Avant d'exister sous la forme numérique que nous connaissons par exemple dans la bibliothèque numérique *Thesaurus Linguae Graecae* (TLG)<sup>130</sup> pour les textes du corpus grec, les premiers concordanciers ont été conçus en version papier :

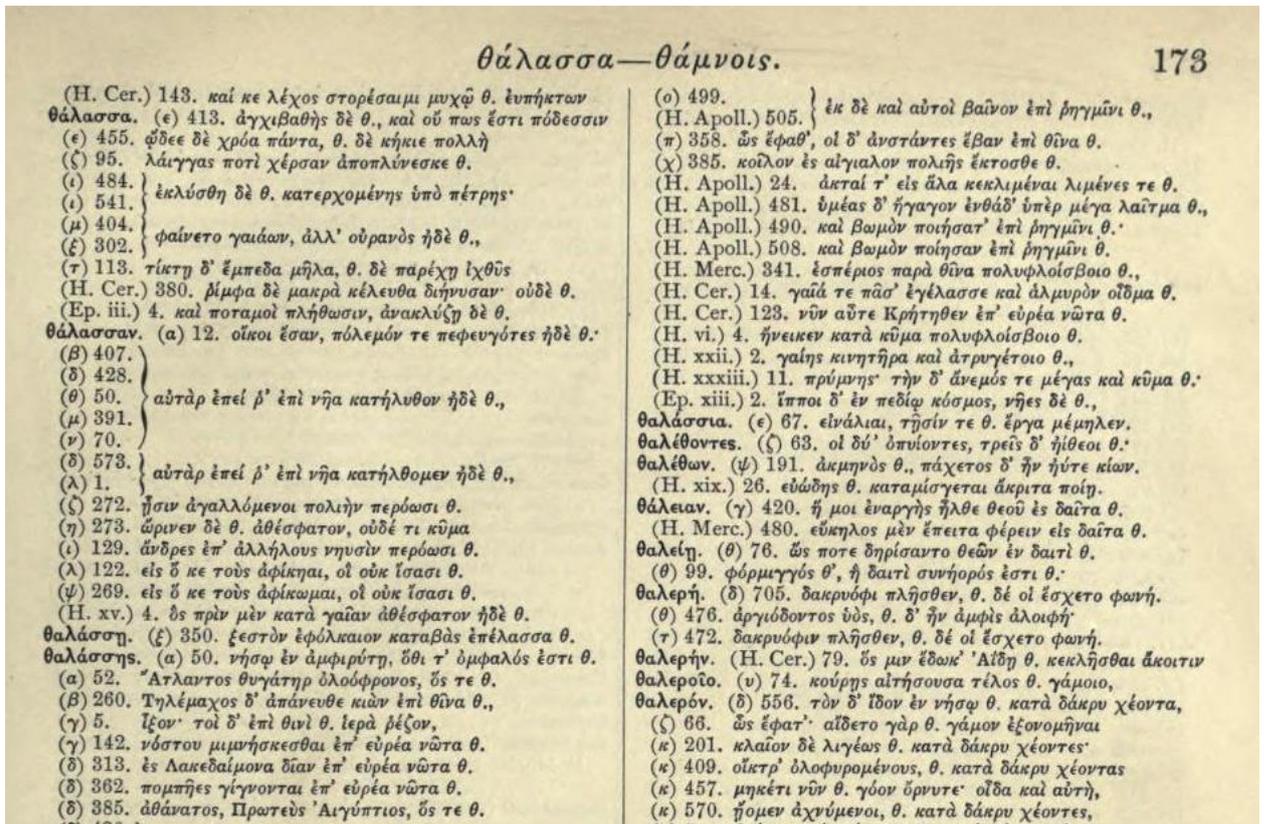


Figure 5 : Extrait de H. Dunbar (1880), *A Complete Concordance to the Odyssey and Hymns of Homer*, Oxford Clarendon Press (p. 173).

Dans sa version papier, le concordancier présente la liste de toutes les formes, classées par ordre alphabétique, d'un texte ou d'un corpus et indique les coordonnées permettant de les localiser ainsi que le contexte d'apparition de chaque occurrence. Dans la figure 5, l'accusatif θάλασσαν « mer » est attesté à quatorze reprises dans le corpus : treize fois dans *L'Odyssee* et une fois dans l'*Hymne xv*. Pour les occurrences de *L'Odyssee*, les coordonnées fournies sont le numéro de chant (en minuscule grecque α = 1, etc.) suivi du numéro de vers. Le contexte fourni est celui du vers entier avec la forme abrégée par son initiale (θ. pour θάλασσαν). Une première

<sup>130</sup> Le TLG est un projet de recherche pionnier de l'Université de Californie (Irvine), qui a débuté en 1972. Il représente une bibliothèque numérique accueillant l'ensemble de textes littéraires écrits en grec des origines (8<sup>e</sup> siècle avant n.è.) à la chute de Byzance (1453).

lecture permet de repérer si la forme apparait dans un vers formulaire que l'on retrouve à l'identique dans plusieurs endroits de l'œuvre (par ex., le vers formulaire αὐτὰρ ἐπεὶ ῥ' ἐπὶ νῆα κατήλυθον ἠδὲ θ. que l'on retrouve dans cinq chants différents).

Tribble (2010, p. 167-168) rappelle que le premier concordancier attesté dans le monde occidental serait celui réalisé, au 13<sup>e</sup> siècle, par des centaines de moines dominicains du couvent Saint-Jacques à Paris, sous la supervision du Cardinal Hugues de Saint-Cher, dans l'objectif de recenser tout le vocabulaire de la Vulgate. Cet exemple illustre bien l'extrême lourdeur de la réalisation de tels ouvrages dont la forme la plus simple est celle de l'index comme le rappelle Sinclair (1991) :

A concordance is a collection of the occurrences of a word-form, each in its own textual environment. In its simplest form it is an index. Each word-form is indexed and a reference is given to the place of occurrence in a text. (p. 32)

C'est ce principe même que nous retrouvons mis en œuvre, après automatisation, dans la bibliothèque numérique *Thesaurus Linguae Graecae* :

The screenshot shows the TLG website interface. At the top, there is a navigation bar with links like HOME, CANON, TEXT SEARCH, BROWSE, LEXICA, N-GRAMS, STATISTICS, VOCAB. TOOLS, and HELP. The main search area is titled 'SEARCH THE TLG CORPUS' and includes options for 'SIMPLE', 'PROXIMITY', and 'ADVANCED PROXIMITY'. The search term 'qalassan' is entered, and the results are displayed as a list of occurrences. The search count is 13. The results are sorted by Author # and displayed in Greek. The first result is from Book 1, line 12, and the subsequent results are from Book 2, line 407, Book 4, line 428, and Book 4, line 573. The interface also includes a 'SELECTION' panel on the left and a 'MY SEARCH SELECTION' panel on the right.

Figure 6 : Concordance de θάλασσαν dans *L'Odyssee* (TLG).

Dans cette fenêtre, la forme est surlignée en jaune et elle est insérée dans son contexte constitué de trois vers (le vers précédant et le vers suivant celui dans lequel elle apparaît). La localisation de la forme est indiquée, en caractères rouge et gras, en tête du groupe de vers. Il s'agit ici d'une concordance élargie (KWIC ou *key word in context*) dans laquelle la forme se trouve dans un contexte suffisamment large pour pouvoir interpréter son usage dans le corpus (Poudat & Landragin 2017, p. 187).

La plus-value de l'automatisation est indéniable pour le chercheur : non seulement, il n'a plus besoin de recenser manuellement toutes les occurrences d'une forme par la lecture intégrale des textes du corpus mais encore les liens hypertextes lui permettent d'accéder en un clic à un contexte encore plus élargi (une vingtaine de vers dans le cas de *L'Odyssée* par exemple), de circuler ensuite dans le texte à sa guise voire d'accéder à une ressource lexicographique en ligne comme c'est le cas dans la base textuelle *Perseus Digital Library* (voir sitographie) qui s'appuie, pour les textes gréco-latins, sur une partie du fonds du *TLG*. Cet outillage lui permet de gagner un temps précieux dans le repérage des formes et de limiter fortement le risque d'oublier des occurrences.

Dans le cas d'une étude en diachronie longue, l'intérêt pour le linguiste est encore plus marqué. La consultation du *TLG*, qui recense les productions littéraires en grec s'étendant sur une période de vingt-deux siècles – soit un corpus de 10 315 documents pour un total de 120 623 284 mots – m'a permis d'identifier précisément tous les contextes d'apparition de la lexie ἡ ῥηγμῖν « brisant, rivage » (Sorba 2014b) et de lancer une investigation sur le devenir de cette lexie homérique et de sa combinatoire. Comme « les concordances facilitent l'étude détaillée des comportement syntaxiques » (Habert et *al.* 1997, p. 133), j'ai pu établir le profil lexico-syntaxique de ce mot rare qui survit, de manière artificielle, tout au long de la période, dans l'oeuvre des lexicographes. J'ai pu distinguer deux types d'emploi en fonction du genre textuel : d'une part, le maintien des constructions au datif en SP héritées de l'épopée homérique (ἐπὶ / παρὰ ῥηγμῖνι « sur le / le long du rivage ») dans les textes poétiques (Apollonios, Pindare, etc.) ; d'autre part, des innovations syntaxiques dans les textes scientifiques (par ex., un nominatif pluriel, αἱ ῥηγμῖνες, avec déterminant « les brisants » dans les *Météorologiques* d'Aristote).

Pour le corpus constitué de romans en prose du 13<sup>e</sup> siècle, les textes sont lemmatisés avec l'outil *LGeRM*, un lemmatiseur initialement conçu pour traiter le moyen français (Souvay & Pierrel, 2009). En travaillant avec son créateur (Gilles Souvay), nous l'avons progressivement

ajusté pour le traitement de ces textes d'ancien français<sup>131</sup>. Cette interface propose classiquement la fonctionnalité concordancier comme la figure 7 le montre :

**atif** ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE

Projet  
Sortie de lemmatisation  
Exploitation des résultats  
Présentation  
Recherche de formes  
Parcours des mots du texte  
Édition en ligne  
Administration du projet  
Utilisateur : julie.sorba  
Projet : GRE\_ArtusdeBretag...

Finalisation Recherche Texte Nomenclature Formes Concordancier Lemmes Étiquettes Gloses Lexique  
Parcours du texte

### Recherche sur les formes

■ Travail sur la forme *chevalier*

■ Résultats

Remise à zéro Appliquer les critères

cotexte gauche (facultatif) pivot (obligatoire) cotexte droite (facultatif)

A Z [-----] forme [-----] A Z [-----]

négation joker 0  négation joker 0

sensible casse  sensible casse

Nombre d'attestations : 455

[1]	mot 44	, Lancelot et maint autre	<b>chevalier</b>	preu et bon,
§ 1				
[2]	mot 86	et haut lignage du vaillant	<b>chevalier</b>	Lancelot, qui tant
§ 1				
[3]	mot 725	de la Terre Deserte I	<b>chevalier</b>	de grant pooir et
§ 4				
[4]	mot 836	lui d'estre le meilleur	<b>chevalier</b>	du monde. Mes
§ 4				
[5]	mot 906	esforcielement sus les biens du	<b>chevalier</b>	et de la dame
§ 4				
[6]	mot 1504	, sire, a un	<b>chevalier</b>	que on apeloit monseigneur
§ 5				
[7]	mot 1516	seigneur d'icelui chastel et	<b>chevalier</b>	fort et puissant,
§ 5				
[8]	mot 3078	roy Artus ot moult bon	<b>chevalier</b>	, si voudroie que
§ 7				
[9]	mot 3862	nichement tenu compagnie a I	<b>chevalier</b>	dont ele est acointe
§ 9				
[10]	mot 4485	s'est meffaitte a I	<b>chevalier</b>	! Me tenéz vous
§ 11				
[11]	mot 4798	et pleniere et feroit Artus	<b>chevalier</b>	et le marieroit.
§ 11				
[12]	mot 5357	, li baron et li	<b>chevalier</b>	, pour aler encontre
§ 12				
[13]	mot 5848	Hector mon cousin soit demain	<b>chevalier</b>	avecques moi. Si
§ 13				
[14]	mot 5880	poi de terre pour devenir	<b>chevalier</b>	. » Et ce
§ 13				

Figure 7 : Concordance du pivot *chevalier* dans *Artus de Bretagne* (LGeRM)

<sup>131</sup> Holgado, Lavrentiev & Constant (2021) ont d'ailleurs pu montrer les performances élevées de cet outil (déjà ancien) pour la lemmatisation du français médiéval qu'ils expliquent par « l'importance de son lexique (indépendant du corpus d'apprentissage) et le fait que les règles de substitution permettent le plus souvent de retrouver une forme attestée dans le lexique. » (p. 158). Voir Bazin-Tacchella & Souvay (2021) pour une présentation actualisée de l'outil. L'étape de lemmatisation avec LGeRM est nécessaire pour désambigüiser le maximum de formes avant l'annotation du corpus en dépendances syntaxiques en vue de son intégration dans le Lexicoscope. L'équipe, renforcée d'une stagiaire (Jingyu Liu, M2 Sciences du Langage – Parcours Industries de la langue), a réalisé un premier travail sur la fusion des sorties de LGeRM et du parseur HOPS (Grobol & Crabbé 2021), en collaboration avec l'institut MIAI et le laboratoire LIG, que nous présentons dans Coavoux et al. (2022). Un premier corpus-test *PhraseoMedieval* a été mis en ligne fin 2021 dans le Lexicoscope avec un total de 457 269 tokens (avec des échantillons des romans en prose *Artus de Bretagne*, *Lancelot*, *La Quête del saint Graal*, *Merlin*. Ce premier corpus-test inclut en outre deux romans en vers *Aucassin et Nicolette* ; *Le Roman de la Rose* (source : SRCMF) qui seront basculés dans le corpus de contraste).

La concordance en ligne de *LGeRM* se présente sous une forme plus usuelle (KWAC ou *key word and context*, pour reprendre le jeu de mots de Pincemin 2006) qui met en valeur dans l’affichage le **pivot** autour duquel sont alignés les contextes gauche et droit. Ce type de concordance est un des outils privilégiés de la linguistique de corpus dont « la puissance repose sur le double alignement vertical (du pivot) et horizontal (des lignes contextuelles) » (Poudat & Landragin 2017, p. 186).

En bref, même si c’est l’outillage le plus commun (et le plus ancien) pour interroger les corpus, la concordance garde tout son intérêt en offrant « un mode de visualisation simple et puissant, ce qui en fait un outil heuristique particulièrement utile et efficace pour explorer une unité linguistique en contexte. » (Poudat & Landragin 2017, p. 185). En complément de ce premier stade, l’exploration des corpus se poursuit à l’aide d’outils plus complexes tel le Lexicoscope.

### 6.1.2 La fouille de corpus élaborée : le Lexicoscope

Créé dans le cadre du projet EMOLEX, le Lexicoscope est un outil consacré à l’étude de la combinatoire lexico-syntaxique pour l’analyse de corpus arborés<sup>132</sup> multilingues contemporains (Kraif & Diwersy 2012 ; Diwersy et al. 2014). Il a été régulièrement développé depuis 2010 :

[...] le Lexicoscope réunit des fonctionnalités complémentaires de concordance, de recherche d’expressions complexes et de tableaux de cooccurrents – la richesse des outils sous-jacents restant toutefois accessible pour l’utilisateur non spécialiste grâce à des principes ergonomiques originaux. (Kraif 2016, p. 91).

À côté de la fonctionnalité classique de concordancier, le Lexicoscope présente d’autres fonctionnalités élaborées pour l’extraction des cooccurrences à partir d’un pivot donné pour en caractériser la combinatoire. Cette extraction repose sur le concept de *lexicogramme* (Heiden & Tournier 2001) : « il s’agit d’établir, pour un pivot donné, la liste de ses cooccurrents les plus fréquents, à gauche et à droite, en faisant l’extraction des fréquences de cooccurrence et en calculant des mesures d’association statistiques (telles que le rapport de vraisemblance ou le t-

---

<sup>132</sup> Les corpus ont été annotés en dépendances syntaxiques à l’aide de différents analyseurs selon la version du Lexicoscope et les langues ou états de langue concernés. L’intérêt de ce type d’annotations est de réduire le bruit et le silence par rapport au repérage des simples cooccurrences de surface (Kraif 2016, p. 93-94). Par exemple, dans la première version du Lexicoscope, *XIP* a été utilisé pour le corpus anglais et *Connexor* pour l’allemand, le français et l’espagnol (Kraif & Diwersy 2012, p. 400). Pour le corpus des prosateurs latins (César, Cicéron, Sénèque le Père, Sénèque, Salluste, Tacite, Tite-Live), nous avons entraîné trois modèles avec *MaltParser* (Autin et al. 2016). Nous testons actuellement le parseur *HOPS* pour les nouveaux corpus (Grobol & Crabbé 2021).

score). » (Kraif & Diwersy 2012, p. 400). Le modèle mis en œuvre est « un modèle de cooccurrence flexible permettant à l'utilisateur de définir lui-même les *unités de cooccurrences* : formes, lemmes, catégories morphosyntaxiques, traits additionnels (p.ex. sémantiques), relations syntaxiques (dans le cas des *colligations*) ou des combinaisons de ces informations. » (*ibid.*). L'espace de cooccurrence ou espace de rencontre entre le pivot et son collocatif peut également être caractérisé de manière flexible car l'utilisateur peut déterminer la largeur de l'empan ou le type de relation fonctionnelle entre les deux. Les deux figures 8 et 9 ci-dessous permettent de visualiser les deux représentations possibles de ce lexicogramme dans le Lexicoscope :

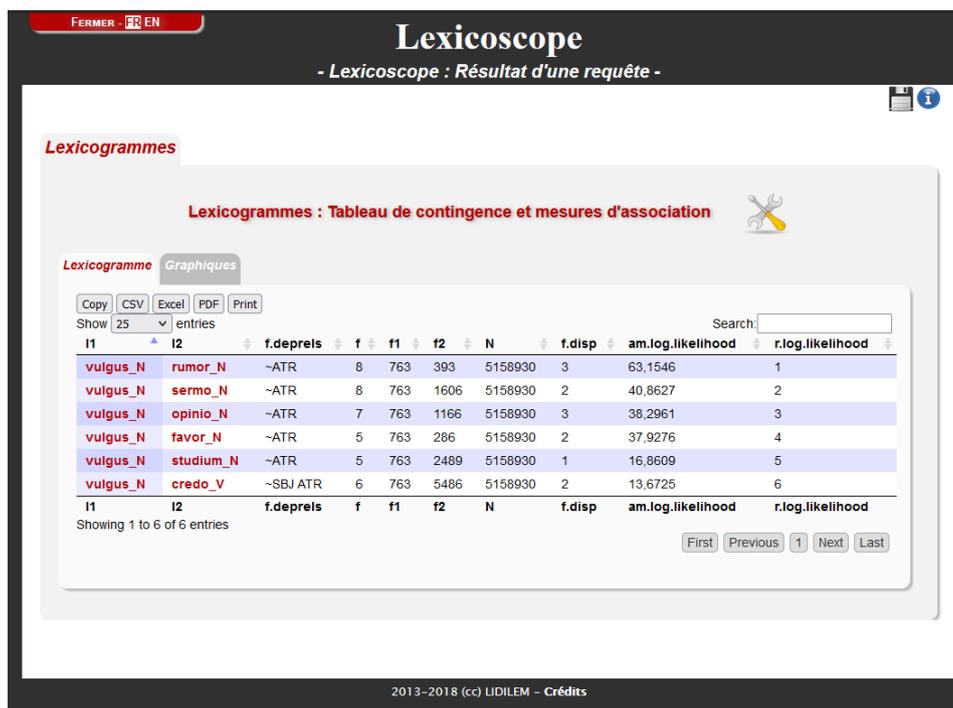


Figure 8 : Lexicogramme du pivot latin *vulgus* (tableau de contingence).

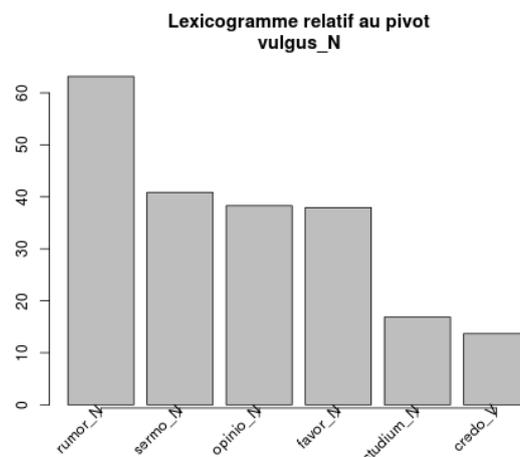


Figure 9 : Lexicogramme du pivot latin *vulgus* (graphique).

Le tableau de contingence de la figure 8 permet de faire figurer les cooccurrents syntaxiques (12 *rumor, sermo, opinio, etc.*) du pivot (11 *vulgus*) par ordre de pertinence décroissante. Celle-ci est quantifiée par la mesure d'association nommée *log likelihood ratio* (ou rapport de vraisemblance ou LLR)<sup>133</sup>. Ce tableau fournit d'autres informations comme les relations de dépendances syntaxiques qui entrent en jeu (f.deprels) et la dispersion de la cooccurrence indiquant le nombre de sous-corpus dans lequel elle apparaît (f.disp). La représentation graphique de ce tableau (figure 9) propose une visualisation sous la forme d'un histogramme de la spécificité des collocatifs (le *log likelihood* sur l'ordonnée, le collocatif sur l'abscisse).

Dans notre étude précisément consacrée à l'analyse de la combinatoire lexico-syntaxique des noms collectifs de la 'foule' en latin *vulgus, multitudo* et *turba* (Autin et al. 2016), il est apparu que le profil combinatoire de *vulgus* se distinguait nettement de celui des deux autres car il apparaît dans « des structures à forte valeur idéologique » (p. 6). En effet, d'une part, cette lexie se combine de manière préférentielle avec des noms abstraits (*rumor, sermo, opinio, favor, studium*) comme le montre le tableau de contingence (figure 8). D'autre part, l'étude des collocatifs indique que *vulgus* désigne systématiquement une foule comme une collection d'êtres animés (les spectateurs dans l'exemple 38) alors que *multitudo* et *turba* peuvent également renvoyer à des collections d'entités inanimés (les armes de jets dans l'exemple 39) :

38) Et ludicro circensium, quod acquirendis **uulgi studiis** edebatur, Britannicus in praetexta, Nero triumphali ueste traecti sunt. (Tacite, *Annales* 12.41). « Et lors des jeux du cirque que l'on donnait pour capter *l'affection de la foule*, Britannicus défila en toge prétexte, Néron avec un vêtement triomphal. »

39) **Multitudine telorum** ex turribus propugnantes deturbant, aggere et cratibus fossas explent, falcibus vallum ac loricam rescindunt. (César, *Bellum Gallicum* 7.86) « Ils chassent *par une foule de traits* ceux qui combattaient du haut des tours. »

Le lexie *vulgus* désigne, de manière très nette, la foule des humains qui pensent, qui parlent et qui s'émeuvent alors que *turba* et *multitudo* s'associent plus volontiers avec des verbes de mouvement (*circumfundo, concito, effundo, convenio, traduco* etc.).

---

<sup>133</sup> Cette mesure d'association est analogue à un *Chi2* et « exprime l'invraisemblance d'obtenir, par le seul jeu du hasard, le tableau de contingence (résumé par la fréquence de cooccurrence *f*, la fréquence du pivot *f1*, la fréquence du collocatif *f2*, et la taille de l'espace de cooccurrence *N*). Des valeurs seuils, modifiables par l'utilisateur, permettent de ne filtrer que les cooccurrences dépassant une certaine fréquence ou un certain degré d'association » (Autin et al. 2016, p. 4).

Les extractions de cooccurrences fournies par le Lexicoscope permettent de mettre au jour les points de rencontre entre lexique et grammaire au sein des corpus. C'est ce que Hunston & Francis (2000) ont développé sous la notion de *pattern* au sein de l'approche *corpus-driven* :

The pattern of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combinaison of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it. (p. 37)

Dans le cadre du projet PhraseoRom, une nouvelle version du Lexicoscope a été développée, toujours avec le même objectif de simplicité pour l'utilisateur qui peut effectuer un requêtage à partir d'un exemple :

Sur le modèle des systèmes de recherche d'information, on peut partir d'une requête sans formalisme (une expression telle que « citer en exemple »), que l'on peut ensuite affiner par expansion proposition de requêtes équivalentes ou plus générales, extension) ou par analogie (recherche de résultats similaires). [...] Quand l'utilisateur entre une chaîne de caractères, notre système fouille le corpus afin de trouver toutes les occurrences qui « matchent » avec cette chaîne. » (Kraif 2019, p. 76)

Cette nouvelle interface, Lexicoscope 2.0, propose une interface d'interrogation plus ergonomique et des nouveautés comme par exemple, la visualisation récapitulative du *wordsketch*<sup>134</sup> pour un pivot donné :

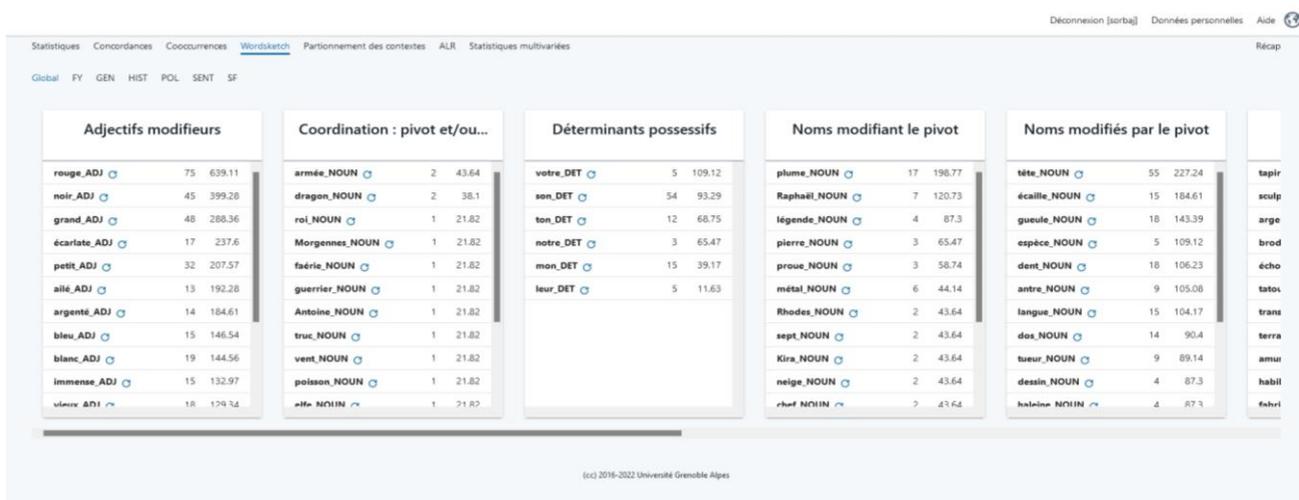


Figure 10 : Lexicoscope 2.0 : *wordsketch* du pivot *dragon* dans le corpus PhraseoRom.

Le *wordsketch* rassemble toutes les informations sur les collocatifs préférentiels du pivot par catégorie morpho-syntaxique (adjectif modifieur, coordination, déterminant possessif, nom,

<sup>134</sup> Sur ce point, voir Walter (2010) : « A *Word Sketch* groups collocates according to their grammatical relation to the search word. » (p. 435-436). Cette invention a vu le jour dans le *Sketch Engine* de Kilgarriff et al. (2004).

etc.) que l'utilisateur peut visualiser de manière synoptique. Ainsi, dans l'ensemble du corpus romanesque français contemporain PhraseoRom, on peut constater que le pivot *dragon* s'associe à de nombreux adjectifs de couleur (*rouge, noir, écarlate, argenté, bleu, blanc*) :

- 40) Le héros, Olorín, se préparait à entrer dans la caverne d'un grand **dragon rouge**.  
(R. Michel, *La Malerume 2 Le dire des Sylfes*, 2003).
- 41) Un grand ver de lave s'y enroulait autour du corps musculeux d'un **dragon noir**. Un dragon, emblème de son père, créature légendaire qui avait disparu depuis des siècles.  
(C. Debien, *Le Cycle de Lahm 1 L'éveil du roi*, 2010)
- 42) Nous ne sommes que deux dans ce monde, mais là d'où nous venons, il y en a des milliers, répondit le **dragon bleu**. (A. Robillard, *Les chevaliers d'émeraude 11 La justice céleste*, 2007)

Cette visualisation du *wordsketch* apporte des informations complémentaires à celle du lexicogramme (figure 11 ci-dessous) :

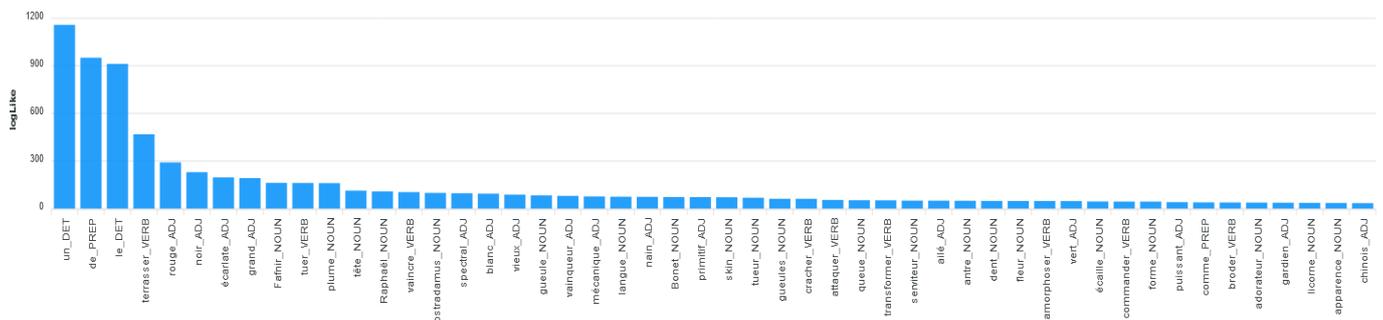


Figure 11 : Lexicoscope 2.0 : lexicogramme du pivot *dragon* dans le corpus PhraseoRom

En effet, cette première consultation du *wordsketch* de *dragon* et des concordances associées nous (Sorba & Novakova 2022 sous presse) a mis sur la piste de deux sous-genres romanesques spécifiques, la fantasy et la science fiction, dans lesquelles le dragon est un personnage décrit précisément (grâce aux noms modifiant le pivot ou modifiés par lui : *plume, tête, écaille, gueule, dent, dos*, etc.) et qui joue un rôle plutôt actif comme le montrent les nombreux verbes d'action dont le pivot *dragon* est sujet (*attaquer, défoncer, persécuter, semer, avaler, terroriser*, etc.). La consultation du lexicogramme, dans un second temps, nous a indiqué quelles unités étaient les plus spécifiques sans catégorisation préalable (par ex., le verbe *terrasser* est le premier collocatif lexical par ordre de pertinence décroissante). Ensuite, par le retour au texte grâce au contexte élargi, nous montrons sur le plan linguistique comment la lexie *dragon* occupe un rôle central dans la caractérisation même de ce sous-genre romanesque, en pleine émergence éditoriale depuis le dernier tiers du 20<sup>e</sup> siècle (Baudou 2005). La forte variation paradigmatique des collocatifs verbaux s'explique par la mise en scène des différentes phases

du combat contre le dragon : *affronter, piéger* (phase initiale) / *combattre* (phase médiane) / *terrasser, vaincre, tuer, détruire, décapiter, abattre, éliminer, exterminer le dragon* (phase finale)<sup>135</sup>.

Dans cette section, la présentation du Lexicoscope dans ses deux versions m'a permis de montrer non seulement le fonctionnement de l'outil mais également son apport dans mes études linguistiques. D'une part, il permet d'établir le profil combinatoire lexico-syntaxique de lexies, très utile par exemple pour distinguer des synonymes (par ex., *vulgus, turba, multitudo* en latin) ; d'autre part, il offre des éléments précieux pour la caractérisation des genres ou sous genres textuels (par ex., le dragon dans la fantasy). C'est d'ailleurs ce dernier point qui a été au cœur du récent projet PhraseoRom et que je vais aborder dans cette ultime section.

## **6.2 Étudier la phraséologie sur corpus outillé pour caractériser les genres textuels**

Mon travail de recherche sur la manière dont la phraséologie permet de caractériser les genres et sous-genres textuels vise à contribuer à « une théorie opératoire des genres » que Rastier appelle de ses vœux (Rastier 2011, p. 72). En effet, dans la lignée de Siepmann (2016), je considère que « la surreprésentation statistiquement significative de certains phraséologismes pourrait jouer un rôle non négligeable dans la construction littéraire du texte » (p. 22). Cette conception repose sur le fait de considérer la cooccurrence comme un aspect central de la textualité à l'instar de Viprey (2006). Ces travaux sur la nature lexico-grammaticale des textes littéraires sont actuellement en plein essor (par ex., Legallois, Charnois & Poibeau 2016 ; Diwersy & Legallois 2019) et bénéficient des nombreuses recherches, en linguistique de corpus outillé, effectuées jusqu'alors sur des textes relevant d'autres genres textuels, scientifiques (par ex., Poudat 2006 ; Tutin 2014 ; Jacques & Tutin 2018) ou professionnels (par ex., Née, Sitri & Véniard 2014).

C'est d'ailleurs précisément dans le cadre des écrits scientifiques que Tutin & Kraif (2016) ont montré que les méthodes fondées sur des corpus arborés permettaient de détecter des unités phraséologiques spécifiques à fort potentiel heuristique dans le cadre de la problématique de caractérisation des genres textuels (6.2.1). Je présenterai ensuite comment nous avons utilisé ces méthodes pour distinguer des sous-genres romanesques sur la base d'une nouvelle unité phraséologique baptisée « motif » (6.2.2).

---

<sup>135</sup> Cette organisation en phases du combat chevaleresque se trouve dans la littérature médiévale au sein du *topos* du « combat singulier » (sur ce point, voir Rychner 1955).

### 6.2.1 De l'arbre lexicosyntaxique récurrent au motif

Dans leur étude consacrée aux routines sémantico-rhétoriques<sup>136</sup> dans l'écrit scientifique, Tutin & Kraif (2016) se sont intéressés à ces séquences qui permettent aux scripteurs de s'inscrire dans « une communauté de discours. » (p. 119). Le repérage de ces unités phraséologiques s'appuie sur une méthode fondée sur l'exploitation d'annotations syntaxiques en dépendance : « par comparaison à des approches purement séquentielles [...] comme la méthode des segments répétés [...] et donc] en s'affranchissant de la linéarité et en exploitant les relations syntaxiques, les méthodes basées sur les corpus arborés permettent l'extraction de motifs plus abstraits » (p. 120). Dans ce cas, ces unités sont nommées « arbre lexicosyntaxique récurrent » (ALR) et sont extraites par le Lexicoscope sur les corpus qui y ont été préalablement intégrés. La figure 12 ci-dessous permet de visualiser un ALR sous la forme d'un schéma syntaxique de dépendance pour le GN *résultats étonnants*. Dans ce cas, le fait de s'affranchir de la linéarité permet d'extraire des suites comme *étonnants résultats*, *résultats étonnants* et *résultats vraiment étonnants* car leur structure profonde N-Adj est la même.

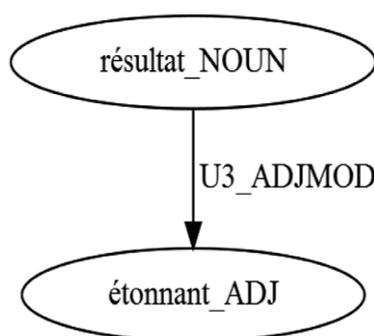


Figure 12 : Schéma syntaxique de dépendance pour le GN *résultats étonnants* (d'après Tutin & Kraif 2016, p. 125).

---

<sup>136</sup> Selon Tutin & Kraif (2016), les routines sémantico-rhétoriques se définissent comme des unités phraséologiques qui associent un patron à une fonction rhétorique spécifique de l'écrit scientifique : « par exemple, le patron [{nous/on/cet article} {reprenons/utilisons} DET {concept/modèle/conception} de X] qui indique le positionnement scientifique par rapport aux pairs. » (p. 120). Jacques & Tutin (2018, p. 83) proposent une définition complète des routines qui sont « a) des motifs récurrents, b) ayant une fonction discursive et rhétorique propre à un genre textuel, c) mettant en jeu une configuration lexico-syntaxique spécifique avec des paradigmes lexicaux, et d) correspondant à un énoncé actualisé dans le texte, c'est-à-dire renvoyant à des référents spécifiques comme l'objet d'étude, l'énonciateur du texte ou son destinataire. ». Elles précisent aussi la distinction entre *routine* et *collocation* : « les collocations constituent un matériel lexical abstrait. Les routines, par opposition, mettent en jeu des référents identifiés dans le discours. Par exemple, dans la routine *comme on l'a vu*, le *on* renvoie à l'auteur et aux destinataires. » (p. 83).

Dans une recherche consacrée aux phraséologismes construits autour du vin et du champagne (Gonon & Sorba 2022), nous avons interrogé le corpus romanesque PhraseoRom (romans de littérature blanche GEN, de science-fiction SF, policier POL, fantasy FY, historique HIST et sentimental SENT). À partir des ALR extraits par le Lexicoscope dans une démarche *corpus-driven*<sup>137</sup>, nous avons sélectionné ceux dans lesquels apparaissaient les mots clés *vin* et *champagne*. Le tableau 12 présente les résultats de cette sélection :

<b>Corpus contrastés</b>	<b>ALR extraits (total 5489)</b>	<b>ALR <i>vin/champagne</i> (LLR)</b>
GEN vs. HIST	1038	<boire du vin> (LLR 14,34) <boire du champagne> (LLR 24,18)
GEN vs. POL	406	<boire du vin> (LLR 41,85)
GEN vs. SF	840	<boire du vin> (LLR 52,87), <il boire du vin> (LLR 13,23), <boire du champagne> (LLR 10 000)
GEN vs. SENT	119	<boire du vin> (LLR 29,87)
GEN vs. FY	1148	<boire du vin> (LLR 12,46) <boire du champagne> (LLR 34,87)
HIST vs. GEN	78	0
POL vs. GEN	725	0
SF vs. GEN	247	0
SENT vs. GEN	315	0
FY vs. GEN	573	0

Tableau 12 : Présentation des ALR extraits par le Lexicoscope contenant les lexies *vin* et *champagne* (d'après Gonon & Sorba 2022, p. 404).

À la lecture du tableau 12, il apparaît qu'aucun ALR contenant les lexies *vin* ou *champagne* n'a été extrait quand les sous-genres romanesques HIST, POL, SF, SENT, FY ont été contrastés avec GEN. C'est seulement lors du contraste GEN vs les autres sous-genres romanesques que des ALR spécifiques ont été extraits (<boire du vin>, <il boire du vin>, <boire du champagne>). Ces unités phraséologiques permettent ici de caractériser un sous-genre romanesque par rapport aux cinq autres : la consommation d'alcool est relevée comme spécifique des romans de

---

<sup>137</sup> Le Lexicoscope permet de comparer automatiquement les sous-corpus deux à deux pour extraire les ALR spécifiques à un sous-genre en regard d'un autre sous-genre, c'est-à-dire les ALR dont la fréquence relative dans un sous-corpus est significativement supérieure à leur fréquence dans un autre sous-corpus. Nous nous sommes intéressées ici au contraste entre la littérature générale (GEN) et les autres sous-genres romanesques. Pour le détail des filtres appliqués, voir Gonon & Sorba (2022, p. 403-404).

littérature générale quand on les compare aux romans de paralittérature<sup>138</sup>. L'étude des variations paradigmatiques (collocatifs pronominaux, circonstants) et syntagmatiques (temps verbaux) révèle que *boire du vin* (ex. 43) est une activité plus quotidienne et moins marquée socialement et festivement que *boire du champagne* (ex. 44) :

43) Plus tard, avec les autres, ils taillèrent dans le pain, le jambon, le saucisson. Ils **burent du vin**. Olivier regarda sa paume où deux ampoules s'étaient formées. (P. Sabatier, *Les Trompettes guerrières*, 2007)

44) On domine le bois de Boulogne. Il fait chaud. On **boit du champagne**. On écoute des sonates de Scarlatti, la Kirkpatrick 209... (Ph. Sollers, *Femmes*, 1983)

Une explication à cette spécificité de GEN par rapport aux autres sous-genres romanesques dits « populaires » pourrait en effet être que « la mimesis du quotidien contemporain y est de fait plus importante, l'action stéréotypée (une enquête policière, une quête fantastique, un complot historique...) des romans populaires empêchant généralement le développement récurrent de scènes quotidiennes. » (*ibid.*, p. 420). Dans les romans de littérature générale, la représentation de la consommation d'alcool contribuerait à construire l'atmosphère du roman, que Bentolila nomme l'« effet de réel » (2016, p. 56), au même titre que la représentation de la consommation de tabac : « By virtue of being part of social rituals, smoking and drinking typically serve to create a mood or an atmosphere in a literary text. » (Grossmann et al. 2020, p. 120).

Comme le soulignent Jacques & Tutin (2018), plusieurs études, sur des corpus variés, se sont intéressées à ces « objets linguistiques plus émergents, au croisement du lexique et du discours, dont la modélisation reste à affiner » (p. 75). Par exemple, dans leur réflexion sur des corpus littéraires, Maingeneau & Philippe (2002) utilisent le terme de *patrons discursifs* pour désigner « les représentations imaginaires de plusieurs types de production langagière précis dont la tradition littéraire a figé les spécificités en une sorte de stéréotype. Le déchiffrement des textes romanesques repose en effet en tout premier lieu sur la reconnaissance de tels « patrons » et exige une compétence interprétative spécifique. » (p. 366-367). Ces patrons récurrents guideraient l'affiliation d'un texte à un genre, en l'occurrence le genre romanesque, lors de la lecture.

---

<sup>138</sup> Selon Boyer (2008), la paralittérature se définit comme « l'ensemble des livres de fiction dont la diffusion est massive, et que le discours critique, le plus fréquemment, ne considère pas, ou pas encore, comme appartenant à la littérature » (p. 7). Cette étiquette correspond aux cinq sous-genres romanesques du corpus PhraseoRom (HIST, POL, SF, SENT, FY). Pour Baroni (2003), la paralittérature se caractérise surtout par sa conformité aux stéréotypes, en privilégiant les « recettes qui marchent » (p. 145).

Dans cette même perspective de caractériser un texte par des unités phraséologiques plus abstraites, Longrée, Luong & Mellet (2008) utilisent le terme de *motif*<sup>139</sup> :

Le motif fournit un cadre formel accueillant toutes sortes d'unités linguistiques et textuelles ; par exemple, dans le domaine lexical, il pourrait englober ce qu'A. Salem a reconnu sous le nom de « segments répétés » ou ce que, dans un autre domaine d'application, on appelle les « unités phraséologiques », ou encore – dans la mesure où le motif admet la présence d'une variable en son sein – ce qu'en milieu anglo-saxon on appelle « collocation ». Mais le motif pourrait aussi rassembler des structures phonologiques ou des structures métriques. (Longrée, Luong & Mellet 2008, p. 736).

La modélisation du cadre est ici très peu contraignante car celui-ci peut accueillir des unités linguistiques de différents niveaux. Dans cette étude, le corpus privilégié est celui des historiens latins et tout particulièrement les séquences narratives. Les deux séries de motifs analysés, les « motifs cadratifs » (le cadre circonstanciel de l'événement relaté est posé au début de la phrase narrative à grand renfort de propositions subordonnées) et les « motifs de rallonge » (le cadre circonstanciel de l'événement relaté est posé à la fin de la phrase narrative), permettent de distinguer nettement, par exemple, les commentaires de César des récits historiques de Tacite. Le motif est donc ici lié à une fonction dans l'économie du texte qu'il caractérise. C'est d'ailleurs, cette idée que les auteurs formulent ultérieurement quand ils proposent d'affiner leur définition du *motif* comme « unité phraséologie englobante » :

Sur le plan fonctionnel, le motif est un « cadre collocationnel » accueillant un ensemble d'éléments fixes et de variables, susceptible d'accompagner la structuration textuelle et, simultanément, de caractériser des textes de genres divers, voire de permettre la détection, au sein d'un même texte, des passages de registres différents. (Longrée & Mellet 2013, p. 66).

Pour illustrer leur définition, les auteurs donnent en exemple un « motif transitionnel parenthétique caractéristique du discours argumentatif » (*ibid.*, p. 67-68) :

ici j'aimerais toutefois préciser / ici je voudrais toutefois préciser / ici je souhaiterais toutefois préciser / ici, toutefois, j'aimerais préciser / j'aimerais toutefois préciser / ici j'aimerais préciser / ici je souhaite toutefois préciser, etc.

Tous ces énoncés présentent une variation qu'elle soit paradigmatique (*aimer/vouloir/souhaiter*) ou syntagmatique (présence ou non de l'adverbe ; déplacement de

---

<sup>139</sup> Le terme est déjà présent dans Habert et *al.* (1997) : « Les constituants de ces motifs sont, ici encore, des structures de traits. Le *motif* (ou *patron*) correspondra au fragment de texte pour lequel les structures de traits de ses composants s'apparient avec celles des éléments du texte. » (p. 39-40). La notion de *motif* est ici théorisée sans être appliquée à un genre textuel particulier. Pour un historique sur cette notion, voir Legallois & Koch (2020).

l'adverbe ; temps du verbe) mais il s'agit bien des réalisations discursives d'un seul et même motif.

Nous avons testé, au sein du sous-corpus POL, la potentialité de l'unité phraséologique qu'est le motif pour distinguer les trois synonymes *crime*, *meurtre* et *assassinat* (Novakova & Sorba 2017). L'expérimentation s'avère concluante dans la mesure où deux motifs se sont révélés spécifiques à la lexie *meurtre* : (i) le motif de la motivation de l'acte (*avoir/sentir/éprouver/donner/nourrir/être en proie à/réprimer (souvent/aussi) une/des envie(s) de meurtre*) et (ii) le motif de la sérialité (*initier/être impliqué dans/devenir le moteur de/avoir quelque chose à voir avec/éclairer/travailler sur la/les série(s) de meurtre /les meurtres en série*). Dans le premier cas, le motif remplit une fonction descriptive en se focalisant sur la description du criminel ; dans le second cas, le motif occupe une fonction narrative structurant la progression de l'enquête après le meurtre qui l'a déclenchée.

C'est cette notion de motif que l'équipe du projet PhraseoRom a approfondie et mise à l'épreuve dans les différentes études portant sur la caractérisation phraséologique du genre romanesque. En s'appuyant sur les différentes études réalisées, Legallois & Koch (2020) ont proposé de définir le motif comme une unité linguistique à part entière :

[...] the motif is a construction – a linguistic sign – considered not only from the grammatical point of view but from the perspective of its discursive dimension in a particular text or a particular genre. (p. 39-40).

Ainsi, les motifs correspondent à des patrons lexico-syntaxiques récurrents, qui présentent des régularités et des variations lexicales et syntaxiques aux niveaux paradigmatique et syntagmatique, associées à des fonctions discursives spécifiques<sup>140</sup>. Le motif est ainsi le lien qui permet de relier les niveaux macro et micro : « The notion of motif lets us fill in the missing link between 'macro-level' notions of script or schema that have traditionally been used in cognitive narratology (e.g. study of the plot, isotopies) and the 'micro-level' elements that go into making up the script (specific phraseological recurrences) » (Novakova & Siepmann 2020, p. 10). C'est pourquoi notre analyse de ces unités phraséologiques s'intéresse non seulement à leur fonctionnement au sein des textes mais aussi au rôle qu'elles y jouent comme nous allons le voir dans les études de cas concernant différents sous-genres romanesques.

---

<sup>140</sup> « [...] the discursive function (DF) of a motif depends on the role it plays in the context, such as in maintaining "textual coherence" and "thematic continuity" » (Novakova & Siepmann 2020, p. 291). Le projet PhraseoRom a proposé une typologie fine des fonctions discursives en huit catégories : narrative, descriptive, affective, indirectement descriptive, infra-narrative, infra descriptive, cognitive, pragmatique (p. 292-293).

## 6.2.2 Des motifs pour contraster des sous-genres romanesques : études de cas

Les études de cas présentées ici sont issues de collaborations fructueuses avec les stylisticiens du projet PhraseoRom, collaborations qui illustrent bien le mot de Sinclair (2008) selon lequel « phraseology is by no means the exclusive province of linguists » (p. xvi). Leur objectif est de montrer comment les ALR extraits du Lexicoscope permettent de repérer des unités phraséologiques, des motifs, permettant de contraster les sous-genres romanesques les uns avec les autres.

Dans l'une des toutes premières études de notre équipe (Gonon et al. 2016), le contraste des sous-genres romanesques a fait apparaître l'ALR <sur la scène de crime> comme spécifique du sous-corpus POL (figure 13) :

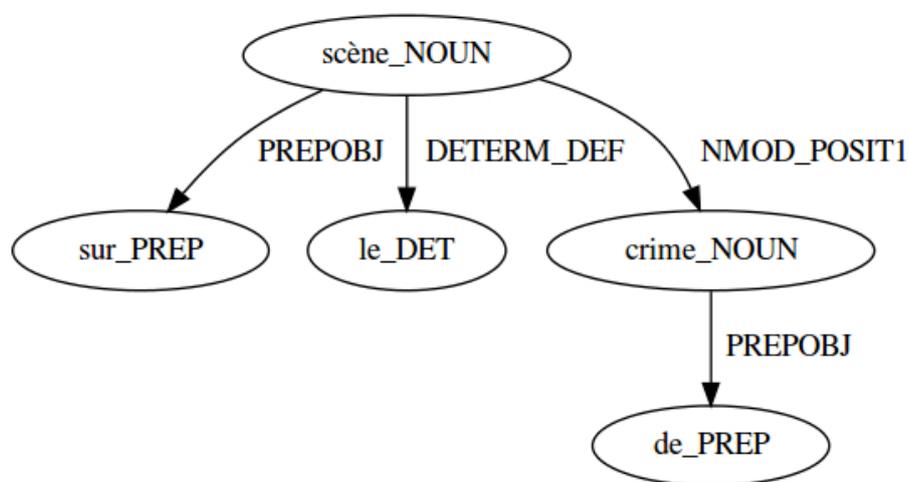


Figure 13 : ALR <sur la scène de crime> (d'après Gonon et al. 2016, p. 3).

Ce résultat – certes prévisible – nous a permis de lancer une étude du fonctionnement et du rôle structurant de cette unité phraséologique dans le roman policier. Sur le plan diachronique, nous avons pu constater que la collocation *scène de crime* apparaissait à partir de la fin des années 1990, ce qui semble coïncider avec son usage récent en criminologie, tout particulièrement dans le sous-sous genre du roman policier qu'est le *thriller* (*ibid.*, p. 4).

L'étude des occurrences de *scène de crime* nous a conduits à regrouper les structures en deux sous-ensembles : le motif de la localisation et le motif de l'expertise. Nous nous focaliserons ici sur le premier dont voici deux exemples (45-46) :

- 45) C'était la première chose à faire **en arrivant sur une scène de crime**. (M. Chattam, *Maléfices*, 2004)
- 46) Quand Volo parvint à leur hauteur, il disait :
  - On pourrait penser à des échantillons du parquet mais ce n'est pas le cas. Pour moi, c'est la même essence que la première fois.

- **Et sur la scène de crime de l'Indien**, boulevard Malesherbes, tu en as trouvé aussi ?
- Dans le couloir, ouais. (J.-C. Grangé, *Miserere*, 2008)

Ces deux exemples (45-46) illustrent le phénomène de variation qui caractérise un motif (ici, variations syntagmatiques avec expansion à gauche ou à droite). L'analyse des données nous a montré que sur le plan discursif, l'unité phraséologique *scène de crime* a non seulement pour fonction d'introduire l'élément initial de la diégèse (ex.45) mais aussi d'en faire le support d'observation et de remémoration (ex.46). On passe ainsi d'une dimension narrative à un discours plus descriptif :

La dimension narrative s'estompe donc, dans la mesure où l'expression *scène de crime* entre petit à petit dans l'énumération des différentes pièces de l'enquête, suivant la logique d'aspectualisation propre au discours descriptif (Hamon, 1993) ; la narration se dilue prototypiquement en discussions entre les protagonistes (légiste, policier, juge) qui échangent des constatations et échafaudent des hypothèses, ainsi qu'en observations et analyses des indices. (Gonon et al. 2016, p. 9).

Dans une autre étude consacrée au roman historique, nous avons pu comparer les ALR de l'interaction verbale entre les deux sous-corpus HIST et GEN (Gonon & Sorba 2020). Les quatre constructions qui ont émergé comme les plus représentatives sont, pour le sous-corpus HIST, <donner l'ordre>, <dire d'une voix>, <je vous prie>, <je vous remercie> et, pour le sous-corpus GEN, <je ne sais pas>, <j'ai oublié>, <j'ai l'impression>, <je me dis>. Une analyse de ces premières données permet de constater que les ALR de l'interaction verbale expriment plutôt une analyse de la situation par un locuteur, une réflexion intérieure, un mouvement d'introspection face à un allocutaire ou à soi-même dans GEN tandis qu'ils révèlent plutôt des rapports de pouvoir attendus dans HIST.

L'étude des occurrences dans lesquelles se réalise l'ALR <donner l'ordre> montre qu'il s'agit bien d'un motif qui occupe une fonction narrative, au service de la progression de l'intrigue, dans des contextes très souvent militaire (un seigneur qui commande à ses soldats, ex. 47) ou relevant aussi de la vie quotidienne (un mari commande à sa femme ou un seigneur à son valet, ex. 48) :

- 47) Quand les maisons entourant le palais de Mutarrif ne furent plus qu'un tas de cendres fumantes, Ubaid Allah Ibn Mohammad Ibn Abi Ibn Abda **donna l'ordre** à ses troupes d'attaquer. (P. Girard, *Abdallah le cruel*, 2007)
- 48) Le valet barbier connaissait vaguement le recteur du collège Saint-Cosme. Il demandait une somme élevée pour arracher une dent, mais assura-t-il, souvent le malade survivait. Olivier lui **donna l'ordre** d'aller le chercher. (J. d'Aillon, *La Guerre des trois Henri t.3*, 2009)

Ce motif *donner l'ordre* renvoie à une société hiérarchisée par des rapports de domination et rend compte de l'ethos de l'homme qui détient le pouvoir en la circonstance. Dans l'exemple 47, le général aux ordres du sultan mène l'assaut contre le palais de Mutarrif et transmet l'ordre à ses soldats ; dans l'exemple 48, le héros, Olivier Hauteville, cherchant à soigner la rage de dents de son hôte, s'adresse au valet de chambre-barbier d'un voisin. Dans ces deux extraits, nous avons constaté que la fonction discursive était également indirectement descriptive puisqu'elle évoque les rapports sociaux entre les personnages de même que, dans le roman policier, le motif *ouvrir la porte* décrit indirectement le détective comme un individu toujours en mouvement (Gonon et al. 2018). C'est bien une caractéristique de la paralittérature d'utiliser, selon le sous-genre romanesque, des personnages qui sont davantage des types que des individualités (Krucic 2007, p. 46). Grâce aux fonctions discursives qu'ils occupent, les motifs permettent de distinguer finement les sous-genres romanesques.

La même étude (Gonon & Sorba 2020) examine le motif construit autour de l'ALR <je vous prie>. Les occurrences de celui-ci oscillent entre la prière empressée (ex. 49) et l'ordre souverain (ex. 50) :

49) Ta fille te sera rendue, dit Clémence, et je la protégerai. J'en veux parler au roi.

– N'en faites rien, Madame, **je vous en prie**, s'écria Eudeline.

– Le roi me comble de cadeaux que je ne souhaite pas ; il peut bien m'en accorder un qui me plaise !

– Non, non, je vous en supplie, n'en faites rien, répéta Eudeline. J'aime mieux voir ma fille sous le voile que de la voir sous terre. (M. Druon, *Les Rois maudits* t.3, 1956)

50) C'est impossible, Monsieur ! Que me reproche-t-on ?

– Vous le saurez en temps voulu. Laissez-nous travailler, **je vous prie**.

L'un des hommes se dirigea vers un secrétaire, l'autre vers une commode. Sophie ne songea même pas à protester. Elle savait, par expérience, qu'il est superflu de parler raison à un policier chargé d'exécuter un ordre. (H. Troyat, *La Lumière des Justes* t.5, 1963)

Ce motif permet de révéler une fonction discursive (FD) que nous avons proposé de nommer pragmatique, le terme *pragmatique* renvoyant ici à la réalisation d'un acte de langage par un personnage du roman historique à l'intérieur du discours romanesque lui-même. En effet, dans l'exemple 49, « le motif *je vous prie* exprime la supplique de la servante, et est d'ailleurs repris par *je vous en supplie*. La FD est pragmatique en ce qu'elle renforce la dimension illocutoire de l'acte de langage et oriente l'interprétation de la phrase dans le contexte dialogal : le motif souligne l'emploi de l'impératif à la forme négative (*N'en faites rien*) et lui donne valeur de prière et non pas d'ordre. » (Gonon & Sorba 2020, p. 12).

De plus, dans le sous-corpus GEN, nous avons pu observer une spécificité du roman de littérature blanche dans le cadre de ces ALR de l'interaction verbale :

- 51) **Je n'ai pas tout à fait oublié** le mal que ces gens m'ont fait. J'y pense mais je suis indifférent. (T. Ben Jelloun, *Amours sorcières*, 2003)
- 52) Je ressentais une joie que je ne pouvais ressentir – dont **j'avais l'impression** de ne plus posséder les instruments pour la ressentir. (P. Quignard, *Le Salon du Wurtemberg*, 1986)
- 53) Il est hyper-doué et il est nul en classe : ça l'emmerde. Mais aux tests de QI, c'est le premier. Résultat : il se fait cogner par tout le monde et il régresse. **Je sais pas** ce qu'il va devenir. J'ai pas les moyens de le changer d'école (D. Van Cauwelaert, *Hors de moi*, 2003)

Dans l'exemple 51, <j'ai oublié> (ici avec variation syntagmatique de la négation) évoque des expériences pénibles laissées derrière soi, un passé qui a marqué ; dans l'exemple 52, le narrateur s'émeut d'une invitation et en rapporte les effets dans son récit avec <j'ai l'impression> ; dans l'exemple 53, la locutrice s'adresse au narrateur et lui avoue son ignorance <je sais pas> en le faisant ainsi entrer dans sa réflexion. Dans ces trois motifs, nous avons proposé de nommer *cognitive* la fonction discursive que ceux-ci remplissent car ils renvoient à des processus cognitifs (hypothèses, réflexions, souvenirs, etc.) du personnage. Cette étude nous a permis, d'une part, de préciser la description des fonctions discursives pour aller au-delà des traditionnelles fonctions narrative et descriptive du roman (Adam 2011, p. 267) et, d'autre part, de montrer le potentiel de ces unités phraséologiques pour affiner la distinction entre les sous-genres romanesques.

Nous avons pu mettre en pratique cette distinction fine entre les fonctions discursives au sein d'une étude contrastive entre les deux sous-corpus GEN de romans français (FR) et anglais (EN) (Sorba et al. 2020). Elle a permis de révéler des ressemblances et des différences concernant le rôle structurant des motifs au sein du sous-genre GEN dans les deux langues. Dans cette étude, nous avons analysé les occurrences des quatre ALR les plus spécifiques comprenant les verbes *écrire/write* et *lire/read*<sup>141</sup> :

<lire\_le\_lettre>, <je\_lire\_livre> ; <read\_the\_letter>, <read\_the\_book>

<écrire\_un\_roman>, <je\_écrire\_lettre> ; <write\_the\_word>, <write\_the\_letter>

En comparant les fonctions discursives de ces motifs, notre objectif était de voir si ceux-ci jouaient le même rôle dans les roman français et anglais de littérature générale ('literary

---

<sup>141</sup> Le choix de ces ALR a été guidé par la découverte, lors d'une précédente étude (Gonon et al. 2018), de deux associations hautement spécifiques dans GEN *écrire + lettre* (LLR 2038) et *écrire + roman* (LLR 763).

fiction’). Je présente ici deux résultats sur les comparaisons entre *lire\_DET\_lettre* et *read\_DET\_letter* et entre *Lire\_DET\_livre* et *read\_DET\_book*.

Dans le cas du motif *lire\_DET\_lettre* / *read\_DET\_letter*, nous avons pu observer que la fonction discursive la plus fréquente dans les deux langues était bien la fonction narrative :

54) Quand Véro est accourue, ce sale matin de janvier 17, venant d’apprendre par quelqu’un du quartier que son amant était mort, Petit Louis lui a fait **lire la dernière lettre de l’Eskimo** [où il exprime ses dernières volontés] et demandé de s’expliquer. Elle était en larmes, à genoux sur le sol, effondrée. (S. Japrisot, *Un long dimanche de fiançailles*, 1991)

55) She turned over the other letter. It was addressed to herself. As she instantly recognised the writing, she stood for several moments very still. Then breathing deeply she moved into her bedroom and closed the door. She sat down on the bed, opened the envelope, and **read the letter through carefully**. (I. Murdoch, *The Green Knight*, 1993)

Dans l’exemple 54, le motif *lire\_DET\_lettre* qui relie les différents personnages (Véro, Petit Louis, l’Eskimo) entre eux construit la séquence narrative en faisant avancer l’intrigue : la lettre révèle un fait qui a une conséquence directe sur son destinataire qui en prend connaissance. C’est exactement le même phénomène dans l’exemple 55 où le motif apparaît au sein d’une liste d’actions qui conduisent naturellement à prendre connaissance du contenu de la lettre, ici redouté par le personnage.

En outre, ce même motif peut occuper une fonction discursive pragmatique dans les deux corpus comme l’illustrent les deux exemples suivants (56-57) :

56) Je pense à Diogène qui offrit des figes à Platon, et à La Rochefoucauld qui envoyait des Maximes inédites à une amie sans prétendre pour autant mériter son potage de carottes. Quand il ne réussissait pas à trouver les truffes qu’elle lui demandait, il lui envoyait à la place quelques Maximes qui, s’excusait-il, ne valaient pas de bonnes truffes. Si nous nous revoyons, je vous **lirai les lettres** qu’il écrivit à cette femme, la marquise de Sablé. (F. Weyergans, *Trois jours chez ma mère*, 2005)

57) Abruptly Fritz leaned towards her and said, “You know that I have only one thing to ask. **Has he read my letter?**” (P. Fitzgerald, *The Blue Flower*, 1988)

Dans l’exemple 56, le motif *lire\_DET\_lettre* créé un lien entre les deux personnages, François qui tente de séduire Juliette, mais se présente sous la forme hypothétique d’une promesse, un acte de langage dont la réalisation n’est pas garantie. Dans l’exemple 57, le motif *read\_DET\_letter* apparaît dans un dialogue sous la forme d’une question directement adressée à un personnage (la mère du locuteur), un acte de langage se présentant sous la forme d’une requête dont la réponse est susceptible d’influer sur le cours de l’intrigue.

En revanche, il est apparu que la fonction discursive descriptive était plutôt spécifique aux romans français pour ce motif :

- 58) Un doigt posé sur la bonne ligne, j’attendais pour continuer ma lecture. En tournant sur elle-même, Edmondsson **lisait des lettres**, classait des documents. Elle s’éloignait du bureau, revenait vers moi. Elle s’asseyait sur le fauteuil et, en bougeant les lèvres, prenait connaissance d’un imprimé ; puis elle décroisait les jambes, se relevait et faisait des commentaires. (J.-P. Toussaint, *La Salle de bain*, 1985)

Dans l’exemple 58, le personnage lit des lettres en effectuant parallèlement des actions insignifiantes du point de vue de l’intrigue sous l’œil du narrateur. Par contraste, dans les romans anglais, le motif *read\_DET\_letter* est employé plutôt de manière dynamique (fonctions narrative ou pragmatique) que comme une justification pour décrire un personnage comme c’est le cas ici en français (Sorba et al. 2020, p. 238).

Le second motif *lire\_DET\_livre* et *read\_DET\_book* se distingue du précédent sur le plan statistique (ses occurrences sont plus nombreuses) et sur le plan stylistique (il permet une mise en abyme). En effet, contrairement à la lettre, le livre n’est pas adressé à un personnage et n’établit pas de lien entre les personnages comme peut le faire la lettre. La lecture d’un livre, activité solitaire par excellence, sans impact sur la progression narrative, encourage une posture plutôt introspective de la part du personnage. C’est ainsi que la fonction discursive cognitive est bien occupée par ce motif aussi bien dans les romans français qu’anglais :

- 59) Il n’aime pas la musique. Il trouve que ça gêne. Que c’est une servitude. Parce que, quand on en écoute, on est obligé d’entendre le morceau sans l’interrompre, alors que quand on **lit un livre**, on peut le lire dans le désordre, sauter des pages, que la liberté est totale. Il aime cette liberté et ne supporte pas d’en être privé. (C. Angot, *Une semaine de vacances*, 2012)

- 60) Will I ever be able to **read a book again**? Will I ever be able to talk like old friends with my wife? (T. Parks, *Europa*, 1997)

L’exemple 59 débute par une description des goûts du personnage masculin et le motif *lire\_DET\_livre* est introduit par l’impersonnel *on* qui introduit du discours rapporté nous livrant ainsi les pensées de ce même personnage. Dans l’exemple 60, le motif *read\_DET\_book* apparaît dans le discours à la première personne du narrateur formulé sous la forme d’une hypothèse. Ces deux exemples illustrent bien la fonction discursive cognitive de ce motif qui permet au lecteur d’accéder au discours intérieur des personnages.

Grâce à cette étude, nous avons pu montrer que la spécificité de ces motifs dans les romans français et anglais de littérature générale révélait la place conséquente que les auteurs

accordaient aux activités d'écriture et de lecture au sein de leur ouvrage comme un reflet de leur propre pratique auctoriale (Sorba et *al.* 2020, p. 248).

Les quelques études de cas présentées dans cette section 6.2.2 ont chacune permis d'illustrer l'apport de l'unité phraséologique qu'est le motif à la caractérisation des sous-genres romanesques. Le travail réalisé dans le cadre du projet PhraseoRom sur un genre littéraire singulier, le roman, présente un apport original dans le dialogue qu'il a permis d'instaurer entre différentes approches du texte comme le souligne Novakova & Gymnich (2021) :

The researchers collaborating in the context of the PhraseoRom project have explored new ways of bridging the gap between quantitative, corpus-linguistic approaches and traditional goals of literary studies. They propose detailed analyses of lexico-syntactic constructions that are specific to subgenres of French and English novels. The usefulness of the concept of the phraseological motif for distinguishing between literary subgenres has been proven by various case studies completed within the PhraseoRom project. (p. 90)

C'est la contribution phraséologique que nous apportons à la « théorie opératoire des genres » textuels chère à Rastier. De nombreuses recherches restent encore à faire dans ce champ en s'appuyant, par exemple, sur les ALR extraits de la comparaison des sous-corpus romanesques les uns avec les autres. La liste de ceux-ci est d'ailleurs librement disponible sur la PhraseoBase sous la forme de ressources téléchargeables.

L'objet de ce sixième et dernier chapitre était de retracer les apports de la linguistique de corpus outillés qui, depuis de nombreuses années, favorise l'émergence de nouveaux observables phraséologiques dont les linguistes se sont emparés. Comme le rappelle Sinclair (2008), « the main reason for the rapid increase in interest in phraseology is the present-day use of text corpora as the principal data-source for language analysis. 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. » (p. xvi).

Depuis les premiers concordanciers électroniques, qui ont considérablement accru les capacités d'observations des linguistes, aux outils de fouille textuelle extrêmement sophistiqués comme le Lexicoscope, la recherche en phraséologie s'est développée largement. Notre équipe a ainsi pu exploiter les séquences extraites par cet outil sous la forme des ALR pour modéliser une nouvelle unité phraséologique, le motif, permettant de distinguer les sous-genres romanesques les uns des autres. Ma perspective est désormais de permettre à ces études phraséologiques qui ont dépassé le cadre du syntagme de se déployer dans la diachronie longue en explorant plusieurs autres genres textuels.

## Conclusion

Au terme de ces réflexions sur les travaux que j'ai réalisés depuis une douzaine d'années, je propose de récapituler les principales lignes de force de ma recherche. Ancrée dans une perspective profondément interdisciplinaire, ma démarche scientifique fait dialoguer différentes traditions, diverses époques et des productions textuelles variées. La structuration de ce mémoire en trois parties, consacrées respectivement à la phraséologie, à la question des genres textuels et à l'apport de la linguistique outillée, m'a permis de montrer la dynamique de ma recherche actuelle et les questionnements qui la sous-tendent pour l'inscrire dans le temps long. C'est une approche holistique et fonctionnelle que je mets en oeuvre pour décrire les phénomènes linguistiques complexes qui font les textes et inscrivent ceux-ci dans des genres.

Dans la première partie, j'ai montré que la phraséologie, une discipline linguistique à part entière, prend comme objet d'étude des unités complexes et peu homogènes. La constitution de cette discipline doit beaucoup à l'utilisation massive récente des corpus comme le rappelle Sinclair (2008) :

[...] the main reason for the rapid increase in interest in phraseology is the present-day use of text corpora as the principal data-source for language analysis. 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 preprocessing is required, no abstractions, no information such as parts of speech added. This is made possible because corpora, especially large corpora, open up new lines of argument, based heavily on the perception of stretches of text as repeated events, and being able to compute frequencies and use statistical measures. (p. xvi)

Néanmoins, j'ai montré que les interrogations sur les séquences préconstruites dans les langues étaient anciennes. La nouveauté a consisté en un changement de perspective : la phraséologie est passé du statut d'élément marginal au devant de la scène comme l'ont rappelé Granger & Meunier (2008) : « Long regarded as a peripheral issue, phraseology is now taking centre stage in a wide range of fields, from natural language processing to foreign language teaching and now [...] it is gradually acquiring the place it deserves in linguistic theory. » (p. xix). Ma recherche a ainsi contribué (et contribuera encore) à modéliser le fonctionnement linguistique des unités phraséologiques dans les textes à différents niveaux en synchronie et en diachronie.

Dans la deuxième partie, j'ai proposé une réflexion sur le renouvellement de l'approche générique. Face à cet objet de haute complexité qu'est le texte, j'ai pu montrer toute la difficulté de l'entreprise typologique. Néanmoins, afin de trouver une issue à ce problème apparemment insoluble, je propose là aussi de changer de perspective en considérant la caractérisation du genre textuel de manière dynamique. C'est ainsi un point commun que le genre partage avec la

phraséologie : « Le système phraséologique d'une langue constitué historiquement est toujours en renouvellement permanent. » (Buridant 1989, p. 128). Afin de rendre compte de la généricité aussi bien sur le plan synchronique de la variation que sur le plan diachronique du changement, la typologie des genres textuels sera forcément ouverte et sujette à évolution.

Enfin, tout l'enjeu de la troisième partie a été de montrer comment saisir au mieux cette réalité doublement mouvante (la phraséologie et le genre textuel). La caractérisation d'un genre textuel par la saisie des unités phraséologiques qui y sont repérées au sein de corpus outillés paraît une solution prometteuse. La collocation est effectivement l'une de ces unités qui permettent de modéliser linguistiquement la spécifique générique. L'apport des arbres lexico-syntaxiques récurrents (ALR) dans ce domaine permet de s'appuyer sur les liens syntaxiques entre les unités lexicales (palier local ou niveau micro) pour accéder au palier global du sens dans le texte (niveau macro). En outre, les résultats de nos recherches sur les motifs, au cœur desquels se trouve la collocation extraite au moyen des ALR, ont montré tout le potentiel heuristique de cette nouvelle unité, notamment par les fonctions discursives qu'elle occupe, pour distinguer les sous-genres au sein d'un genre littéraire singulier, le roman.

Le travail en cours de constitution de corpus de genres variés, en synchronie et en diachronie, permettra, je l'espère, d'apporter des éléments nouveaux pour contribuer à cette « théorie opératoire des genres » que j'ai posée en fil conducteur de ma recherche.

## **Annexes**

## I. Liste des travaux scientifiques

La liste de mes travaux est présentée ci-dessous par ordre chronologique décroissant. Pour chaque année, sont indiqués tous les travaux scientifiques (publication, communication, brevet, conférence invitée). L'ensemble des publications est disponible sur HAL dans le respect de la législation des droits d'auteur <<https://cv.archives-ouvertes.fr/julie-sorba>>. Un graphique, à la fin de cette annexe, permet de visualiser l'ensemble de mes travaux par catégorie et année.

### 2022

BOCH F., RINCK F. & **SORBA J.** (soumis, 2022). Former les enseignants du supérieur à la révision des textes des étudiants, *Pratiques*.

COAVOUX M., DENOYELLE C., KRAIF O. & **SORBA J.** (2022, accepté). Phraséologie du roman médiéval en prose. Communication présentée au Colloque *Diachro X – le français en diachronie*, Paris, 17-19 janvier 2022 (reporté au 31 mai – 2 juin 2022).

DENOYELLE C. & **SORBA J.** (2022, accepté). Analyse phraséologique du marqueur discursif *il m'est avis* et de ses variantes dans un corpus d'ancien français. Communication présentée au Colloque *La phraséologie dans les interactions orales et écrites*, Université Grenoble Alpes, 14-15 juin 2022.

FABRY I., LIGNEREUX C. & **SORBA J.** (2022). Un automatisme rhétorique : les structures comparatives dans les lettres de compliments. Communication présentée au Colloque *La corrélation en diachronie longue (1450-1800). Phrase, texte et discours*, Université Grenoble Alpes, 14-15 janvier 2022.

GONON L. & **SORBA J.** (2022). Phraséologismes autour du vin et du champagne dans le roman contemporain. In C. Feyrer, C. Konzett-Firth & E. Lavric (éd.), *Le vin et ses émules. Discours oenologiques et gastronomiques* (p. 399-423). Frank & Timme.

LI Q. & **SORBA J.** (soumis, 2022). L'organisation des expressions ritualisées dans les lettres en français et en chinois. Communication présentée au Colloque *Informer, déformer, re-former les genres : entre contraintes et variations*, Université de Poitiers, 20-21 juin 2022.

**SORBA J.** (2022 sous presse). Note de lecture. *Grande Grammaire Historique du Français*, Marchello-Nizia C., Combettes B., Prévost S. & Scheer T. (dir.), De Gruyter, 2020, 2185 p., *LIDIL*, 65.

**SORBA J.** & NOVAKOVA I. (2022 sous presse). Le dragon dans le roman contemporain : un motif pour distinguer les sous-genres romanesques ? *Iris*, 42.

## 2021

BOCH F., RINCK F. & SORBA J. (2021). Linguistics errors in students' academic writing: how to adapt a linguistic typology for pedagogical uses? Communication présentée au Colloque WRAB 2020 - *Writing Research Across Borders V*, initialement prévu à Xi'an International Studies University, Xi'an (Chine), 11-14 mars 2020, annulé pour raisons sanitaires et reporté 5-6 mars 2021 (à distance, <<https://vimeo.com/520139011>>).

DENOYELLE C., SORBA J. & KRAIF O. (2021). *Phraséologie et genres textuels : le cas du roman médiéval (Projet Formules)*. Communication présentée à la journée d'étude *Corpus et généricité*, Université Grenoble Alpes, 18 mai 2021 (à distance).

DIWERSY S., GONON L., GOOSSENS V., KRAIF O., NOVAKOVA I., SORBA J. & VIDOTTO I. (2021). La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase. *Corpus*, 22 <<https://doi.org/10.4000/corpus.6101>>

FABRY I., LIGNEREUX C. & SORBA J. (2021). Approche diachronique des marques d'appartenance dans les formules de clôture. Communication présentée au Colloque international *Épistolâtries : mutations contemporaines et nouvelles approches d'étude de la lettre*, Université Paris 8 et Université Paris Nanterre (France), 2-3 décembre 2021.

LIGNEREUX C., SORBA J. & FABRY I. (2021). *Phraséologie du genre épistolaire (Projet PhraséoCorr)*. Communication présentée à la journée d'étude *Corpus et généricité*, Université Grenoble Alpes, 18 mai 2021 (à distance).

ROUKHOMOVSKY B., SORBA J. & DUFOUR É. (2021). *ENCHRE / Phraséobase : un exemple de synergie*. Communication présentée à la journée d'étude *Corpus et généricité*, Université Grenoble Alpes, 18 mai 2021 (à distance).

SORBA J. (2021a). Note de lecture. *La phraséologie entre fixité et congruence. Hommage à Salah Mejri*, Inès Sfar & Pierre-André Buvet (dir.), Éditions Academia – L'Harmattan, Louvain-la-Neuve, 2018, 374 p. *LIDIL*, 63. <<https://doi.org/10.4000/lidil.9103>>.

SORBA J. (2021b). Note de lecture. *Grammaticalia. Hommage à Bernard Colombat*, Jean-Marie Fournier, Aimée Lahaussais & Valérie Raby (dir.), ENS Éditions, 2019, 318 p. *LIDIL*, 64. <<https://doi.org/10.4000/lidil.9403>>.

SORBA J. (2021c). Compte-rendu. Nicole Tersis & Pascal Boyeldieu (éds.). — Le langage de l'émotion. Paris, Peeters, 2017, 655 p. *Bulletin de la Société de linguistique de Paris*, 116 : 99-108.

SORBA J. (2021d). Conférence invitée. Profil combinatoire de *constituere* chez les historiens latins : une approche en linguistique de corpus outillés. *Journée d'étude Rem publicam*

constituer : le paradoxe d'une chose publique à (re)fonder ?, 23 novembre 2021, UGA-LUHCIE.

## 2020

BOCH F. & SORBA J. (2020). Tester la compétence lexicale des adultes francophones : réflexion sur le choix des distracteurs dans un test à choix multiple. *LIDIL*, 62. <<https://doi.org/10.4000/lidil.8023>>

BOCH F., RINCK F. & SORBA J. (2020). Les acquisitions tardives en français écrit : une base de données sur les erreurs et maladroites à un niveau avancé. *7<sup>e</sup> Congrès Mondial de Linguistique Française*, SHS Web Conf., vol. 78. <[10.1051/shsconf/20207806005](https://doi.org/10.1051/shsconf/20207806005)>

DENOYELLE C. & SORBA J. (2020). L'approche phraséologique du roman médiéval : une voie de caractérisation générique ? *7<sup>e</sup> Congrès Mondial de Linguistique Française*, SHS Web Conf., vol 78. <[10.1051/shsconf/20207805005](https://doi.org/10.1051/shsconf/20207805005)>

GONON L. & SORBA J. (2020 [2019]). Phraséologismes spécifiques dans les romans historiques et les romans de littérature blanche. *Journal of French Language Studies*, 30.1 : 1-20. <<https://doi.org/10.1017/S0959269519000140>>

SORBA J. (2020a). Les expressions spatiales : des unités phraséologiques pour distinguer les genres romanesques ? In L. Fesenmeier & I. Novakova (éd.), *Phraséologie et stylistique de la langue littéraire / Phraseology and Stylistics of the Literary Language* (p. 31-48). Peter Lang.

SORBA J. (2020b). Note de lecture. Wendy Ayres-Bennett, Anne Carlier, Julie Glikman, Thomas M. Rainsford, Gilles Siouffi & Carine Skupien Dekens (dir.), *Nouvelles voies d'accès au changement linguistique*. Classiques Garnier, Paris, 2018, 548 p. *LIDIL*, 62. <<https://doi.org/10.4000/lidil.8191>>

SORBA J., GONON L., DYKA S. & GOOSSENS V. (2020). Reading and Writing as Motifs in English and French Literary Fiction. In D. Siepman & I. Novakova (éd.), *Phraseology and Style in Subgenres of the Novel. A Synthesis of Corpus and Literary Perspectives* (p. 223-249). Palgrave Macmillan.

SORBA J. & NOVAKOVA I. (2020). Complexité sémantique et scénarii discursifs du repentir. *Travaux de Linguistique*, 81 : 31-56.

## 2019

BOCH F., SORBA J. & BESSONNEAU P. (2019). Quelle conscience de mes (in)compétences à l'écrit ? Étude du "degré de lucidité" chez les étudiants. In M. Niwese, J. Lafont-Terranova & M. Jaubert (éd.), *Écrire et faire écrire dans l'enseignement postobligatoire. Enjeux, modèles et pratiques innovantes* (p. 79-94). Presses Universitaires du Septentrion.

BOCH F. & SORBA J. (2019). L'évaluation de la compétence lexicale : étude des distracteurs. Communication présentée au Colloque *Recherche en didactique du lexique : avancées, réflexions, méthodes*, 87<sup>e</sup> congrès de l'ACFAS, Université du Québec en Outaouais, Gatineau (Canada), 29 mai 2019.

DENOYELLE C. & SORBA J. (2019a). La phraséologie appliquée aux dialogues. Communication présentée au Colloque *Diachro IX Le français en diachronie*, Universidad de Salamanca (Espagne), 28-30 mars 2019.

DENOYELLE C. & SORBA J. (2019b). Phraséologie du roman médiéval : une approche sur corpus outillés. Communication affichée au *Global Challenges Science Week. International interdisciplinary days of Grenoble Alpes*, Grenoble – World Trade Center, 3-6 juin 2019.

DENOYELLE C. & SORBA J. (2019c). Phraséologie et genres textuels : étude des phraséologismes construits autour des verbes *doner* et *metre* dans le roman médiéval. Communication présentée au Colloque *JLC 2019 – 10<sup>e</sup> Journées Internationales de la Linguistique de Corpus*, Université Grenoble Alpes (France), 26-28 novembre 2019.

GONON L. & SORBA J. (2019). Phraséologismes autour du vin et du champagne dans le roman contemporain. Communication présentée au Colloque *Terminologies et discours gastronomiques et œnologiques. Le vin et les autres produits*, Faculté de philologie de l'Université d'Innsbrück (Autriche), 19-20 septembre 2019.

SORBA J. (2019a). Les expressions spatiales : des unités phraséologiques pour distinguer les genres romanesques ? Communication présentée au Colloque *Phraseology and Stylistics of Literary Language*, Friedrich-Alexander Universität Erlangen – Nürnberg (Allemagne), 13-15 mars 2019.

SORBA J. (2019b). Note de lecture. Jean-Michel Adam, 2018 : *Le paragraphe : entre phrases et texte*. Paris : Armand Colin. 224 p. *LIDIL*, 59. <<https://doi.org/10.4000/lidil.6170>>

SORBA J. (2019c). Compte-rendu. Raluca NITA & Freiderikos VALETOPOULOS (éds.). — *L'expression des sentiments. De l'analyse linguistique aux applications*. Rennes, Presses Universitaires de Rennes, 2018, 379 p. *Bulletin de la Société de linguistique de Paris*, 114.2.

SORBA J. (2019d). Note de lecture. Georgeta Cislaru & Thierry Olive, 2018 : *Le processus de textualisation. Analyse des unités linguistiques de performance écrite*. Louvain-la-Neuve : De Boeck Supérieur. 272 p. *LIDIL*, 60. <<https://doi.org/10.4000/lidil.6753>>

SORBA J. (2019e). Table ronde 1 : L'enseignement des LCA dans un établissement : donner du sens au parcours de l'élève. *Séminaire L'antiquité pour demain*, Académie de Grenoble, Université Grenoble Alpes, 19 mars 2019.

## 2018

CAVALLA C. & SORBA J. (2018). Étude diachronique du figement : collocations verbo-nominales. In O. Soutet, I. Sfar & S. Mejri (éd.), *Phraséologie et discours* (p. 131-142). Honoré Champion.

DENOYELLE C. & SORBA J. (2018). Le Lexicoscope & première expérimentation sur les formules de congé. Communication présentée au Colloque *Nouvelles perspectives autour de la phraséologie de la langue littéraire médiévale*, Université Grenoble Alpes (France), 29-30 novembre 2018.

GONON L., GOOSSENS V., KRAIF O., NOVAKOVA I. & SORBA J. (2018). Motifs textuels spécifiques au genre policier et à la littérature “blanche”. In F. Neveu, B. Harmegnies, L. Hriba & S. Prévost (éd.), *6<sup>e</sup> Congrès Mondial de Linguistique Française*, vol. 46, SHS Web of Conferences. EDP Sciences. <<https://doi.org/10.1051/shsconf/20184606007>>

GONON L. & SORBA J. (2018). Phraséologismes spécifiques dans les romans historiques et les romans de littérature blanche. Communication présentée au Colloque *LTT 2018, 11<sup>e</sup> journées du réseau Lexicologie, Terminologie, Traduction : « Lexique(s) et genre(s) textuel(s) : approches sur corpus »*, Université Grenoble Alpes (France), 25-28 septembre 2018.

KRAIF O. & SORBA J. (2018). Spécificités des expressions spatiales et temporelles dans quatre sous-genres romanesques (policier, science-fiction, historique et littérature générale). In D.F. Iezzi, L. Celardo & M. Misuraca (éd.), *JADT 2018 Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*, vol. 1 (p. 392-399).

NOVAKOVA I. & SORBA J. (2018). La construction du sens autour des lexies d’affect : proposition d’un modèle fonctionnel. *Langages*, 210 : 55-70.

NOVAKOVA I., SORBA J. & TUTIN A. (2018). Les noms généraux *sentiment* et *émotion*. *Langue française*, 198 : 83-97.

SORBA J. (2018a). Changements linguistiques dans les collocations verbo-nominales construites autour de *feu*. In P. Blumenthal & D. Vigier (éd.), *Études diachroniques du français et perspectives sociétales* (p. 269-294). Peter Lang.

SORBA J. (2018b). *La Langue des émotions - XVI<sup>e</sup>-XVIII<sup>e</sup> siècle*, Véronique Ferrer et Catherine Ramon (éd.), Paris, Garnier, 2017. *Le français préclassique*, 20 : 170-176.

SORBA J. (2018c). Note de lecture. Francis Grossmann, Salah Mejri & Inès Sfar (dir.), *La phraséologie : sémantique, syntaxe, discours*, Honoré Champion, 2017, 284 p. *LIDIL*, 58. <<https://doi.org/10.4000/lidil.5350>>

**SORBA J. & NOVAKOVA I.** (2018). Les stratégies argumentatives autour des émotions : le cas de *jubilation* et de *dédain*. In T. Hermann, J. Jacquin & S. Oswald (éd.), *Les mots de l'argumentation* (p. 139-157). Peter Lang.

## 2017

BOCH F. & **SORBA J.** (2017a). Le projet CertiRedac. Communication affichée au 3<sup>e</sup> salon *Innovatives SHS* CNRS, Marseille-Parc des Expositions, 17-18 mai 2017.

BOCH F. & **SORBA J.** (2017b). Le projet CertiRedac. Conférence invitée à la Journée d'étude *L'accompagnement des étudiants en matière d'expression écrite*, Université de Toulon, 8 juin 2017.

BOCH F., **SORBA J.** & BESSONNEAU P. (2017a). L'évaluation de la compétence lexicale chez les adultes francophones Comment intégrer l'inférence ? Communication présentée au *Colloque en didactique du lexique*, 85<sup>e</sup> congrès de l'ACFAS, Université McGill, Montréal (Canada), 8-9 mai 2017.

BOCH F., **SORBA J.** & BESSONNEAU P. (2017b). Travailler la syntaxe dans les formations à l'écrit professionnel. Communication présentée au Colloque *Les écrits professionnels : enjeux socioéconomiques et enjeux de formation*, 85<sup>e</sup> congrès de l'ACFAS, Université McGill, Montréal (Canada), 11 mai 2017.

CAVALLA C. & **SORBA J.** (2017). *Prendre un bain, des risques ou la fuite* : étude diachronique du figement. In F. Grossmann, S. Mejri & I. Sfar (éd.), *Phraséologie : sémantique, syntaxe, discours* (p. 185-199). Honoré Champion.

NOVAKOVA I. & **SORBA J.** (2017). À la recherche des motifs autour de *meurtre, crime, assassinat* dans le roman policier contemporain. *Synergies Pologne*, 14 : 29-42. Disponible en ligne <[http://gerflint.fr/Base/Pologne14/novakova\\_sorba.pdf](http://gerflint.fr/Base/Pologne14/novakova_sorba.pdf)>

**SORBA J.** (2017). Note de lecture. Pierre Chiron & Charles Guérin (dir.), *L'infraction stylistique et ses usages théoriques de l'Antiquité à nos jours*, Presses Universitaires de Rennes, 2016, 332 p. *LIDIL*, 56. <<https://doi.org/10.4000/lidil.4717>>

**SORBA J.** (2017). Du grec dans le Lexicoscope ! Communication présentée au Colloque international *Tragédie grecque et numérique : pour une exploration phraséologique du corpus tragique*, Université Grenoble Alpes (France), 15 décembre 2017.

## 2016

AUTIN L., BOUZIDI K., KRAIF O. & **SORBA J.** (2016). *Sine nomine vulgus* : étude contrastive des profils combinatoires des noms de la foule à partir d'un corpus arboré latin. *JADT 2016, Journées internationales d'Analyse statistiques des Données Textuelles*, Nice (France), 7-10 juin 2016.

BOCH F. & **SORBA J.** (2016). Le projet CertiRedac. Conférences invitées à 7<sup>e</sup> rencontres Pôle Grenoble Cognition « Sortie de laboratoire » (Grenoble, 2 juin 2016) ; Séminaire annuel LIDILEM « Mon projet en 180 secondes » (Grenoble, 28 juin 2016) ; Journée de Valorisation de la recherche en SHS Université Grenoble Alpes (Grenoble, 18 juin 2016) ; Journée de Valorisation « Inauguration Linksum » (Grenoble, 2 décembre 2016).

BOCH F., **SORBA J.** & BESSONNEAU P. (2016a). Évaluer les compétences rédactionnelles : que tester ? *Le Français aujourd'hui*, 193 : 128-142.

BOCH F., **SORBA J.** & BESSONNEAU P. (2016b). Les étudiants et leur perception de leurs compétences à l'écrit : étude du "degré de lucidité". Communication présentée au Colloque international *Enseignement et apprentissage de l'écriture de la maternelle à l'université et dans les formations tout au long de la vie*, Bordeaux, ÉSPÉ d'Aquitaine – Université de Bordeaux (France), 19-21 octobre 2016.

GONON L., KRAIF O., NOVAKOVA I., PIAT J. & **SORBA J.** (2016). *Sur la scène de crime...* Enquête sur les enjeux linguistiques et stylistiques de motifs récurrents dans le thriller contemporain. *5<sup>e</sup> Congrès Mondial de Linguistique Française*, SHS Web of Conferences, vol. 27. EDP Sciences. <<https://doi.org/10.1051/shsconf/20162706006>>

KRAIF O., NOVAKOVA I. & **SORBA J.** (2016). Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction. *LIDIL*, 53 : 143-159. <<https://doi.org/10.4000/lidil.3976>>

NOVAKOVA I. & **SORBA J.** (2016a). Les lexies d'émotion et la construction du sens. Communication présentée à la Journée Consilia *Vers une sémantique discursive : propositions théoriques et méthodologiques*, Université de Paris 3 – Sorbonne Nouvelle (France), 5 février 2016.

NOVAKOVA I. & **SORBA J.** (2016b). À la recherche de motifs autour de *meurtre, crime, assassinat* dans le roman policier contemporain. Communication présentée à la journée d'étude *De la phraséologie aux genres textuels : état des recherches et perspectives méthodologiques*, Université Pédagogique de Cracovie (Pologne), 18 octobre 2016.

**SORBA J.** (2016a). Review on Marie-Claire Beaulieu, *The Sea in the Greek Imagination*. Philadelphia (PE): University of Pennsylvania Press, 2016. *The Ancient History Bulletin, Online Reviews*, Vol. 6: 81-84. <<http://ancienthistorybulletin.org/online-reviews-vol-6>>

**SORBA J.** (2016b). Changements linguistiques dans les collocations verbo-nominales construites autour de *feu(x)*. Communication présentée au Colloque international *Changements linguistiques et phénomènes sociétaux*, ENS-Lyon (France), 9-11 mars 2016.

**SORBA J.** (2016c). Le vocabulaire du corps en latin et en grec : une approche linguistique. Conférence invitée lors de la *Journée d'études autour du programme de Culture Antique aux concours d'entrée des ENS*, Université Grenoble Alpes, 18 mars 2016.

**SORBA J. & GOOSSENS V.** (2016). Le rôle du figement dans le traitement de la synonymie au sein du champ de la colère. *Linguisticæ Investigationes*, 39.1 : 1-27.

## 2015

AUTIN L., BOUZIDI K., KRAIF O. & **SORBA J.** (2015). Phraseotext : annotation syntaxique et mise en ligne d'un corpus latin (stylistique et phraséologie). Communication affichée au Colloque international *Humanités numériques et Antiquité*, Grenoble MSH-Alpes, 2-4 septembre 2015.

BOCH F. & **SORBA J.** (2015). *Le Certificat en Compétences Rédactionnelles (CCR)*. Invention ayant fait l'objet d'un dépôt de sept enveloppes Soleau et d'un dépôt AAP pour la plateforme logicielle. <<https://www.youtube.com/watch?v=TFxxU3XhE44>>

BOCH F., **SORBA J.** & BESSONNEAU P. (2015). Évaluer les compétences rédactionnelles. Que tester ? Communication présentée aux Journées d'étude *L'écriture de haut niveau : de la conception du texte à sa réalisation*, Marseille (France), 3-4 septembre 2015.

KRAIF O., NOVAKOVA I. & **SORBA J.** (2015). Unités polylexicales spécifiques dans deux sous-genres littéraires : le roman policier et la science-fiction. Communication présentée à *4<sup>th</sup> International Conference on Grammar & Text - GRATO 2015*, Lisbonne (Portugal), 2-4 juillet 2015. Présentée également à la Journée franco-polonaise, lundi 19 octobre 2015, organisée par l'Axe 1 du LIDILEM et l'Institut de Lettres et de Langues Modernes de l'Université pédagogique de Cracovie.

**SORBA J.** (2015a). Variations sémantiques autour de la mer, d'Homère à Eschyle. *GAIA Revue interdisciplinaire sur la Grèce archaïque*, 18. *Πολυφόρβη Γαίη. Mélanges de littérature et de linguistique offerts à Françoise Létoublon* : 311-323.

**SORBA J.** (2015b). Note de lecture. Pierre-Yves Testenoire [éd.], *Ferdinand de Saussure. Anagrammes homériques ; Ferdinand de Saussure à la recherche des anagrammes*, Limoges,

Lambert-Lucas, 2013. *Texto ! Textes et cultures*, vol. XX.1, coordonné par Carine Duteil-Mougel. <<http://www.revue-texto.net/index.php?id=3622>>

**SORBA J.** (2015c). Compte-rendu. *Bilinguisme gréco-latin et épigraphie*. Édités par Frédérique Biville, Jean-Claude Decourt et Georges Rougemont. Lyon. Maison de l'Orient et de la Méditerranée. 2008. *Agora*, 7, *Les comptes rendus de GAIA*.

**SORBA J. & NOVAKOVA I.** (2015). Les stratégies argumentatives autour des émotions : le cas de *jubilation* et de *dédain*. Communication présentée au Colloque international : *Argumentation & Langage : marqueurs linguistiques, processus discursifs, opérations cognitives*, Université de Lausanne (Suisse), 9-11 septembre 2015.

## 2014

**BOCH F. & SORBA J.** (2014). Le projet CertiRedac. Conférence invitée à la Journée de Valorisation en LLASHS, Université Stendhal Grenoble 3, 11 juin 2014.

**CAVALLA C. & SORBA J.** (2014). Étude diachronique du figement : collocations verbo-nominales. Communication présentée à la *Conférence de l'Europhras 2014 : La phraséologie : ressources, descriptions et traitements informatiques*, Paris (France), 10-12 septembre 2014.

**NOVAKOVA I. & SORBA J.** (2014a). L'évaluation à travers les émotions : le cas d'*estime* et de *déception*. *Langue Française*, 184 : 75-91.

**NOVAKOVA I. & SORBA J.** (2014b). Les adjectifs d'affect : complexité sémantique et profils discursifs, *4<sup>e</sup> Congrès Mondial de Linguistique Française*, vol. 8, SHS Web of Conferences, (p. 637-649). EDP Sciences. <<https://doi.org/10.1051/shsconf/20140801193>>

**NOVAKOVA I. & SORBA J.** (2014c). L'émotion dans le discours. À la recherche du profil discursif de *stupeur* et de *jalousie*. In P. Blumenthal, I. Novakova & D. Siepmann (éd.), *Les émotions dans le discours. Emotions in discourse* (p. 161-173). Peter Lang.

**SORBA J.** (2014a). La polysémie comme artefact des poètes. Variations sémantiques de *samudrá-* dans la *Ṛksaṃhitā*. *Études Romanes de Brno*, 35.2 : 39-51. <<http://digilib.phil.muni.cz/handle/11222.digilib/132857>>

**SORBA J.** (2014b). Le rivage de la mer ou la ligne des brisants ? Vie et mort du lexème grec ῥηγμῖν. In D. Ablali, S. Badir & D. Ducard (éd.), *Documents, textes, œuvres : perspectives sémiotiques* (p. 249-261). Presses Universitaires de Rennes.

**SORBA J. & BROQUET S.** (2014). Présentation : Sens multiple(s) et polysémie. Regards d'Orient. *Études Romanes de Brno*, 35.2 : 5-11. <<http://digilib.phil.muni.cz/handle/11222.digilib/132855>>

SORBA J. & CUSIMANO C. (2014). Présentation : Sens multiple(s) et polysémie. Regards d'Occident. *Études Romanes de Brno*, 35.1 : 5-8. <<http://digilib.phil.muni.cz/handle/11222.digilib/130367>>

SORBA J. & NOVAKOVA I. (2014). Complexité sémantique et scenarii discursifs du repentir. Communication présentée au Colloque international *Approches pluridisciplinaire de repentir : formes, médiums, enjeux*, Chambéry (France), 9-11 avril 2014.

## 2013

CAVALLA C. & SORBA J. (2013). Étude diachronique du figement : trois collocations du quotidien. Communication présentée au Colloque international *Premières rencontres phraséologiques. Des collocations aux séquences figées*, Grenoble MSH-Alpes (France), 13-15 novembre 2013.

NOVAKOVA I. & SORBA J. (2013a). *Stupéfier* et *jalouser* dans les séquences textuelles journalistiques : quel profil discursif pour quelle stratégie argumentative ? *Le discours et la langue. Revue de linguistique française et d'analyse du discours*, 4.1 : 203-220.

NOVAKOVA I. & SORBA J. (2013b). Argumentation et émotion dans les séquences textuelles journalistiques. Le cas de *stupeur* et de *jalousie*. In T. Muryn, S. Mejri, W. Prazuch & I. Sfar (éd.), *La phraséologie entre langues et cultures. Structures, fonctionnements, discours* (p. 137-149). Peter Lang.

NOVAKOVA I. & SORBA J. (2013c). Argumentation et émotion dans les séquences textuelles journalistiques. À la recherche du profil discursif de *stupeur* et de *jalousie*. Communication présentée au Colloque international *Nouvelles perspectives en sémantique lexicale et en organisation du discours*, Osnabrück (Allemagne), 6-8 février 2013.

SORBA J. (2013a). La polysémie comme artefact des poètes. Variations sémantiques de *samudrá-* dans la *Ṛksamhitā*. Communication présentée au Colloque international *Sens multiple(s) et polysémie : perspectives croisées, Orient & Occident*, Aix-en-Provence (France), 4-6 juin 2013.

SORBA J. (2013b). Note de lecture. *Revue Nordiques n°24 (Les langues nordiques à l'ère de la mondialisation)*. *LIDIL*, 48 : 209-210. <<https://doi.org/10.4000/lidil.3334>>

SORBA J. & GOOSSENS V. (2013). *Vert de rage, noire fureur*. Propriétés des figements au sein du champ de la colère. Communication présentée au XXXV<sup>e</sup> Colloque international de linguistique fonctionnelle, Brno (République Tchèque), 18-22 septembre 2013.

## 2012

NOVAKOVA I. & SORBA J. (2012). *Stupéfier et jalouser* dans les séquences textuelles journalistiques : quel profil discursif pour quelle stratégie argumentative ? Communication présentée aux Journées d'étude *L'émotion argumentée*, Lyon (France), 6-7 septembre 2012.

SORBA J. (2012a). Le rivage de la mer ou la ligne des brisants ? Vie et mort du lexème grec ῥηγμῖν. Communication présentée au Colloque international *Textes, documents, œuvres. Perspectives sémiotiques. Autour de Fr. Rastier*, Cerisy-la-Salle (France), 6-13 juillet 2012.

SORBA J. (2012b). La langue qui ne voulait pas de l'écriture : le cas de l'indo-aryen ancien. Conférence invitée à *Atelier de linguistique comparée des langues indo-européennes : « Rapports entre systèmes graphiques et systèmes linguistiques : quelques exemples de regards neufs en védique, tokharien, celtibère »*, HISOMA - Université Lyon 2 – MOM, 6 décembre 2012.

## 2011

SORBA J. (2010-2011). L'océan des origines ? Étude du lexème *salilá-* dans les *saṃhitā* du *R̥gveda* et de l'*Atharvaveda* recension *śaunaka*. *Bulletin d'Études Indiennes*, 28-29 : 171-185.

SORBA J. (2011a). Présentation de thèse. Le vocabulaire de la mer. Étude comparée en indo-aryen ancien, grec ancien et latin. *L'information grammaticale*, 131 : 52-53.

SORBA J. (2011b). L'océan des origines ? Étude du lexème *salilá-* dans les *saṃhitā* du *R̥gveda* et de l'*Atharvaveda* recension *śaunaka*. Communication présentée à la *I<sup>ère</sup> Journée de l'Inde*, Université de Provence – MMSH Aix-en-Provence (France), 28 mai 2011.

## 2010

SORBA J. (2010a). *Le vocabulaire de la mer. Étude comparée en indo-aryen ancien, grec ancien et latin*. Thèse de doctorat, dir. G.-J. Pinault. École Pratique de Hautes Études, Paris, 796 p.

SORBA J. (2010b). Le vocabulaire de la mer dans la *Rudens* de Plaute. *L'information grammaticale*, 127 : 24-27.

SORBA J. (2010c). Le vocabulaire de la mer dans quelques *saṃhitā* védiques : le cas du lexème *arṇavá-*. Communication présentée à la *V<sup>e</sup> Journée Monde Indien*, Ivry-sur-Seine UMR 7528 Mondes iranien et indien, 19 mars 2010.

## *Avant 2010*

SORBA J. (2008-2009). Le vocabulaire de la mer dans quelques *saṃhitā* védiques : le cas du lexème *arṇavá-*. *Bulletin d'Études Indiennes*, 26-27 : 193-209.

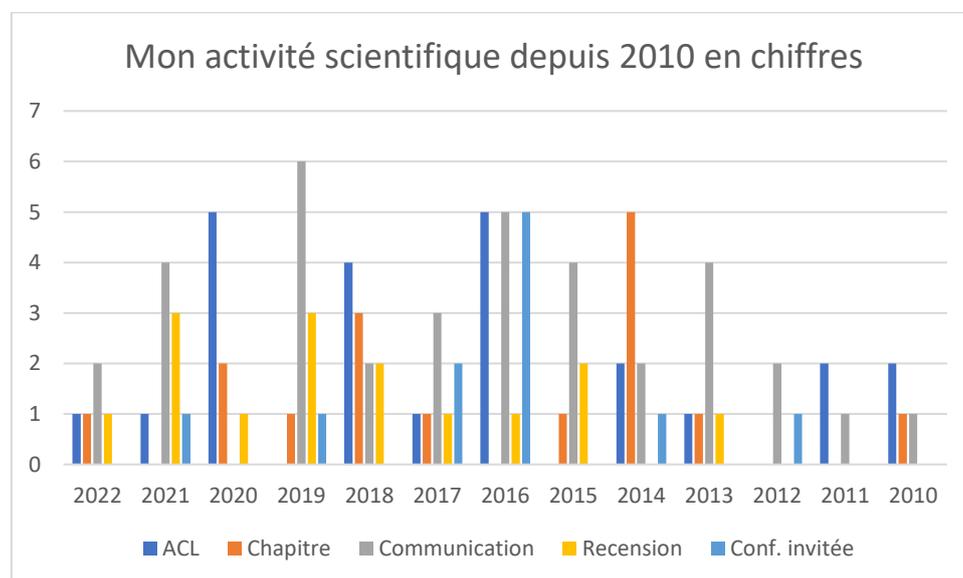
**SORBA J.** (2008a) (dir.). *L'Explication de textes en langues anciennes : préparation aux concours*. Ellipses (240 p.). Direction de l'ouvrage et réalisation du préambule, des annexes et des exercices p. 9-14, 21-26, 61-66, 74-79, 86-103.

**SORBA J.** (2008b). La mer tragique et l'héritage homérique. Étude des lexèmes ἄλς, θάλασσα, πέλαγος et πόντος dans les tragédies d'Eschyle. *Recherches et Travaux hors-série n°16 : L'Antiquité en ses confins. Mélanges en l'honneur de Benoît Gain* (p. 139-149). ELLUG. Republié dans *Texto !* <<http://www.revue-texto.net/index.php?id=2084>>

**SORBA J.** (2005). La transmission d'Homère de l'Antiquité à la Renaissance. Conférence invitée au *Séminaire Mondes Anciens et Médiévaux*, HISCANT-MA, Université Nancy 2.

**SORBA J.** (2002). *Étude du vocabulaire de la mer et de ses représentations dans quatre œuvres de la littérature indienne*. Mémoire de DEA Civilisations des mondes anciens, dir. C. Chojnacki, Université Jean Moulin – Lyon 3, 147 p.

**SORBA J.** (1998). *La perception de l'espace marin dans la poésie grecque archaïque*. Mémoire de Maitrise de Lettres classiques, dir. F. Letoublon, Université Stendhal – Grenoble 3, 237 p.



## II. Références bibliographiques

### A

Aarts B. (2000). Corpus linguistics, Chomsky and Fuzzy Tree Fragments. In C. Mair & M. Hundt (éd.), *Corpus Linguistics and Linguistic Theory* (p. 5-13). Rodopi.

Ablali D. (2013). Types, genres et généricité en débat avec Jean-Michel Adam. *Pratiques* 157-158 : 216-232.

Ablali D., Bouhouhou A. & Tebbaa O. (dir.) (2015). *Les genres textuels, une question d'interprétation ?* Lambert-Lucas.

Adam J.-M. (1993). Le texte et ses composantes. Théorie d'ensemble des plans d'organisation. *Semen* 8 <<https://doi.org/10.4000/semen.4341>>

Adam J.-M. (1997). Genres, textes, discours : pour une reconception linguistique du concept de genre. *Revue belge de philologie et d'histoire*, 75 : 665-681.

Adam J.-M. (1998). Les genres du discours épistolaires. De la rhétorique à l'analyse pragmatique des pratiques discursives. In J. Siess (éd.), *La lettre entre réel et fiction* (p. 37-53). SEDES.

Adam J.-M. (2011). *Les textes : types et prototypes* (3<sup>e</sup> éd.). Armand Colin.

Adam J.-M. (dir.) (2015). *Faire texte. Frontières textuelles et opérations de textualisation*. Presses Universitaires de Franche-Comté.

Adam J.-M. & Heidmann U. (2004). Des genres à la généricité. L'exemple du conte (Perrault et les Grimm). *Langages*, 153 : 62-72.

Alaux J. (2006). *Lectures tragiques d'Homère*. Belin.

Agulhon M. (1980). 1830 dans l'histoire du XIX<sup>e</sup> siècle français. *Romantisme. Revue du dix-neuvième siècle*, 28-29 : 15-27.

Amossy R. (2006). *L'argumentation dans le discours*. 2<sup>e</sup> éd. Armand Colin.

Andries L. (2015). La Bibliothèque bleue, une littérature éphémère ? *Fabula / Les colloques, Les éphémères, un patrimoine à construire*. <<https://www.fabula.org/colloques/document2919.php>>

Anscombe J.-C. (1995). Morphologie et représentation événementielle : le cas des noms de sentiment et d'attitude. *Langue française*, 105 : 40-54.

Anscombe J.-C. (2011). Figement, idiomaticité et matrices lexicales. In J.-C. Anscombe & S. Mejri (éd.), *Le figement linguistique : la parole entravée* (p. 17-40). Honoré Champion.

Anscombe J.-C. & Mejri S. (2011). *Le figement linguistique : la parole entravée*. Honoré Champion.

Apothéloz D. (2002). *La construction du lexique français*. Ophrys.

## B

Bakhtine, M. (1984 [1952-1953]). *Esthétique de la création verbale*, trad. du russe par A. Aucouturier. Gallimard.

Bally C. (1909). *Traité de stylistique française*. Premier volume (2<sup>nd</sup>e éd.). Klincksieck. En ligne <<https://archive.org/details/traitdestylist01ball/page/66/mode/2up>>

Baroni R. (2003). Genres littéraires et orientations de la lecture. Une lecture modèle de « La mort et la boussole » de J. L. Borges. *Poétique*, 134 : 141-157.

Baroni R. (2009). Généricité ou stéréotypie ? *Cahiers de Narratologie. Analyse et théories narratives*, 17. <[10.4000/narratologie.1090](https://doi.org/10.4000/narratologie.1090)>

Baroni R. & Macé M. (dir.) (2007). *Le savoir des genres*. Presses Universitaires de Rennes.

Baudou J. (2005). *La fantasy*. Presses Universitaires de France.

Bazin-Tacchella S. & Souvay G. (2021). Lemmatisation et construction automatique de ressources lexicographiques : les développements du lemmatiseur LGeRM. *Diachroniques*, 8 : 121-143.

Beebee T.O. (1994). *The ideology of genre. A comparative study of generic instability*. Penn State Press.

Benson M., Benson E. & Ilson R. (1997). *The BBI Dictionary of English Word Combinations*. (2<sup>nd</sup>e éd.). John Benjamin.

Bentolila, É. (2016). *Le roman policier français de 1970 et 2000 : une analyse littéraire*. Thèse de doctorat. Université Grenoble Alpes.

Benveniste E. (1969). *Le vocabulaire des institutions indo-européennes* (2 tomes). Les éditions de Minuit.

Biber D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

Biber D. (1989). A typology of English texts. *Linguistics*, 27 : 3-43.

Biber D. (1993a). Using Register-diversified Corpora for General Language Studies. *Computational Linguistics*, 19: 219-241.

Biber D. (1993b). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4): 243-257.

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

Billardeau M. & Porcher M.-C. (dir.) (1999). *Le Rāmāyaṇa de Vālmīki*. Gallimard.

Blumenthal P. (2002a). Les péchés capitaux : profil combinatoire et dimension textuelle. In M. Keşik (éd.), *Référence discursive dans les langues romanes et slaves* (p. 29-45). Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.

Blumenthal P. (2002b). Profil combinatoire des noms. Synonymie distinctive et analyse contrastive. *Zeitschrift für französische Sprache und Literatur*, 112.2 : 115-138.

Blumenthal P. (2006). *Wortprofil im Französischen*. Niemeyer.

Blumenthal P. (2014). Caractéristiques et effets de la complexité sémantique de noms d'affect. In P. Blumenthal, I. Novakova & D. Siepmann (éd.), *Les émotions dans le discours / Emotions in Discourse* (p. 175-186). Peter Lang.

Bouquet S. (2004). Linguistique générale et linguistique des genres. *Langages*, 153 : 3-14.

Boyer A.-M. (2008). *Les paralittératures*. Armand Colin.

Bozzetto R. (2007). *La science-fiction*. Armand Colin.

Bronckart J.-P. (1996). Genres de textes, types de discours et opérations psycholinguistiques, *Voies livres*, 78 : 1-20.

Bronckart J.-P., Bain D., Schneuwly B., Davaud C. & Pasquier A. (1985). *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse*. Delachaux & Niestlé.

Brunetière F. (1889). *Questions de critique*. Calmann Lévy.

Burger, H. (1998). *Phraseologie: Eine Einführung am Beispiel des Deutschen*. Erich Schmidt.

Buridant C. (1989). L'approche diachronique en phraséologie : quelques aspects de l'ancien et du moyen français. *Travaux de linguistique et de philologie*, XXVII : 127-149.

## C

Carel M. (2011). La polyphonie linguistique. *Transposition. Musique et sciences sociales*, 1. <<https://doi.org/10.4000/transposition.365>>

Chantraine P. (1968). *Dictionnaire étymologique de la langue grecque. Histoire des mots*. Klincksieck.

Charaudeau P. & Maingueneau D. (dir.) (2002). *Dictionnaire d'analyse du discours*. Seuil.

Charolles, M. (1989). Problèmes de la cohérence textuelle. In M. Charolles, J.-F. Halté, C. Masseron & A. Petitjean (éd.), *Pour une didactique de l'écriture* (p. 49-83). Centre d'analyse syntaxique de l'Université de Metz.

Chupeau J. (1987). Puget de la Serre et l'esthétique épistolaire : les avatars du « Secrétaire de la Cour ». *Cahiers de l'Association internationale des études françaises*, 39 : 111-126. <<https://doi.org/10.3406/caief.1987.2428>>

Combettes B. (2003) (dir.). *Évolution et variation en français préclassique : études de syntaxe*. Honoré Champion.

Combettes B. (2012). Linguistique textuelle et diachronie. *3<sup>e</sup> Congrès Mondial de Linguistique Française*. SHS Web of Conferences, vol. 1 (p. 3-10). EDP Sciences. <<https://doi.org/10.1051/shsconf/20120100344>>

Compagnon A. (2000). La notion de genre : cours d'Antoine Compagnon. *Fabula* <<http://www.fabula.org/compagnon>>

Cori M. & David S. (2008). Les corpus fondent-ils une nouvelle linguistique ? *Langages*, 171 : 111-129.

Coseriu E. (2007). *Lingüística del texto. Introducción a la hermenéutica del sentido*, édition et annotation d'Oscar Loureda Lamas. Arco/Libros.

Coulet H. (1991). *Le roman jusqu'à la Révolution*. Armand Colin.

## D

Dain A. (1965). *Traité de métrique grecque*. Klincksieck.

Derrida J. (1986). La loi du genre, *Parages*. Galilée.

Dijk, T.A. van (1972). *Some Aspects of Text Grammar*. Mouton

Dion R., Fortier F. & Haghebaert É. (dir.) (2001). La dynamique du genre, suite conclusion. In R. Dion, F. Fortier & É. Haghebaert (éd.), *Enjeux des genres dans les écritures contemporaines* (p. 351-361). Nota Bene. <<https://numerique.banq.qc.ca/patrimoine/details/52327/2437133>>

Diwersy S. & Legallois D. (2019). L'apport de la méthodologie des motifs aux analyses phraséologiques en discours. In M. Kauffer & Y. Keromes (éd.), *Theorie und Empirie in der Phraseologie. Approches théoriques et empiriques en phraséologie* (p. 115-132). Stauffenburg Verlag.

Diwersy S., Goossens V., Grutschus A., Kern B., Kraif O., Melnikova E. & Novakova I. (2014). Traitement des lexies d'émotion dans les corpus et les applications d'EmoBase. *Corpus*, 13 : 269-293. <<https://doi.org/10.4000/corpus.2537>>

Diwersy S., Gonon L., Goossens V., Kraif O., Novakova I., Sorba J. & Vidotto I. (2021). La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase. *Corpus*, vol.22 <<https://doi.org/10.4000/corpus.6101>>

Droysen J.G (1967). *Historik*. R. Hübner.

Dubois J., Giacomo M., Guespin L., Marcellesi C., Marcellesi J.-B. & Mével J.-P. (1994). *Dictionnaire de linguistique et des sciences du langage*. Larousse.

Dubreil E. (2008). Collocations : définitions et problématiques. *Texte !*, XIII.1. <<http://www.revue-texto.net/index.php?id=126>>

## É

Equoy Hutin S. (2007). Places et fonctions du genre épistolaire dans les lettres d'allocataires d'aides sociales. In R. Baroni & M. Macé (éd.), *Le savoir des genres* (p. 119-138). Presses Universitaires de Rennes.

## F

Fesenmeier L. (2010). « Se souvenir » en français et en italien : différence(s) de centrage. In M. Iliescu, H. Siller-Runggaldier & P. Danler (éd.), *Actes du XXV<sup>e</sup> Congrès International de Linguistique et de Philologie Romanes*, Innsbrück 2007, tome III (p. 85-95). De Gruyter.

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

Fillmore C., Kay P. & Kay O'Connor M. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, 64 : 501-38.

Firth J.R. (1957 [1951]). Modes of Meaning. In *Essays and Studies*, The English Association. Reprinted in *Firth J.R. (1957), J. Papers in Linguistics 1934–1951* (p. 190-215). Oxford University Press.

Fløttum K., Dahl T. & Kinn T. (2006). *Academic Voices: Across languages and disciplines*. John Benjamin.

Fónagy I. (1982). *Situation et signification*. John Benjamins.

Fontanier P. (1977 [1821]). *Les figures du discours*. Flammarion.

Foucault M. (1969). *L'archéologie du savoir*. Gallimard.

Francis W.N. (1992). Language Corpora B.C. In J. Svartvik (éd.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991* (p. 7-32). Mouton de Gruyter.

François J. (2004). Le fonctionnalisme linguistique et les enjeux cognitifs. In C. Fuchs (éd.), *La linguistique cognitive* (p.99-133). Ophrys.

Frath P. & Gledhill C. (2005). Qu'est-ce qu'une unité phraséologique ? In C. Bolly, J. Klein & B. Lamiroy (éd.), *La phraséologie dans tous ses états* (p. 11-25). Peeters.

## G

Genette G. (1982). *Palimpsestes. La littérature au second degré*. Seuil.

Gingras F. (2015). Les réécritures de « Lancelot du Lac » dans la longue durée : de la tradition manuscrite aux premiers imprimés. In D. Kullmann & S. Lalonde (éd.), *Réécritures*.

*Regards nouveaux sur la reprise et le remaniement de textes, dans la littérature française et au-delà, du Moyen Âge à la Renaissance* (p. 13-27). Pontifical Institute of Mediaeval Studies.

Gingras F. (2017). *Le Bâtard conquérant. Essor et expansion du genre romanesque au Moyen-Âge*. Honoré Champion.

González Rey I. (2015). *La phraséologie du français*. Presses Universitaires du Midi.

Gardes-Tamine J. & Hubert M.-C. (1993). *Dictionnaire de critique littéraire*. Armand Colin.

Granger S. & Meunier F. (2008). Introduction. The many faces of phraseology. In S. Granger & F. Meunier (éd.), *Phraseology: an interdisciplinary perspective* (p. XIX-XXVIII). John Benjamins.

Greimas A. J. (1986 [1966]). *Sémantique structurale*. Presses Universitaires de France.

Gries S. T. (2008). Phraseology and linguistic theory. A brief survey. In S. Granger & F. Meunier (éd.), *Phraseology: an interdisciplinary perspective* (p. 3-25). John Benjamins.

Grobol L. & Crabbé B. (2021). Analyse en dépendances du français avec des plongements contextualisés. *Actes de la 28<sup>e</sup> Conférence sur le Traitement Automatique des Langues Naturelles* (p. 106-114). ATALA.

Gross G. (1996). *Les expressions figées en français : noms composés et autres locutions*. Ophrys.

Gross G. (2012). *Manuel d'analyse linguistique. Approches sémantico-syntaxiques du lexique*. Presses Universitaires du Septentrion.

Gross M. (1998). La fonction sémantique des verbes supports. *Travaux de linguistique*, 37 (1) : 25-46.

Grosse E.-U. (2001). Évolution et typologie des genres journalistiques. Essai d'une vue d'ensemble. *Semen*, 13. <<https://doi.org/10.4000/semn.2615>>

Grossmann F., Mejri S. & Sfar I. (2017). Présentation. Phraséologie : sémantique, syntaxe, discours. In F. Grossmann, S. Mejri & I. Sfar (éd.), *La phraséologie : sémantique, syntaxe, discours* (p. 7-11). Honoré Champion.

Grossmann F. & Tutin A. (2005). Présentation. Sémantique des noms et adjectifs d'émotion. *LIDIL*, 32. <<https://doi.org/10.4000/lidil.83>>

Grossmann F., Gymnich M. & Siepmann D. (2020). Alcohol and tobacco consumption in English and French novels since the 1950s: A corpus-stylistic analysis. In D. Siepmann & I. Novakova (éd.), *Phraseology and Style in Subgenres of the Novel. A Synthesis of Corpus and Literary Perspectives* (p. 115-150). Palgrave Macmillan.

## H

Habert B. (2000). Des corpus représentatifs ; de quoi, pour quoi, comment ? *Cahiers de l'Université de Perpignan*, 31 : 11-58.

Habert B., Nazarenko A. & Salem A. (1997). *Les linguistiques de corpus*. Armand Colin.

Hamon P. (1993). *Du descriptif*. Hachette Supérieur.

Halliday M.A.K. & Hasan R. (1976). *Cohesion in English*. Longman

Hamburger K. (1977). *Die Logik der Dichtung*. Ernst Klett Verlag.

Hamburger K. (1986). *Logique des genres littéraires*, trad. fr. P. Cadiot. Seuil.

Haroche-Bouzinac G. (1995). *L'épistolaire*. Hachette.

Hausmann F. J. (2002). Isotopie, scénario, collocation et exemples lexicographiques. In M. Heinz (éd.), *L'exemple lexicographique dans les dictionnaires français contemporains. Actes des premières journées allemandes des dictionnaires*, Klingenberg am Main, 25-27 juin 2004 (p. 283-292). Niemeyer.

Hausmann F.J. & Blumenthal P. (2006). Présentation : collocations, corpus, dictionnaires. *Langue française*, 150 : 3-13.

Hegel G.W.F. (1979). *Esthétique*, trad. fr. S. Jankélévitch. Flammarion.

Heiden S. & Tournier M. (2001). Lexicométrie textuelle, sens et stratégie discursive. In J. J. de Bustos Tovar (éd.), *Simposio internacional de análisis del discurso* (p. 2287-2300). Visor.

Herman T. & Lugin G. (1999). La hiérarchie des rubriques : un outil de description de la presse. *Communication et langage*, 122 : 72-85.

Hoey M. (2005). *Lexical Priming. A new theory of words and language*. Routledge.

Holgado C., Lavrentiev A. & Constant M. (2021). Évaluation de méthodes et d'outils pour la lemmatisation automatique du français médiéval. *Actes de la 28<sup>e</sup> Conférence sur le Traitement Automatique des Langues Naturelles* (p. 153-161). ATALA.

Howarth, P. A. (1996). *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making*. Niemeyer.

Hunston S. & Francis G. (2000). *Pattern Grammar. A Corpus Driven Approach to the Lexical Grammar of English*. John Benjamins.

## J

Jacques M.-P. (2005). Pourquoi une linguistique de corpus ? In G. Williams (éd.), *La linguistique de corpus* (p. 21-30). Presses Universitaires de Rennes.

Jacques M.-P. & Tutin A. (dir.) (2018). *Lexique transversal et formules discursives des sciences humaines*. ISTE Editions.

Jacquin J. & Micheli R. (2012). Entre texte et interaction : propositions méthodologiques pour une approche discursive de l'argumentation en sciences du langage. *3<sup>e</sup> Congrès Mondial de Linguistique Française*. SHS Web of Conferences, vol. 1 (p. 599-610). EDP Sciences.  
<<https://doi.org/10.1051/shsconf/20120100128>>

Jauss H.R. (1970). Littérature médiévale et théorie des genres. *Poétique*, 1 : 79-101.

## K

Kern B. & Grutschus A. (2014). Surprise vs étonnement : comportement discursif et perspectives contrastives. In P. Blumenthal, I. Novakova & D. Siepmann (éd.), *Les émotions dans le discours / Emotions in Discourse* (p. 187-198). Peter Lang.

Kilgarriff A., Rychly P., Smrz P. & Tugwell D. (2004). The Sketch Engine. *Proceedings Euralex 2004* (p. 105-115). European Association for Lexicography.

Koelb C. (1975). The problem of "Tragedy" as a genre. *Genre: a quarterly devoted to generic criticism*, 8 : 248-266.

Kraif O. (2016). Le Lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de lexicologie*, 108 : 91-106.

Kraif O. (2019). Explorer la combinatoire lexico-syntaxique des mots et expressions avec le Lexicoscope. *Langue française*, 203 : 67-83.

Kraif O. & Diwersy S. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. *Actes de la conférence conjointe JAP-TALN-RECITAL 2012*, volume 2 : TALN (p. 399-406). ATALA & AFCEP.

Krulic B. (2007). *Fascination du roman historique. Intrigues, héros et femmes fatales*. Autrement.

## L

Lagorgette D. (2003). Termes d'adresse, insulte et notion de détachement en diachronie : quels critères d'analyse pour la fonction d'adresse ? *Cahiers de praxématique*, 40 : 43-70.  
<<https://doi.org/10.4000/praxematique.2696>>

Landowski E. (1988). La lettre comme acte de présence. In A. J. Greimas (éd.), *La lettre. Approches sémiotiques. Les actes du VI<sup>e</sup> colloque interdisciplinaire* (p. 19-25). Éditions Universitaires Fribourg Suisse.

Leblanc J.-M. (2016). *Analyses lexicométriques des vœux présidentiels*, ISTE Editions.

Leech G. (1991). The State of the Art in Corpus Linguistics. In K. Aijmer & B. Altenberg (éd.), *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, (p. 8-29). Longman.

Leech G. (1992). Corpora and Theories of Linguistic Performance. In J. Svartvik (éd.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*. (p. 105-122). Mouton de Gruyter.

Legallois D. & François J. (2006). Introduction. In D. Legallois & J. François (éd.), *Autour des grammaires de constructions et de patterns* (p. 1-3). *Cahier du CRISCO*, 21.

Legallois D. & Gréa P. (2006). La Grammaire de Construction. In D. Legallois & J. François (éd.), *Autour des grammaires de constructions et de patterns* (p. 5-27). *Cahier du CRISCO*, 21.

Legallois D. & Koch S. (2020). The Notion of Motif Where Disciplines Intersect: Folkloristics, Narrativity, Bioinformatics, Automatic Text Processing and Linguistics. In I. Novakova et D. Siepmann (éd.), *Phraseology and Style in Subgenres of the Novel. A Synthesis of Corpus and Literary Perspectives* (p. 17-46). Palgrave Macmillan.

Legallois D. & Tutin A. (2013). 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 : 95-117. <<https://doi.org/10.4000/lidil.3950>>

Lehmann A. & Martin-Berthet F. (1998). *Introduction à la lexicologie. Sémantique et morphologie*. Dunod.

Léon J. (2008). Aux sources de la « Corpus Linguistics » : Firth et la London School. *Langages*, 171 : 12-33.

Letourneux M. (2016). Le genre comme pratique historique. *Belphégor. Littérature populaire et culture médiatique*, 14. <<https://doi.org/10.4000/belphegor.732>>

Lignereux C. (2016). ATELIER. Les typologies des manuels d'art épistolaire. *Exercices de rhétorique*, 6. <<https://doi.org/10.4000/rhetorique.439>>

Lipovetsky G. (2006). *La société de déception. Entretien mené par B. Richard*. Éditions Textuel.

Longrée D., Luong X. & Mellet S. (2008). Les motifs : un outil pour la caractérisation topologique des textes. In *Actes JADT'2008. 9<sup>e</sup> Journées internationales d'analyse statistique des données textuelles*. <<http://lexicometrica.univ-paris3.fr/jadt/jadt2008/tocJADT2008.htm>>

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. <<https://doi.org/10.3917/lang.189.0065>>

Lundquist, L. (1980). *La cohérence textuelle : syntaxe, sémantique, pragmatique*. Nyt Nordisk Forlag Arnold Busk.

## M

- Maingueneau D. (1991). *L'analyse du discours. Introduction à la lecture de l'Archive*. Hachette.
- Maingueneau D. (2007a). Genres de discours et modes de généricité. *Le français aujourd'hui*, 159 : 29-35.
- Maingueneau D. (2007b). Modes de généricité et compétence générique. In R. Baroni & M. Macé (éd.), *Le savoir des genres* (p. 57-71). Presses Universitaires de Rennes.
- Maingueneau D. (2009). *Les termes clés de l'analyse du discours*. Seuil.
- Maingueneau D. & Philippe G. (2002). Les conditions d'exercice du discours littéraire. In E. Roulet & M. Burger (éd.), *Les modèles du discours au défi d'un « dialogue » romanesque : l'incipit du roman de R. Pinget « Le Libera »* (p. 351-377). Presses Universitaires de Nancy.
- Mallory J.-P. (1997 [1989]). *À la recherche des Indo-Européens. Langue, archéologie, mythe*, trad. fr. J.-L. Giribone. Seuil.
- Malrieu D. & Rastier F. (2001). Genres et variations morphosyntaxiques. *Traitement automatique des langues*, 42.2 : 547-577.
- Marchello-Nizia C. (1995). *L'évolution du français. Ordre des mots, démonstratifs, accent tonique*. Armand Colin.
- Marchello-Nizia C., Combettes B., Prévost S. & Scheer T. (dir.) (2020). *Grande Grammaire Historique du Français*. De Gruyter.
- Mayaffre D. (2002). Les corpus réflexifs : entre architextualité et hypertextualité. *Corpus*, 1. <<https://doi.org/10.4000/corpus.11>>
- Meillet A. (1906). Comment les mots changent de sens. *L'Année sociologique 1904-1905* : 6-43.
- Meillet A. (1923). *Les origines indo-européennes des mètres grecs*. Presses universitaires de France. <<https://archive.org/details/lesoriginesindoe00meiluoft/page/60/mode/2up>>
- Mejri S. (1997). *Le figement lexical. Descriptions linguistiques et structuration sémantique*. Publications de la Faculté des Lettres de la Manouba.
- Mel'čuk I. (1993). La phraséologie et son rôle dans l'enseignement/apprentissage d'une langue étrangère. *Études de Linguistique Appliquée*, 92 : 82-113.
- Mel'čuk I. (2011). Phrasèmes dans le dictionnaire. In J.-C. Anscombre & S. Mejri (éd.), *Le figement linguistique : la parole entravée* (p. 41-61). Honoré Champion.
- Mel'čuk I. (2013). Tout ce que nous voulons savoir sur les phrasèmes, mais... *Cahiers de lexicologie*, 102 : 129-149.

Micheli R. (2010). *L'émotion argumentée. L'abolition de la peine de mort dans le débat parlementaire français*. Cerf.

Micheli R., Hekmat I. & Rabatel A. (2013). Les émotions : des modes de sémiotisation aux fonctions argumentatives. *Semen*, 35. <<https://doi.org/10.4000/semen.9790>>.

Mogorrón Huerta P. (2011). Les expressions figées le sont-elles vraiment ? In J.-C. Anscombre & S. Mejri (éd.), *Le figement linguistique : la parole entravée* (p. 217-234). Honoré Champion.

Mortureux M.-F. (2008). *La lexicologie entre langue et discours* (2<sup>e</sup> éd.). Armand Colin.

## N

Née É., Sitri F. & Veniard M. (2014). Pour une approche des routines discursives dans les écrits professionnels. *4<sup>e</sup> Congrès Mondial de Linguistique Française*, vol. 8, SHS Web of Conferences (p. 2113-2124). EDP Sciences. <<https://doi.org/10.1051/shsconf/20140801195>>

Nelson M. (2010). Building a written corpus. In A. O'Keeffe & M. Mc Carthy (éd.), *The Routledge Handbook of Corpus Linguistics* (p. 53-65). Routledge.

Noailly M. (2005). Mon cher trésor. *Cher* et son rôle dans l'interpellation, tant à l'oral qu'à l'écrit, *Modèles linguistiques*, XXVI-2 (52) : 33-44.

Nølke H. (2002). Pour un traitement modulaire de la syntaxe transphrastique. *Verbum*, 24, 1-2 : 179-192.

Nourrisson D. (1988). Tabagisme et antitabagisme en France au XIX<sup>e</sup> siècle. *Histoire, économie et société*, 4 : 535-547.

Novakova I. (2010). *Syntaxe et sémantique des prédicats (approche contrastive et fonctionnelle). Volume 1 : Synthèse des travaux de recherche*. Dossier présenté en vue de l'Habilitation à Diriger des Recherches. Université Stendhal Grenoble 3.

Novakova I. & Gymnich M. (2021). Extended phraseological units and literary genres: A contrastive analysis of French and English lexico-syntactic constructions. *Lexique*, 28 : 87-112.

Novakova I., Grossmann F. & Goossens V. (2013). Interactions entre profil discursif et structure actancielle : l'exemple des noms de surprise et de respect. In F. Baider & G. Cislaru (éd.), *Cartographie des émotions. Propositions linguistiques et sociolinguistiques* (p. 71-83). Presses universitaires de la Sorbonne Nouvelle.

Novakova I. & Tutin A. (dir.). (2009). *Le lexique des émotions*. ELLUG.

Novakova I. & Siepmann D. (dir.) (2020). *Phraseology and Style in Subgenres of the Novel. A Synthesis of Corpus and Literary Perspectives*. Palgrave Macmillan.

## P

Pamies Bertrán A. (2018). Aux limites du limitrophe : à propos des catégories phraséologiques. In I. Sfar & P.-A. Buvet (éd.), *La phraséologie entre fixité et congruence. Hommage à Salah Mejri* (p. 221-263). Éditions Academia – L'Harmattan.

Parry M. (1928a). *Les formules et la métrique d'Homère*. Les Belles Lettres.

Parry M. (1928b). *L'épithète traditionnelle dans Homère. Essai sur un problème de style homérique*. Thèse pour le doctorat ès lettres présentée à la faculté des Lettres de l'université de Paris. Les Belles Lettres.

Pecman M. (2004). L'enjeu de la classification en phraséologie. In *Actes du congrès EUROPHRAS 2004*. Société européenne de phraséologie, Université de Bâle, 26-29 août 2004 (p.127-146). Schneider Hohengehren Verlag.

Perpillou J.-L. (1973). *Les substantifs grecs en -εύς*. Klincksieck.

Pinault G.-J. (1995-1996). Distribution des particules comparatives dans la *Ṛk-saṃhitā*. *Bulletin d'Études Indiennes*, 13-14 : 307-367.

Pincemin B. (2006). Concordances et concordanciers : de l'art du bon KWAC. In C. Duteil-Mougel & B. Foulquié (éd.), *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation* (p. 33-42). Texto.

Plantin C. (1997). L'argumentation dans l'émotion. *Pratiques*, 96 : 81-100.

Plantin C. (2011). *Les bonnes raisons des émotions. Principes et méthodes pour l'étude du discours émotionné*. Peter Lang.

Plantin C. (2012). Les séquences discursives émotionnées : définition et application à des données tirées de la base CLAPI. *3<sup>e</sup> Congrès Mondial de Linguistique Française*. SHS Web of Conferences, vol. 1 (p. 629-642). EDP Sciences.

Polguère A. (2008). *Lexicologie et sémantique lexicale. Notions fondamentales* (2<sup>e</sup> éd.). Les Presses de l'Université de Montréal.

Poiron D. & Walter P. (dir.) (2001). *Le Livre du Graal, I. Joseph d'Armathie. Merlin. Les Premiers Faits du roi Arthur*. Gallimard.

Poudat C. (2006). *Étude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*. Thèse de doctorat, Université d'Orléans.

Poudat C. (2020). *Explorer un corpus textuel dans une perspective d'analyse contrastive des genres et des discours. Des observables à la méthode, du texte aux discussions de la CMR*. Mémoire de synthèse, volume 1, présenté pour l'Habilitation à diriger des recherches. Université Côte d'Azur.

Poudat C. & Landragin F. (2017). *Explorer un corpus textuel. Méthodes – pratiques – outils*. De Boeck Supérieur.

Prévost S. (2015). Diachronie du français et linguistique de corpus : une approche quantitative renouvelée. *Langages*, 197 : 23-45.

## R

Rabatel A. (2010). Texte, dynamique interprétative et communauté discursive. In L.-S. Florea, C. Papahagi, L. Pop & A. Curea (éd.). *Directions actuelles en linguistique du texte. Actes du Colloque International Le Texte : modèles, méthodes, perspectives* (p. 177-188). Casa Cărții de Știință.

Rastier F. (1987). *Sémantique interprétative*. Presses Universitaires de France.

Rastier F. (1989). *Sens et textualité*. Hachette.

Rastier F. (2003). Formes sémantiques et textualité. *Cahiers du CRISCO*, 12 : 99-114.

Rastier F. (2005a). Enjeux épistémologiques de la linguistique de corpus. In G. Williams (éd.). *La linguistique de corpus* (p. 31-45). Presses Universitaires de Rennes.

Rastier F. (2005b). La microsémantique. *Texte ! Textes & Cultures*. <<http://www.revue-texto.net/index.php?id=536>>

Rastier F. (2006). Formes sémantiques et textualité. *Langages*, 163 : 99-114.

Rastier F. (2011). *La mesure et le grain. Sémantique de corpus*. Honoré Champion.

Renou L. (1951). *L'hindouisme*. Presses Universitaires de France.

Renou L. (1956). *Hymnes spéculatifs du Véda*. Gallimard / Unesco.

Ricoeur P. (1986). *Du texte à l'action*. Seuil.

Ryan M.-L. (1981). Introduction. On the why, what and how of generic taxonomy. *Poetics* 10 : 109-126.

Rychner J. (1955). *La chanson de geste. Essai sur l'art épique des jongleurs*. Droz.

## S

Schaeffer J.-M. (1989). *Qu'est-ce qu'un genre littéraire ?* Seuil.

Schaeffer J.-M. (1995). Genres littéraires. In O. Ducrot & J.-M. Schaeffer (éd.), *Nouveau dictionnaire encyclopédique des sciences du langage* (p. 626-637). Seuil.

Schaeffer J.-M. (2007). Des genres discursifs aux genres littéraires : quelles catégorisations pour quels faits textuels ? In R. Baroni & M. Macé (éd.), *Le savoir des genres* (p. 357-364). Presses Universitaires de Rennes.

Scholes R. (1986 [1977]). Les modes de la fiction. In G. Genette et al., *Théorie des genres* (p. 77-88). Seuil.

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.

Siepmann D. (2016). Lexicologie et phraséologie du roman contemporain : quelques pistes pour le français et l'anglais. *Cahiers de Lexicologie*, 108 : 21-41.

Sinclair J. (1965). When is a Poem like a Sunset? *A Review of English Literature*, 6.2 : 76-91.

Sinclair J. (1991). *Corpus Concordance Collocation*. Oxford University Press.

Sinclair J. (2008). Preface. In S. Granger & F. Meunier (éd.), *Phraseology: an interdisciplinary perspective* (p. XV-VXIII). John Benjamins.

Sonntag I. (2018). L'« image » des Turcs dans les débats des temps modernes sur la tolérance. *Annales de Bretagne et des pays de l'ouest*, 125.1 : 83-95.  
<<https://doi.org/10.4000/abpo.3775>>

Sorlin S. (2015). Le genre : ligne de force de la théorie et de la pratique littéraire ? In D. Ablali, A. Bouhouhou & O. Tebbaa (éd.), *Les genres textuels, une question d'interprétation ?* (p. 103-112). Lambert-Lucas.

Soutet O., Sfar I. & Mejri S. (dir). (2018). *Phraséologie et discours*. Honoré Champion.

Souvay G. & Pierrel J.-M. (2009). LGeRM Lemmatisation des mots en Moyen Français. *Traitement automatique des langues, ATALA*, 50.2.

Stefanowitsch A. & Gries S. T. (2003). Collocations: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8.2 : 209-43

Stubbs M. (2014). Patterns of Emotive Lexis and Discourse Organization in *Short Stories* by James Joyce. In P. Blumenthal, I. Novakova & D. Siepmann (éd.), *Les émotions dans le discours / Emotions in Discourse* (p. 237-253). Peter Lang.

Stubbs M. & Barth I. (2003). Using Recurrent Phrases as Text-type Discriminators: A Quantitative Method and some Findings. *Functions of Language*, 10 : 65-108.

## T

Tersis N. & Plantin C. (2017). Les émotions dans la langue et les chants inuit. In N. Tersis & P. Boyeldieu (éd.), *Le langage de l'émotion. Variations linguistiques et culturelles* (p. 93-122). Peeters.

Todorov T. (1970). *Introduction à la littérature fantastique*. Seuil.

Tognini-Bonelli E. (2001). *Corpus Linguistics at Work*. John Benjamins.

Tribble C. (2010). What are concordances and how are they used? In A. O'Keefe & M. Mc Carthy (éd.), *The Routledge Handbook of Corpus Linguistics* (p. 167-183). Routledge.

Tutin A. (2010). *Sens et combinatoire lexicale : de la langue au discours*. Dossier en vue de l'habilitation à diriger des recherches. Volume 1 : Synthèse. Université Stendhal – Grenoble 3.

Tutin A. (2013). Les collocations lexicales : une relation essentiellement binaire définie par la relation prédicat-argument. *Langages*, 189 : 47-63.

Tutin A. (2014). La phraséologie transdisciplinaire des écrits scientifiques : des collocations aux routines sémantico-rhétoriques. In F. Grossmann & A. Tutin (éd.), *L'écrit scientifique : du lexique au discours* (p. 27-44). Presses Universitaires de Rennes.

Tutin A. & Kraif O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents. *LIDIL*, 53 : 119-141. <<https://doi.org/10.4000/lidil.3966>>

Tutin A., Novakova I., Grossmann F. & Cavalla C. (2006). Esquisse de typologie des noms d'affect à partir de leurs propriétés combinatoires. *Langue française*, 150 : 32-49.

## V

Van Valin R. D. & LaPolla R.J. (1997). *Syntax. Structure, meaning and function*. Cambridge University Press.

Vaugeois D. (2007). Un genre est un genre parce que c'est un genre. Réflexion sur les critères minimaux du genre littéraire. In R. Baroni & M. Macé (éd.), *Le savoir des genres* (p. 211-225). Presses Universitaires de Rennes.

Veniard M. (2013). *La nomination des événements dans la presse : essai de sémantique discursive*. Presses universitaires de Franche-Comté.

Viprey J.-M. (2006). Structure non séquentielle des textes. *Langages*, 163 : 71-85.

## W

Walter E. (2010). Using corpora to write dictionaries. In A. O'Keeffe & M. Mc Carthy (éd.), *The Routledge Handbook of Corpus Linguistics* (p. 428-443). Routledge.

Weinrich H. (1973 [1964]). *Le temps*, trad. fr. M. Lacoste. Éditions du Seuil.

Weinrich H. (1989). *Grammaire textuelle du français*. Didier/Hatier.

Werlich E. (1975). *Typologie der Texte, Entwurf eines Textlinguistischen Modells zur Grundlegung einer Textgrammatik*. Quelle und Meyer.

Williams G. (dir.) (2005). *La linguistique de corpus*. Presses Universitaires de Rennes.

Wittgenstein L. (1958 [1953]). *Philosophical Investigations*, tr. angl. G.E.M. Anscombe. Basil Blackwell.

Wodtko D.S., Irslinger B. & Schneider C. (2008). *Nomina im Indogermanischen Lexikon*. Universitätsverlag Winter.

## Z

Zaharieva R. & Kaldieva-Zaharieva S. (2017). Des principaux termes dans le domaine de la phraséologie. In F. Grossmann, S. Mejri & I. Sfar (éd.), *La phraséologie : sémantique, syntaxe, discours* (p. 15-37). Honoré Champion.

### III. Références sitographiques

- CTH - CATALOGUE OF HITTITE TEXTS. E. Laroche, continued by S. Košak and G.G.W. Müller with the collaboration of S. Görke and Ch. Steitler <[https://www.hethport.uni-wuerzburg.de/CTH/index\\_en.php](https://www.hethport.uni-wuerzburg.de/CTH/index_en.php)>
- EMAN – Édition de Manuscrits et d’Archives Numériques <<https://eman-archives.org/EMAN/projetseman>>
- EMOLEX. *Le lexique des émotions dans cinq langues européennes*. <<http://phraseotext.u-grenoble3.fr/emolex/spip.php?article1>>
- Frantext <<https://www.frantext.fr/>>
- Gallica <<https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop>>
- GENESIS Manuscrits - Recherche – Invention <<https://journals.openedition.org/genesis/>>
- HumaNum La TGIR des humanités numériques <<https://www.huma-num.fr/>>
- Le Petit Robert de la langue française*. Dictionnaire Le Robert. Version numérique 5.4 (2021).
- Lexicoscope <<http://phraseotext.univ-grenoble-alpes.fr/lexicoscope/>>
- Lexicoscope 2.0 <[http://phraseotext.univ-grenoble-alpes.fr/lexicoscope\\_2.0/](http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/)>
- LGeRM Plate-forme de lemmatisation <<http://stella.atilf.fr/LGeRM/>>
- LiDiLEM Laboratoire de linguistique et didactique des langues étrangères et maternelles <<https://lidilem.univ-grenoble-alpes.fr/>>
- LIG Laboratoire d’Informatique de Grenoble <<https://www.liglab.fr/>>
- Litt&Arts UMR 5316 <<https://litt-arts.univ-grenoble-alpes.fr/>>
- MIAI Multidisciplinary Institute in Artificial Intelligence <<https://miai.univ-grenoble-alpes.fr/>>
- Orthodidacte <<https://www.orthodidacte.com/>>
- Perseus Digital Library <<http://www.perseus.tufts.edu/hopper/>>
- PhraseoRom <<https://phraseorom.univ-grenoble-alpes.fr/fr>>
- PhraseoBase <<http://phraseotext.univ-grenoble-alpes.fr/phraseobase/index.html>>
- Projet Ecritures : <<http://www.univ-paris3.fr/corpus-du-projet-98190.kjsp?RH=1295620465180>>
- Corpus Sermo : <<https://www.lextechinstitute.ch/presentation-du-corpus-sermo-quest-ce-quun-corpus-de-sermons-protestants-du-xvi-xviii-siecle-peut-nous-apprendre-sur-la-langue/>>
- SRCMF : Prévost S. & Stein A. (2013). *Syntactic Reference Corpus of Medieval French* (SRCMF) [version number]. ENS de Lyon / ILR Stuttgart. <<http://www.srcmf.org/>>
- TACT <<https://tact.demarre-shs.fr/>>
- TEI Text Encoding Initiative <<https://tei-c.org/>>
- Thesaurus Linguae Graecae: a Digital Library of Greek Literature, University of California <<http://stephanus.tlg.uci.edu/>>
- UGA Éditions <<https://www.uga-editions.com/>>
- Wikisource <<https://fr.wikisource.org/wiki/Wikisource:Accueil>>

## IV. Curriculum vitae

Julie Sorba

Maitresse de conférences hors classe, Université Grenoble Alpes, laboratoire LIDILEM.

[\[https://lidilem.univ-grenoble-alpes.fr/membre/julie-sorba\]](https://lidilem.univ-grenoble-alpes.fr/membre/julie-sorba)

Le *curriculum vitae* détaillé a été supprimé de la version mise en ligne sur HAL. Néanmoins, la pagination du mémoire original a été maintenue.

*Page originale supprimée*

## **V. Une expérience de la valorisation en SHS : le projet Certiredac et son prolongement dans écrit +**

Je propose de revenir ici sur une partie de mon activité scientifique qui a pris naissance à partir d'une réflexion autour d'une de mes pratiques pédagogiques. Le « plan licence », lancé en 2009, nous a permis de mettre en place une formation de remédiation en langue française, baptisée « Orthographe & Grammaire », à destination des étudiants de la licence Sciences du Langage. L'originalité de notre démarche reposait alors sur la répartition des étudiants volontaires en groupes de niveau en fonction du résultat que ceux-ci avaient obtenu au test de positionnement en langue française que nous faisons passer à tous les étudiants de L1 lors de la matinée d'accueil. Ce test de positionnement testait les compétences grammaticales basiques (nature et fonctions des mots, types de proposition, conjugaison, orthographe grammaticale, etc.).

J'ai alors commencé à me documenter sur les solutions numériques disponibles sur le marché. Une entreprise de l'agglomération grenobloise, Zeugmo, venait tout juste de lancer sa plateforme Orthodidacte qui affichait comme ambition d'améliorer la maîtrise du français écrit. Pendant l'année 2012-2013, j'ai ainsi mis en place un entraînement suivi sur cette plateforme pour mes étudiants en parallèle de cours plus traditionnels et de l'utilisation de la plateforme institutionnelle Moodle pour des exercices d'auto-évaluation. Parallèlement, lors d'une formation assurée par ma collègue linguiste didacticienne Françoise Boch (Université Grenoble Alpes, LiDiLEM), j'ai découvert, avec enthousiasme, la « grammaire en couleurs »<sup>142</sup> que j'ai immédiatement mise en pratique dans mes cours de remédiation. Les bases de notre collaboration venaient d'être posées.

À l'automne 2013, l'équipe d'Orthodidacte nous contactait toutes les deux pour nous proposer de créer une certification en langue française afin de se positionner sur le même segment de marché que le projet Voltaire alors en pleine expansion<sup>143</sup>. Nous avons soumis un dossier de candidature à l'appel à projet de l'ANR dans le cadre de l'instrument de financement « partenariat public-privé » mais celui-ci n'a pas été retenu. Notre expertise pédagogique et didactique nous avait amenées à constater que les (rares) tests en langue française sur le marché (autres que les certifications FLE) étaient insatisfaisants car ils ne traitaient que de

---

<sup>142</sup> La grammaire en couleurs est une méthode fondée sur une approche pédagogique inductive et conçue par C. et M. Laurent, à l'origine pour un public d'enseignement secondaire. Voir <<https://lagrammaireencouleurs.com/>>.

<sup>143</sup> Voir <<https://www.projet-voltaire.fr/>>.

l'orthographe et n'utilisaient que des questions fermées. En effet, les compétences rédactionnelles ne se limitant pas à une maîtrise de l'orthographe, pour évaluer celles-ci, il convient d'évaluer (1) toutes les composantes relevant de la compétence scripturale (orthographe, syntaxe, ponctuation, lexique, cohérence textuelle) et (2) la capacité d'un candidat à rédiger, c'est-à-dire à mettre en œuvre toutes ces composantes, et pas seulement à cocher des cases. Après avoir décliné l'offre de prestation intellectuelle de l'entreprise Zeugmo, nous avons décidé d'explorer les (maigres) possibilités institutionnelles qui s'offraient alors à nous.

Grâce au soutien de l'équipe présidentielle de l'Université Stendhal – Grenoble 3, nous avons pu modéliser un premier prototype de certificat (version papier) avec l'aide d'une chercheuse postdoctorante, Tiphaine Mout. En octobre 2014, nous disposions de suffisamment d'éléments pour présenter CertiRedac, notre projet de « technologie innovante » devant le comité d'investissement de Gate 1<sup>144</sup>, pour une candidature à l'entrée en phase de maturation<sup>145</sup> au sein de la structure qui était en train de devenir la SATT Linksium. L'originalité du projet était de proposer un outil pour valoriser les compétences rédactionnelles dans le monde professionnel :

Afin de satisfaire les besoins des différents acteurs professionnels, nous avons tenté d'articuler le point de vue du candidat avec celui du recruteur. Par conséquent, ce certificat poursuit deux objectifs :

- fournir à tout adulte francophone une évaluation fine de son niveau à l'écrit pour l'aider, d'une part, à se positionner et à se former au besoin et, d'autre part, à valoriser ses compétences en milieu professionnel ;
- fournir au recruteur et à l'employeur les moyens d'identifier les compétences à l'écrit de son personnel (actuel ou futur). (Boch, Sorba & Bessonneau 2016, p. 128)

---

<sup>144</sup> La structure Gate 1 (Grenoble Alps Technology and Entrepreneurship) était précédemment née de la fusion des équipes grenobloises de Gravit et de Grain (incubateurs d'entreprise) et de PETALE (accompagnateur de croissance d'entreprise). Voir <<https://www.gate1.fr/>>. L'écosystème français du transfert de technologie vers la société civile était alors en pleine restructuration avec la création des SATT (Société d'Accélération du Transfert de Technologies) dans le cadre du PIA2 (Programme Investissement d'Avenir). Voir <<https://www.satt.fr/>>.

<sup>145</sup> Dans le cadre d'une opération de valorisation de la recherche publique, la phase de maturation s'appuie sur une « preuve de concept » issue de la recherche d'un laboratoire public. Elle propose d'accompagner les chercheurs dans la sécurisation de la propriété intellectuelle de leur invention et dans la réalisation d'un « démonstrateur applicatif répondant aux attentes du marché » (voir <<https://www.linksium.fr/profil-chercheur>>).

Une fois lauréat, notre projet a pu bénéficier, sur l'année 2015, d'un accompagnement financier (103 k€) et humain (notre chargée d'affaires, Mme Gisela Schach) pour faire la « preuve du concept ».

Côté laboratoire, nous avons finalisé le prototype de certificat après l'avoir testé sur des cohortes de 366 étudiants volontaires (Licence Biologie, Université Grenoble 1 Joseph Fourier ; Classes Préparatoires scientifiques, Lycée Vaugelas, Chambéry ; Licences Sciences du Langage, Lettres, Langues, Université Stendhal Grenoble 3). L'appui d'un statisticien, Pascal Bessonneau, nous a permis de valider nos choix par le recours à des tests psychométriques et d'en ajuster certains. Le *Certificat en Compétences Rédactionnelles* (ou CCR<sup>®</sup>) venait de naître (pour le détail, voir Boch, Sorba & Bessonneau 2016). Dans la foulée, nous présentions ces premiers résultats lors d'une journée d'étude à Marseille (Boch, Sorba & Bessonneau 2015)

Côté valorisation, nous avons travaillé, avec notre chargée d'affaires, pour élaborer la stratégie de positionnement la plus pertinente de notre produit à travers les actions suivantes :

- Réalisation d'une étude de marché auprès de professionnels (une centaine d'entretiens) ;
- Création d'une identité visuelle (baseline et logos, figure 14) avec une graphiste (entreprise Kangooroad) ;
- Dépôts des marques et de la propriété intellectuelle (déclaration d'invention et 7 enveloppes Soleau<sup>146</sup>) avec l'aide du service juridique de l'Université Stendhal – Grenoble 3 (Mme Sophie Beaubron) ;
- Création d'un teaser (entreprise Minimento), disponible en ligne : <https://www.youtube.com/watch?v=TFxxU3XhE44> ;
- Élaboration du cahier des charges de la plateforme pour créer la version numérique du certificat et choix du prestataire (entreprise Kreatys) ;
- Conférence pour présenter notre projet (Boch & Sorba 2014).



Figure 14 : Deux exemples de l'identité graphique projet CertiRedac et du CCR<sup>®</sup>

---

<sup>146</sup> L'enveloppe Soleau est un moyen de preuve de création auprès de l'INPI. Pour plus de détail, voir <https://www.inpi.fr/fr/enveloppe-soleau>.

Cette intense activité de valorisation dans le domaine des Sciences humaines et sociales, une première sur le site grenoblois, a eu des retombées inattendues. Lors de l'étude de marché, l'une des entreprises démarchées, McDonald's France, a manifesté son intérêt pour une action de formation certifiante à l'égard de son personnel encadrant. L'Université Grenoble Alpes a alors conclu, via son service commun de Formation Continue, un contrat de deux ans et nous avons pu mettre en place et assurer, à Paris, deux sessions de formation expérimentale à l'automne 2016 et à l'automne 2017 auprès de deux groupes de stagiaires (managers et directeurs de restaurant). Parallèlement, les maisons d'édition Nathan et Hatier nous ont contactées car elles souhaitaient se positionner sur le segment de la certification en langue française comme concurrentes au projet Voltaire. Même si nous avons décliné l'offre de prestation intellectuelle proposée, ces deux manifestations d'intérêt nous ont permis alors de conforter notre projet de développer notre certificat.

À la fin de la phase de maturation, l'étape suivante s'est donc présentée assez logiquement : construire un dossier de candidature pour la phase d'incubation auprès de la SATT Linksium<sup>147</sup>. La stratégie de valorisation imposée par la SATT était alors d'accorder une licence d'exploitation à une entreprise existante ou à créer. La demande d'incubation devait être portée par l'entreprise intéressée. Début 2016, nous avons donc publié une offre de recherche d'un porteur ou d'une porteuse de projet puis organisé deux auditions (4 candidats entrepreneurs à la première session dont 3 retenus pour la seconde). La candidate retenue, Mme Mélanie Viénot, a rejoint l'équipe et nous l'avons donc préparée pour la présentation du dossier de candidature à la phase d'incubation. Lauréat de ce second financement (116 k€), notre projet a pu poursuivre sa route au sein de la SATT Linksium et bénéficier ainsi de l'accompagnement d'une nouvelle chargée d'affaire (Mme Catherine Cayuela). En juillet, l'équipe s'agrandissait d'une co-porteuse, Mme Virginie Kauffman, docteure de l'Université Stendhal – Grenoble 3. La réflexion sur la création d'une start-up est lancée : elle s'appellera Novascrib. Nous aidons activement en tant que conseillères scientifiques les deux co-porteuses dans l'élaboration de leur stratégie et dans la conception de la plateforme de passation du test<sup>148</sup>. La filiale de valorisation de l'université, Floralis, nous apporte son soutien. Durant l'année 2016, nous réalisons quatre conférences sur le site grenoblois pour présenter notre exemple de valorisation

---

<sup>147</sup> La phase d'incubation a pour objectif d'amener la technologie développée sur le marché. Les actions proposées sont variées : formations pour devenir entrepreneur, aide à la constitution d'équipe, recherche de financements, sourcing de prestataires, etc. Voir <<https://www.linksium.fr/creer-une-startup-technologique>>.

<sup>148</sup> Voir la description du projet sur le site du prestataire : <<http://blog.kreatys.com/etude-de-cas-novascrib/>>.

de la recherche en Sciences Humaines et Sociales. Côté recherche, nous présentons une communication à Bordeaux sur l'une des spécificités de notre certificat qui demande aux candidats d'indiquer leur degré de lucidité<sup>149</sup> et son intérêt dans la formation (Boch, Sorba & Bessonneau 2016b). Il s'agit d'envisager le développement dans la durée des compétences rédactionnelles et de fournir aux étudiants et aux formateurs des leviers utilisables dans la formation :

Inclure dans le CCR la mesure du degré de lucidité permet ainsi de proposer un test complet qui n'évalue pas seulement les compétences métacognitives mises en œuvre dans les réponses aux exercices, mais aussi le savoir métacognitif, ce « savoir déclaratif d'une personne, au sujet d'interactions entre une personne, la tâche et les caractéristiques de la stratégie » (Veenman, Van Hout-Wolters & Afflerbach, 2006 : 4). En formulant un jugement sur sa propre réponse, via le degré de lucidité, nous faisons l'hypothèse que le candidat entre dans un processus d'autorégulation : « la métacognition peut aboutir à un jugement (non habituellement exprimé) sur la qualité des activités mentales en question ou de leur produit et éventuellement à une décision de modifier l'activité cognitive, son produit ou même la situation qui l'a suscitée. » (Noël, 1997 : 19). (Boch, Sorba & Bessonneau 2019, p. 83-84)

Malheureusement, les acteurs institutionnels (UGA, Floralis, Linksium) peinent à se mettre rapidement d'accord sur le plan juridique pour contractualiser leurs relations entre eux, avec nous et avec les co-porteuses. Pendant ce temps, nous continuons à transférer notre savoir-faire auprès des co-porteuses qui nous tiennent de moins en moins informées de l'avancée du projet et nous mettent progressivement à l'écart de leur réflexion. Nous alertons nos tutelles (UGA, Floralis, SATT Linksium) de cette dégradation des relations au sein de l'équipe de la phase d'incubation mais rien ne change. En avril 2017, la start-up est créée mais la confiance s'est effritée. Nous participons néanmoins au 3<sup>e</sup> salon *Innovatives SHS* organisé par le CNRS à Marseille (Boch & Sorba 2017a), à une conférence invitée à l'université de Toulon (Boch & Sorba 2017b) et présentons des résultats de recherche en didactique du français écrit à deux sessions du 85<sup>e</sup> congrès l'ACFAS à Montréal (Boch, Sorba & Bessonneau 2017a et b). Jusqu'à la fin 2017, nous nous rendons à de nombreuses réunions infructueuses avec les différentes tutelles. Nous nous retirons alors définitivement du projet de participer à l'entreprise Novascrib

---

<sup>149</sup> « Le “degré de lucidité” est un indicateur du niveau de conscience de ses compétences. Est considéré comme lucide quelqu'un qui sait qu'il sait ou qui sait qu'il ne sait pas. À l'inverse, n'est pas considéré comme lucide quelqu'un qui ne sait pas qu'il sait (il sous-évalue ses compétences) ou qui ne sait pas qu'il ne sait pas (il surévalue ses compétences). Afin de calculer ce degré de lucidité, nous proposons dans le CCR<sup>®</sup> deux choix formulés au moyen d'une consigne. » (Boch, Sorba & Bessonneau 2019, p. 84).

qui commercialise désormais, sous licence exclusive, notre savoir-faire qui est actuellement mis en œuvre dans le certificat Bescherelle<sup>150</sup>.

Après d'âpres négociations, nous avons tout de même réussi à obtenir de conserver la possibilité d'une utilisation (non commerciale) pédagogique et scientifique de notre savoir-faire transféré. Sur le plan pédagogique, nous avons donc pu répondre favorablement à la demande de l'UFR LLSH de l'Université Savoie Mont Blanc d'utiliser notre certificat comme test de positionnement pour ses premières années de Licences Histoire et Sociologie à la rentrée 2017 dans le projet de créer des cours de « Compétences rédactionnelles »<sup>151</sup>. Le partenariat institutionnel ainsi établi nous a permis de recueillir de nouvelles données en 2017 et en 2018 que nous avons exploitées dans une étude originale sur le choix des distracteurs<sup>152</sup> dans l'évaluation de la compétence lexicale des scripteurs adultes en L1 (communication Boch & Sorba 2019 parue dans Boch & Sorba 2020). Notre objectif était d'identifier les caractéristiques des bons distracteurs dans un test QCM évaluant la compétence lexicale L1 en réception, à partir des résultats obtenus par les étudiants à deux exercices de notre test *CCR*<sup>®</sup> :

Nos observations suggèrent que les distracteurs influencent les choix des candidats de manière complexe : si les résultats obtenus convergent avec la littérature sur la question de l'efficacité des distracteurs reliés sémantiquement à la lexie (que ce soit par une relation de synonymie ou d'antonymie ou des emplois similaires entre lexie cible et distracteur), ils mettent également en lumière l'efficacité de distracteurs non reliés sémantiquement : la paronymie en particulier semble exercer une influence notable. (Boch & Sorba 2020, § 34)

En 2017 également, sur le plan scientifique, nous avons été sollicitées pour participer au dépôt de projet écri+ « évaluer, former, certifier en français », en compagnie d'une quinzaine d'établissements d'enseignement supérieur français et porté par l'Université Ouverte de

---

<sup>150</sup> Voir <<https://moncoach.bescherelle.com/la-certification>>.

<sup>151</sup> Ce partenariat a été porté initialement par le directeur du département d'Histoire (Laurent Guichard). Il a contribué à créer la dynamique qui a conduit l'Université Savoie Mont Blanc à rejoindre tout récemment le projet écri+ (avec pour référente Rafaële Fouillet, voir *infra*).

<sup>152</sup> « Dans un test comportant des questions à choix multiple (dorénavant QCM), chaque question est composée d'une réponse correcte et d'autres réponses incorrectes ou "distracteurs" (Revuelta, 2004, p. 217). Ceux-ci se définissent comme des réponses plausibles mais incontestablement fausses, et, selon l'UTC (2015), destinées à identifier si la personne testée est capable de faire la différence entre la bonne et la mauvaise réponse. Un "distracteur" est ainsi nommé car il opère une "distraction" au sens étymologique de "détournement" : "*The incorrect options are called distractors because they are considered to be 'distracting' to students with partial knowledge due to their plausibility to yield the correct option.*" (Gierl, Bulut, Guo & Zhang, 2017, p. 1084). » (Boch & Sorba 2020, § 4).

Humanités (Strasbourg), dans le cadre du PIA3 « Nouveaux Cursus Universitaires ». Notre expertise était sollicitée sur le volet « création d'une certification en langue française » mais nous avons découvert, en consultant les services juridiques de l'université, que le précédent transfert de notre savoir faire portant sur ce même domaine nous empêchait de contribuer à l'élaboration d'un produit (une certification en langue française) qui se positionnait potentiellement, de manière concurrente, sur le même segment de marché (même si l'intention de commercialisation n'était pas affichée). Nous avons alors choisi de contribuer, d'une autre manière, au projet *écri+* en proposant la création d'une plateforme recensant les dysfonctionnements relevés dans les écrits des étudiants dans l'objectif de la mettre à la disposition des enseignants et des formateurs pour leur fournir des extraits authentiques sur lesquels travailler dans leurs classes. Le projet *écri+* a été lauréat de cet appel<sup>153</sup> et l'équipe grenobloise s'est constituée, début 2018, pour réaliser son livrable.

La première phase a été la collecte des dysfonctionnements auprès de formateurs et d'enseignants :

[...] à la différence de la démarche de constitution de corpus de textes intégraux, notre protocole demande à des enseignants volontaires de l'enseignement supérieur francophone de saisir des extraits comportant un dysfonctionnement linguistique. Ces extraits d'au moins une phrase sont déposés sur un formulaire en ligne, pourvus de métadonnées et accompagnés d'une identification du dysfonctionnement repéré par l'attribution d'une étiquette prédéfinie et éventuellement par un commentaire libre. (Boch, Rinck & Sorba 2020, p. 6)

Ainsi, notre collecte visait un double objectif : d'une part, recueillir des extraits avec des dysfonctionnements pour alimenter la plateforme ; d'autre part, constituer une base de données originales concernant les « attentes des enseignants » et « leur seuil d'acceptabilité quand ils rencontrent un dysfonctionnement linguistique » (*ibid.*). La collecte de ces données nous a permis de réfléchir sur la typologie des dysfonctionnements à adopter dans le cadre d'une utilisation pédagogique (Boch, Rinck & Sorba 2020 ; 2021) mais aussi de nous interroger, dans une perspective didactique, sur la manière de former les enseignants du supérieur à la révision des textes des étudiants, à partir de leurs commentaires qu'ils ont déposés en même temps que les extraits contenant les dysfonctionnements (Bock, Rinck & Sorba 2022 soumis).

En parallèle, depuis l'automne 2020, nous menons la deuxième phase, celle de la réalisation opérationnelle de la plateforme numérique. Nous avons soumis notre cahier des charges à une promotion d'étudiants du Master Sciences du Langage Parcours Industrie de la Langue

---

<sup>153</sup> Voir <<https://ecriplus.fr/>>.

(Université Grenoble Alpes) dans le cadre d'un cours de programmation web supervisé par Claude Ponton. Les étudiants nous ont présenté, au printemps 2021, plusieurs prototypes, fort aboutis pour certains, à partir desquels un ingénieur, en cours de recrutement, réalisera l'interface définitive.

Je tenais à retracer, en annexe, mon expérience inédite dans le domaine de la valorisation des sciences humaines et dans le champ des littéracies avancées car j'y consacre une partie de mon activité professionnelle depuis 2013. Celle-ci m'a, tout d'abord, permis de mieux comprendre les enjeux et les processus de l'activité de transfert de technologie, qui prend de plus en plus de place dans le champ des STS (Sciences Technologies Santé), champ scientifique majeur de mon établissement. Je participe ainsi de manière éclairée aux débats de la commission recherche et de ses différents groupes de travail en comprenant mieux la culture scientifique de mes collègues STS. Ensuite, mes incursions dans le champ des littéracies avancées ont contribué à faire évoluer mes pratiques pédagogiques pour viser une plus grande efficacité dans l'amélioration des compétences rédactionnelles de mes étudiants, de la licence au doctorat. Enfin, la mise en place d'un réseau national, via *écri+*, me permet de défendre inlassablement l'idée que les enseignants et les chercheurs de l'enseignement supérieur public sont les mieux qualifiés pour mener à bien la formation et la certification en langue française de leurs étudiants.