La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase

The phraseology of the contemporary fiction in the PhraseoBase's corpora and applications

Sascha Diwersy, Laetitia Gonon, Vannina Goossens, Olivier Kraif, Iva Novakova, Julie Sorba et Ilaria Vidotto

Édition électronique

URL : <https://journals.openedition.org/corpus/6101>

DOI : 10.4000/corpus.6101

ISSN : 1765-3126

Éditeur

Bases ; corpus et langage - UMR 6039

Référence électronique

Sascha Diwersy, Laetitia Gonon, Vannina Goossens, Olivier Kraif, Iva Novakova, Julie Sorba et Ilaria Vidotto, « La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase », *Corpus* [En ligne], 22 | 2021, mis en ligne le 02 février 2021, consulté le 04 mars 2022. URL : <http://journals.openedition.org/corpus/6101> ; DOI : <https://doi.org/10.4000/corpus.6101>

Ce document a été généré automatiquement le 4 mars 2022.

© Tous droits réservés

La phraséologie du roman contemporain dans les corpus et les applications de la PhraseoBase

The phraseology of the contemporary fiction in the PhraseoBase's corpora and applications

Sascha Diwersy, Laetitia Gonon, Vannina Goossens, Olivier Kraif, Iva Novakova, Julie Sorba et Ilaria Vidotto

1. Présentation générale du projet PhraseoRom

- 1 Le principal apport du projet ANR DFG PhraseoRom¹ a été d'élaborer une typologie structurelle et fonctionnelle des constructions lexico-syntaxiques (CLS) spécifiques au discours romanesque francophone, anglophone et germanophone à partir des années 1950. Nous avons exploré, en procédant de manière entièrement inductive, selon l'approche *corpus driven*, la pertinence des phénomènes phraséologiques pour la caractérisation de la langue et des sous-genres littéraires.
- 2 Ce projet interdisciplinaire en Humanités numériques a été construit autour de trois principaux volets : TAL, linguistique et stylistique. Le volet informatique (Corpus et TAL) nous a permis d'extraire, grâce à des approches textométriques² innovantes, des expressions polylexicales sous forme d'Arbres Lexico-syntaxiques Récurrents (ALR), à partir de corpus littéraires trilingues, comparables et parallèles (voir sections 2.1 et 2.2). Dans le volet linguistique, ces ALR ont été classés, annotés et analysés selon une méthodologie spécifique (voir section 2.3), inspirée des modèles fonctionnels et contextualistes britanniques (Sinclair 2004 ; Hoey 2005). Dans le volet stylistique (section 2.4), il a été procédé à des annotations fines des fonctions discursives (FD) des ALR, qui génèrent des motifs textuels (Legallois 2006 ; Longrée & Mellet 2013 ; Novakova & Siepmann 2020). Dans la section 3, nous proposons deux applications de la méthode des ALR pour distinguer les sous-genres romanesques, d'une part, au sein de la littérature de l'imaginaire (section 3.1 : la science-fiction vs. la fantasy), et d'autre

part, au sein de la littérature réaliste (section 3.2 : le roman policier vs. le roman sentimental). Enfin, nous présentons, dans la section 4, l'outil Lexicoscope 2.0³ qui permet d'interroger les corpus de la PhraseoBase⁴, ainsi que les fonctionnalités de l'application PhraseoLing regroupant, via une interface ergonomique, les annotations linguistiques et stylistiques des données extraites des corpus romanesques français et anglais.

2. Le corpus PhraseoRom. La méthodologie d'extraction et d'analyse des données linguistiques et stylistiques

- 3 Les œuvres littéraires des corpus PhraseoRom ont été réparties en 6 sous-genres : romans de littérature générale (GEN), sentimentaux (SENT), policiers (POL), historiques (HIST), de science-fiction (SF), fantasy (FY). Ces œuvres ont été sélectionnées en fonction de critères chronologiques (romans postérieurs à 1950) et éditoriaux (pour GEN, la reconnaissance des auteurs par la critique, notamment lors de l'attribution de prix littéraires, et pour les autres sous-genres, la collection dans laquelle elles sont publiées).

2.1. Les corpus comparables et parallèles intégrés dans la PhraseoBase

- 4 Pour les corpus comparables, l'objectif était d'arriver à un volume d'environ 100 millions de mots par langue, tout en respectant un certain équilibre entre les genres. Pour les corpus parallèles français-anglais et anglais-français, les œuvres originales ont été alignées avec leur traduction. Le Tableau 1 ci-dessous montre la répartition des auteurs, des textes et des tokens, en fonction des langues dans les deux types de corpus :

Tableau 1. Répartition des corpus PhraseoRom selon les langues (auteurs, textes, tokens)

Corpus comparables			
Langue	Auteurs	Textes	Tokens
EN	376	849	112 610 090
FR	418	1131	103 819 368
DE	237	746	89 451 097
Corpus parallèles			
EN → FR	230	495	68 512 492
FR → EN	139	298	33 754 332

- 5 En fonction de la disponibilité des textes, certains genres sont légèrement surreprésentés (par exemple, GEN pour le français et HIST pour l'allemand) comme le montrent le Tableau 2 et la Figure 1 :

Tableau 2. Répartition en sous-genres des corpus comparables⁵

Sous genres	Tokens FR	Tokens EN (GB)	Tokens EN (US)	Tokens DE
FY	13 966 573	24 590 831	6 467 561	16 774 462
GEN	34 593 337	15 696 519	283 291	9 068 325
HIST	14 193 643	14 606 939	92 843	25 454 115
POL	18 020 732	16 117 528	2 891 981	11 117 326
SENT	9 516 638	12 836 609	0	18 689 013
SF	13 528 445	16 507 028	2 518 960	8 347 856
Total	103 819 368	100 355 454	12 254 636	89 451 097

- 6 La Figure 1 ci-dessous montre les proportions des différents sous-genres au sein des corpus anglais (EN), français (FR) et allemand (DE) :

Figure 1. Répartition en fonction des genres (en pourcentage du nombre de tokens)

- 7 Tous ces corpus ont d'abord été rendus accessibles sur l'interface du Lexicoscope v1 (Kraif & Diwersy 2012 ; Kraif 2016), avant leur déploiement progressif vers le Lexicoscope 2.0.

2.2. L'extraction des Arbres Lexico-syntaxiques Récurrents (ALR)

- 8 En s'appuyant sur l'annotation syntaxique des corpus, le Lexicoscope permet d'extraire des associations statistiques significatives reliant des unités lexicales par une relation de dépendance syntaxique. À la différence des motifs séquentiels émergents de Quiniou *et al.* (2012), ces ALR correspondent à des structures hiérarchiques (ou arbres de dépendance). La Figure 2 illustre l'expression récurrente, spécifique à POL, <allumer une nouvelle cigarette> qui se présente sous forme d'Arbre Lexico-syntaxique Récurrent (ALR) :

Figure 2. L'ALR *allumer une nouvelle cigarette*, spécifique au sous-genre policier (POL)

9 Les fréquences des ALR dans chacun des différents corpus sont comparées afin de mesurer leur spécificité : un ALR est spécifique quand sa fréquence relative dans l'un des sous-corpus est significativement supérieure à sa fréquence dans l'ensemble des autres sous-corpus. Le seuil de spécificité statistique (l'indice LLR, Dunning 1993) a été fixé à 10,83⁶. Selon la méthodologie développée dans le cadre de PhraseoRom, les critères retenus pour la sélection des ALR sont les suivants :

- le critère de fréquence absolue : le nombre d'occurrences d'un ALR doit être supérieur à 10 ;
- le critère de dispersion : l'ALR apparaît au moins chez 20 % des auteurs d'un sous-corpus pouvant aller jusqu'à 50 % des auteurs du corpus ;
- le critère morpho-syntaxique : l'ALR doit comporter obligatoirement un verbe, ce qui exclut d'emblée des expressions exclusivement référentielles (<à la tombée de la nuit>, <Monsieur le Procureur>, <les nains de jardin>) ;
- le critère de longueur : les ALR doivent comporter au moins 3 mots dont un verbe pivot, et ce jusqu'à 7 mots.

10 Grâce à l'ensemble de ces critères, ont été extraits des six sous-genres romanesques 8415 ALR pour le français et 1955 ALR pour l'anglais.

2.3. Méthodologie de l'analyse linguistique : la grille sémantique

11 Le principal enjeu du projet était d'identifier et de décrire les genres littéraires non pas en termes de rhétorique, de style individuel des auteurs ou d'univers de référence thématique, comme c'est souvent le cas en stylistique classique, mais à travers des constructions lexico-syntaxiques (CLS) récurrentes. Une fois extraites automatiquement sous forme d'ALR, les données brutes ont été classées et analysées sur quatre niveaux : lexical, syntaxique, sémantique et discursif. La grille conçue pour l'annotation sémantique⁷ des items retenus pour le français et l'anglais (Goossens,

Jacquot & Dyka 2020) comporte huit dimensions sémantiques auxquelles sont associées, pour certaines dimensions (colonne 1 du Tableau 3), un ensemble de valeurs (colonne 2 du Tableau 3).

Tableau 3. Grille d'analyse sémantique des ALR

Dimension	Valeur	Exemple d'ALR
action	indéfini	<je vais le faire>
	mouvement	<descendre de sa monture>, <held her in his arms>
	déplacement	<il traversa le hall>, <walk along the beach>
	autre	<éteindre la lumière>, <blow out the candle>
événement		<l'écran s'éteint>, <the door opened>
état	indéfini	<être en danger>, <was in the house>
	animé	<être trempé de sueur>, <be in love>
	inanimé	<l'endroit était désert>, <the body lay>
	abstrait	<la situation est grave>, <there was in the sky>
communication	indéfini	<répondre avec un sourire>, <accept the offer>
	verbale	<échanger quelques mots>, <have a chat>
	physique	<adresser un sourire>, <give a smile>
cognition		<je ne me souviens pas>, <think for a moment>
qualia	indéfini	<elle ne ressentait rien>, <like the smell>
	affect	<se mettre en colère>, <I'm shocked>
	sensation	<avoir froid>, <my heart is beating>
	perception	<je sentis l'odeur>, <breathing the scent>
temps		<c'est la première fois>, <spend a week>
lieu		<l'appartement était petit>, <he was in town>
autre		<mon père est>, <the driver was>

- 12 La catégorie « autre » permet de classer les ALR correspondant à nos critères mais qui ne rentrent dans aucune de ces huit dimensions. La valeur « indéfini » permet de catégoriser les ALR auxquels plusieurs valeurs peuvent être attribuées sans qu'il soit possible d'en sélectionner une seule⁸.
- 13 Cette grille d'analyse a été appliquée manuellement et systématiquement aux ALR sélectionnés. Les fichiers dans lesquels ont été codés les ALR dans les deux langues ont été implémentés dans l'application PhraseoLing (cf. section 4.2). Les tableaux comportant les annotations linguistiques regroupent plusieurs informations⁹ dont les plus importantes sont les suivantes : l'ALR extrait, la requête d'extraction¹⁰, le LLR, le genre pour lequel il est spécifique, la fréquence, la dispersion et cinq exemples représentatifs (avec les métadonnées).

2.4. Méthodologie de l'analyse stylistique

- 14 Cette première étape d'annotation sémantique des ALR a permis ensuite de les regrouper automatiquement¹¹ et de faire émerger des *motifs textuels*. Cette notion¹², récente en phraséologie étendue (Legallois 2006 ; Longrée, Luong & Mellet 2008), désigne des patrons lexico-syntaxiques récurrents présentant (1) des régularités et des variations lexicales et syntaxiques (aux niveaux paradigmatique et syntagmatique), établies avec des méthodes textométriques, et (2) des fonctions discursives spécifiques. Par son caractère à la fois structurant et caractérisant (Mellet & Longrée 2012 ; Longrée & Mellet 2013), le motif s'est révélé une notion opératoire pour la distinction des sous-genres littéraires. Plus généralement, il a été montré (Novakova & Siepmann, 2020) que la notion de motif assure le lien entre le micro-niveau (les récurrences phraséologiques spécifiques) et le macro-niveau discursif (celui du script narratif ou fictionnel, Baroni 2007).
- 15 Dans notre méthodologie, les ALR constituent des guides pour identifier des motifs. Ensuite, un système d'annotations fines des fonctions discursives (FD) a été élaboré afin de prendre en compte la dimension textuelle de ces expressions lexico-syntaxiques, à savoir leur rôle dans la structuration des textes littéraires.

2.4.1. Les fonctions discursives

- 16 Les FD désignent les fonctions stylistiques qu'occupent les CLS dans le discours fictionnel romanesque. Elles contribuent ainsi à en assurer la cohérence (Martin 1983 : 100). La FD dépend du contexte, mais aussi du sous-genre romanesque (voir Baroni 2015). Dans un texte narratif comme le roman, les FD sont principalement (a) *narratives* ou (b) *descriptives* (Adam 2011 : 267), la CLS permettant à l'action de progresser, ou bien s'intégrant dans une description. L'analyse détaillée des données linguistiques a permis d'affiner cette première typologie. Ainsi, ont été rajoutées deux nouvelles FD, notamment l'*infranarrative* (lorsque les CLS renvoient à des actions minimales, sans rapport direct avec l'intrigue principale) et l'*infradescriptive*, qui s'applique à une action ou à un geste répété, caractérisant le personnage (qui fume beaucoup par exemple, comme c'est le cas des personnages dans POL). De plus, un cas particulier de la FD descriptive, la FD *affective*, a été identifié lorsque la CLS décrit l'état d'esprit du personnage (par exemple dans « Sarah écrasa nerveusement sa cigarette »).
- 17 Enfin, deux autres FD ont également été modélisées. D'une part, la FD *pragmatique* se rapporte aux actes de langage entre les personnages (*N'en faites rien, madame, je vous en prie, s'écria Fabienne*). D'autre part, la FD *cognitive* s'applique à des CLS impliquant des processus cognitifs comme les hypothèses ou les réflexions (*Je sais pas ce qu'il va devenir*).

2.4.2. La grille stylistique

- 18 La première étape de l'annotation stylistique a débuté par le repérage des ALR susceptibles de former un motif phraséologique. En raison de leur spécificité, ces ALR ont été appelés « leaders » (par ex. l'ALR <apparaître_sur_écran> dans SF), dans la mesure où c'est à partir de leurs composantes syntaxiques de base (« syntaxe cœur », par ex. V+SNPrép pour l'ALR <apparaître_sur_écran>), qu'ils présentent des variations syntagmatiques ou paradigmatiques et peuvent constituer, de ce fait, un même motif.

- 19 Une fois l'ALR « leader » et les ALR similaires identifiés, l'annotateur a renseigné l'expression régulière qui caractérise l'ALR dans le Lexicoscope, afin d'observer le contexte dans lequel le motif est susceptible d'apparaître. L'examen des extraits de texte constitue une étape essentielle pour la détermination des FD associées au motif et, plus largement, pour l'interprétation stylistique du rôle que le segment phraséologique joue aussi bien dans les textes où il figure que dans le sous-genre dont il est spécifique. Les informations recueillies à partir de l'observation de ces exemples sont renseignées dans le fichier de l'annotation : l'identifiant (Id) attribué à chaque motif ; son étiquette *<apparaître_sur_écran>* ; le sous-genre pour lequel motif est spécifique (SF) ; la syntaxe cœur (le motif dans sa configuration syntaxique minimale V+SNprép) ; la position du motif dans la phrase et au-delà ; la distribution des éléments autour de la syntaxe cœur ; au moins 5 exemples extraits du Lexicoscope avec les métadonnées (auteur, titre, année de publication) ; la FD attribuée au motif, l'interprétation stylistique et le commentaire de l'annotateur. Des fichiers d'annotation stylistique¹³, constitués pour stocker ces données, ont servi à l'élaboration de la PhraseoBase et, en particulier, de l'application PhraseoLing présentée dans la section 4.

3. Des constructions lexico-syntaxiques (CLS) polylexicales aux motifs textuels

- 20 Dans cette section, nous présentons les résultats obtenus à partir de l'analyse sémantique et syntaxique des ALR qui forment des motifs permettant de distinguer les sous-genres SF vs. FY (3.1) et POL vs. SENT (3.2).

3.1. Les motifs spécifiques à la science-fiction (SF) et à la fantasy (FY)

- 21 L'analyse des ALR de la SF et de la FY en français et en anglais (Goossens, Jacquot & Dyka 2020), deux genres proches souvent englobés sous l'étiquette « littérature de l'imaginaire », s'est révélée très productive.
- 22 Tout d'abord, certaines catégories sémantiques ont fait émerger des « marqueurs génériques », c'est-à-dire des ALR spécifiques à l'un ou l'autre genre, et qui relèvent d'actions ou de descriptions permettant immédiatement l'identification du genre. C'est le cas, par exemple, de la catégorie « déplacement » en français et des catégories « lieu » et « état » en anglais, qui mettent notamment en avant l'espace, spécifique de la SF, et la forêt, spécifique de la FY. De même, la catégorie « mouvement », en français comme en anglais, regroupe des ALR qui renvoient à des actes guerriers et à des actes de déférence très spécifiques de la fantasy.
- 23 L'analyse sémantique des ALR a également permis de mettre en lumière des catégories nettement moins attendues : tout d'abord, la cognition, très spécifique à la SF dans les deux langues, qui s'oppose à un processus de décision beaucoup plus intuitif en FY. Les ALR relevant de cette catégorie mettent l'accent sur le raisonnement dans les romans de SF (*<il essaye de comprendre>*, *<il en déduit>*, *<mettre au point>*, *<trouver la solution/le moyen>*) alors que dans le corpus de FY, ils relèvent plus de l'imagination et du rêve (*<plongé/perdu dans ses pensées>*, *<il reprend ses esprits>*). La catégorie « communication » met elle aussi en évidence des caractéristiques plutôt inattendues de ces deux genres.

Cette catégorie est surtout exploitable pour le français, l'écrasante domination du verbe *to say* en anglais (voir Diwersy et al. 2020) rendant la dimension moins saillante en anglais. Même si la communication physique est peu présente dans ces deux genres, elle y révèle une différence notable : elle n'est véritablement dirigée vers autrui que dans la FY (<hausser les épaules>, <je secoue la tête> (SF) vs <adresser/jeter/échanger/accorder un regard/signe de tête>, <déposer un baiser>, <caresser sa joue> (FY)). La communication verbale est encore plus saillante dans les deux sous-genres : elle révèle des modes de communication très différents. La SF offre des ALR relevant de la communication à distance (<laisser/recevoir un message>, <établir le contact/couper la communication>) et des ALR très factuels construits autour de verbes génériques permettant la structuration de l'interaction (<dire d'une voix/sur un ton>, <je veux parler>, <poser la question>). La FY offre un panorama de la prise de parole beaucoup plus diversifié, notamment en termes de volume sonore, allant du silence (<garder/imposer le silence>) aux hurlements (<pousser un cri de guerre/raie/douleur/un hennissement/rugissement/grognement>, <hurler de terreur/raie>), en passant par les murmures (<des murmures s'élèvent>). Cette surreprésentation des affects et des attitudes des personnages est spécifique de la FY dans son opposition avec la SF.

3.2. Le motif de la cigarette dans POL vs SENT

- 24 Nos résultats ont révélé comme hautement spécifique de POL l'ALR <allumer une cigarette>, même si on le rencontre aussi dans SENT (Gonon, Goossens & Novakova 2020). Le Tableau 4 récapitule la fréquence, la dispersion et le LLR dans les deux sous-corpus :

Tableau 4. Spécificités de l'ALR <allumer une cigarette>

<allumer une cigarette>	POL	SENT
Fréquence	343	124
Dispersion	124	43
LLR	151,8	9,24

- 25 L'ALR <allumer une cigarette>, cœur du motif, présente une forte variation paradigmatique sur le nom (*cigarette*, *cigare*, *pipe*, *clop(e)*, *tige*, *Gitane*, *Camel*, *Craven*) ou sur le verbe relevant de différences de nature aspectuelle : début (*rouler*, *allumer*), déroulement (*fumer*, *griller*) et fin (*éteindre*, *terminer*, *écraser*) du procès. Le motif, sous sa forme minimaliste, agrège aussi de nombreuses extensions syntagmatiques (adjectifs, adverbes, SN prép). Ainsi, nous avons identifié le schéma suivant pour ce motif :
- X allume (fume, écrase, éteindre, prendre) (nerveusement) une (nouvelle) cigarette (mentholée) (contre le mur, sous le bureau)
- 26 Souvent dans POL, le motif de la cigarette remplit une double FD, à la fois descriptive et affective, et ce, grâce aux différentes extensions syntagmatiques de la syntaxe cœur. Ainsi, en (1), pour le personnage féminin, Diane :
- (1) Diane *alluma une cigarette d'agacement*. Il y avait les cigarettes de calme, les cigarettes d'exaspération, les cigarettes de plaisir – les plus rares –, les cigarettes de

tension, les cigarettes de réflexion, les cigarettes d'ennui. Une nomenclature précise que seuls les fumeurs connaissent. (Andrea H. Japp, *Dans la tête, le venin*, 2009.)

- 27 L'auteur évoque toute une panoplie de sentiments à travers les SN prép qui étendent le N cigarette (*de calme, d'exaspération, de plaisir, de tension*, etc.). Dans d'autres cas, le SN prép est complément circonstanciel de lieu (2) :

(2) *J'écrasai ma cigarette contre un mur, jetai le mégot sur le sol.* (O. Gay, *Les talons hauts rapprochent les filles du ciel*, 2012)

- 28 Le motif ainsi étendu fait partie du script « fumer une cigarette », ici décliné avec variations : plutôt que d'écraser sa cigarette *dans un cendrier* et de jeter le mégot à la poubelle comme il est d'usage, le héros narrateur transgresse les règles du script (et celles de la propreté) et se représente ainsi comme un *bad boy* (FD descriptive dans ce cas).
- 29 À la différence de POL, le motif de la cigarette dans SENT est plutôt associé à l'élégance des personnages (3), ou au jeu de séduction¹⁴ entre eux (4) :

(3) *Elle s'alluma une cigarette avec un certain style.* (G. Musso, *Demain*, 2013).

(4) *Juan est silencieux, il allume une cigarette, elle la lui prend aussitôt des doigts et la porte à ses lèvres.* (M. Levy, *Où es-tu ?*, 2001)

- 30 Le motif <*allumer une cigarette*> n'est jamais un motif de premier plan narratif : en (3) et en (4), la fonction est infra-narrative. Le motif revêt néanmoins un rôle différent dans les deux sous-genres : les unités linguistiques variables qu'il regroupe contribuent à créer l'atmosphère tendue dans POL ou l'ambiance plus feutrée dans SENT.

4. Présentation générale de la Phraseobase

- 31 La PhraseoBase a pour vocation de donner accès libre à l'ensemble des ressources élaborées dans le cadre du projet PhraseoRom, c'est-à-dire les corpus et les codages linguistiques et stylistiques des données extraites grâce à la méthodologie précédemment décrite. Cette section présente l'interface utilisée, le Lexicoscope 2.0 (4.1) puis la base de données PhraseoLing qui permet d'accéder aux résultats des analyses (4.2).

4.1. L'interface du Lexicoscope 2.0

- 32 Les corpus n'étant pas, en l'état, diffusables en ligne, pour des questions de droits, nous avons choisi de les intégrer à une interface telle que le Lexicoscope afin d'en permettre l'interrogation sous forme de statistiques lexicométriques et de concordances, qui ne donnent accès qu'à des contextes restreints (Kraif 2019). Dans le but de comparer entre elles les différentes subdivisions du corpus, notamment au niveau des principaux sous-genres, il a été nécessaire de développer une nouvelle version, le Lexicoscope 2.0, accessible en ligne depuis 2020. Cette nouvelle version de l'interface, tout comme la précédente, permet d'extraire des concordances et des tableaux de cooccurrents autour d'une expression pivot, qui peut être un mot simple ou une expression complexe.

4.1.1. Définition du corpus de travail

- 33 Pour contraster différents sous-corpus, l'interface permet de définir deux types de corpus de travail : les collections prédéfinies et les corpus personnalisés. Les collections prédéfinies sont des ensembles de corpus déjà subdivisés en six sous-corpus. Concernant le projet PhraseoRom, les collections sont au nombre de 3 pour l'anglais et le français, et de 1 pour l'allemand. Pour l'anglais et le français, dans la mesure où une partie du corpus correspond au corpus parallèle, on a le choix entre les collections suivantes : intégral, langue source FR ou EN seulement, et parallèle EN-FR ou FR-EN. Chacune de ces collections est subdivisée par sous-genre : FY, GEN, HIST, POL, SENT et SF.
- 34 Si l'on cible un ensemble de textes plus restreints, il est aussi possible de construire un corpus personnalisé. Dans ce cas, la définition se fait d'abord par le choix d'une langue d'interrogation et, éventuellement, d'une langue alignée (si l'on se situe dans le corpus parallèle). Ensuite, l'utilisateur peut ajouter un ou plusieurs sous-corpus, définis en fonction d'une combinaison de critères : sous-genre, année de publication, titre, auteur, langue source du document. Par exemple, pour interroger uniquement les œuvres écrites en anglais britanniques et traduites en français, il faudra définir un corpus personnalisé. Une fois ce corpus constitué, il est possible d'en sauvegarder la définition, à condition d'utiliser l'outil en tant qu'utilisateur connecté¹⁵.

4.1.2. Statistiques textométriques des collections prédéfinies

- 35 Lors du choix d'une collection prédéfinie, on peut accéder à des statistiques concernant le vocabulaire du corpus ainsi que les collocations et colligations. Ces statistiques permettent d'avoir la liste des 5 000 lemmes les plus fréquents apparaissant dans au moins deux sous-genres différents (voir Figure 3).

Figure 3. Statistiques globales pour les sous-corpus (ici HIST)

- 36 Comme la plupart des index hiérarchiques, cette liste est triée par fréquence décroissante, et un simple clic sur un lemme aura pour effet de déclencher une requête sur ce lemme. Ce tableau permet par ailleurs, par simple clic sur la colonne « Comparer », d'accéder rapidement à la comparaison des spécificités (LLR) par sous-genre, afin de faire contraster les corpus de manière synthétique. Par exemple, on voit dans la Figure 4 ci-dessous la répartition des spécificités pour le verbe *demander*, qui se trouve être spécifique à SENT et à POL, et sous-employé dans GEN ou SF.

Figure 4. Spécificités par sous-corpus pour le lemme *demander*

- 37 Ces données fréquentielles et ces calculs de spécificité sont également accessibles pour les parties du discours (où l'on voit, par exemple, que les signes de ponctuation sont spécifiques à GEN), ainsi que pour les collocations binaires et les étiquettes de relation syntaxique. Pour les collocations, une mesure d'association est fournie en supplément de la fréquence. Nous avons choisi l'information mutuelle spécifique, qui favorise les associations entre unités de basse fréquence et qui se trouve en meilleure complémentarité avec la fréquence brute. Ainsi, comme on le voit dans la Figure 5, un tri fondé sur cette mesure permet de mettre en avant des entités nommées (*Garbo*, *Giorgio* etc.), des expressions en langue étrangère (*feel like*, *all right* etc.) ou des unités polylexicales (*congestion pulmonaire*) :

Figure 5. Index hiérarchique des collocations binaires, trié par PMI (information mutuelle spécifique)

Rang	lemma1	lemma2	Fréquence	Dispersion	association	Comparer
901	Garbo	Greta	83	5	11.82	<input type="button" value="cmp"/>
902	like	feel	7	2	11.82	<input type="button" value="cmp"/>
903	Giorgio	Maggiore	13	2	11.82	<input type="button" value="cmp"/>
904	Jérémias	Mistère	6	2	11.81	<input type="button" value="cmp"/>
905	Heydrich	Reinhard	19	4	11.81	<input type="button" value="cmp"/>
906	Quid	quo	4	2	11.8	<input type="button" value="cmp"/>
907	land	mans	4	2	11.8	<input type="button" value="cmp"/>
908	Martini	dry	18	3	11.8	<input type="button" value="cmp"/>
909	all	right	17	4	11.8	<input type="button" value="cmp"/>
910	Buster	Keaton	40	3	11.8	<input type="button" value="cmp"/>
911	pulmonaire	congestion	13	3	11.8	<input type="button" value="cmp"/>

- 38 La comparaison des spécificités (bouton 'cmp') montre qu'alors que *congestion pulmonaire* est spécifique de GEN et HIST, *feel like* est spécifique de SENT et *Martini dry* est spécifique de POL.

4.1.3. Requêtes

- 39 Comme dans la première version du Lexicoscope, le fait d'entrer une expression simple ou complexe permet d'accéder à sa définition sous forme de requête formalisée précisant les lemmes, parties du discours, et éventuellement les relations syntaxiques entre les mots – afin de mettre en œuvre une forme de requête basée sur l'exemple (voir Figure 6), comme chez Augustinus, Vandeghinste & Vanallemeersch (2016) ou Wang (2017).

Figure 6. Suggestion de requête avancée pour 'fume une cigarette'

- 40 Le résultat de la requête s'affiche dans un nouvel onglet, distinct de l'onglet « tableau de bord » qui permet de préciser corpus et requête. Le résultat contient tout un ensemble de données synthétiques (voir Figure 7) :
- les statistiques de l'expression cherchée, globalement et par corpus ;
 - les concordances de l'expression cherchée, avec les métadonnées et les contextes élargis ;
 - les cooccurents de l'expression cherchée, sous forme de diagrammes à barre et de tableaux contenant l'ensemble des données de cooccurrence (fréquence de la base et du cooccurrent, cofréquence, relation syntaxique, indice du loglike, fréquence de document, etc.) ;
 - un tableau synoptique des principaux cooccurents sous la forme de Word Sketch, inspiré du Sketch Engine (Kilgariff 2004).
- 41 Enfin, si l'utilisateur le désire, et s'il a les droits requis¹⁶, il peut lancer en complément l'extraction des ALR autour de l'expression cherchée ainsi que le calcul de statistiques multivariées sur les sorties (clustering hiérarchique ascendant, analyse en composantes principales, analyse des correspondances).

Figure 7. Extrait de Word Sketch pour l'ALR <fumer une cigarette>

Statistiques Concordances Cooccurrences <u>Word Sketch</u> Statistiques multivariées			
Global FY GEN HIST POL SENT SF			
Adjectifs modificateurs			
roulé_ADJ	4	47.27	
dernier_ADJ	9	43.02	
américain_ADJ	5	37.08	
anglais_ADJ	4	28.65	
autre_ADJ	8	22.26	
Adverbes modificateurs du verbe			
tranquillement_ADV	6	55.21	
dehors_ADV	6	38.62	
à la main_ADV	3	19.28	
encore_ADV	7	17.07	
plus tard_ADV	2	7.11	
parfois_ADV	2	6.42	
Noms modifiant le pivot			
bout_NOUN	7	27.81	
tabac_NOUN	3	18.08	
fenêtre_NOUN	5	13.41	
filtre_NOUN	1	8.08	
Noms modificateurs du verbe			
trottoir_NOUN	6	31.71	
couloir_NOUN	6	29.17	
seuil_NOUN	4	21.40	
filtre_NOUN	2	18.94	
balcon_NOUN	3	18.62	
terrasse_NOUN	3	12.33	

- 42 Là encore, dans le souci de faire apparaître les contrastes de manière synthétique, chaque type de résultat apparaît soit de manière globale pour l'ensemble du corpus, soit dans un onglet séparé pour chaque sous-corpus, avec les mesures de spécificité afférentes.

4.2. L'application PhraseoLing

- 43 Pour donner accès aux résultats issus des analyses sémantiques, discursives et stylistiques, nous avons mis en place une base de données relationnelle interrogeable via une interface en ligne¹⁷, PhraseoLing. Celle-ci comprend deux formulaires, l'un pour les données syntactico-sémantiques, l'autre pour les données stylistiques, permettant de concevoir des requêtes à différents niveaux d'abstraction. Ainsi, dans le cas des données syntactico-sémantiques, on peut rechercher des ALR non seulement en fonction de la langue (français ou anglais), du genre littéraire et de la classe sémantique dont ils relèvent, mais aussi en fonction des mots qu'ils impliquent, de la classe grammaticale de ceux-ci ou encore de la fonction syntaxique que ces mots occupent au sein de l'ALR.
- 44 L'application principale associée à ces requêtes est l'affichage d'entrées lexicographiques, classées soit par groupe d'ALR ou genre littéraire, pour ce qui est des données syntactico-sémantiques, soit par motif, fonction discursive ou genre, pour ce qui est des données stylistiques. La Figure 8 montre l'une des entrées obtenues à la suite de la requête visant les ALR contenant l'adjectif *heureux* spécifiques à FY.

Figure 8. Entrée lexicographique résultant de la requête visant les ALR contenant l'adjectif *heureux* spécifiques à FY

FY
Je suis heureux de revoir

Statistiques

Fréquence : 28 Score de spécificité : 18.8767 Dispersion : 13

Fonctions et exemples

Catégorie sémantique : qualia:affect

Exemples (corpus): <1=de,c=PREP,#1>&&<1=heureux,c=ADJ,#2>&&<1=je,c=PRON,#3>&&<1=revoir,c=VERB,#4>&&<1=être,c=VERB,#5>:: (OBJ_SPRED,5,2) (PREPOBJ,4,1) (ADJMOD,2,4) (SUBJ,5,3) ↻

Exemples (base): Ange - "Les trois lunes de Tanjor 1 Le peuple turquoise" (9961) :) [Je] [suis] si [heureux] [de] vous [revoir] vivante.\n
 Ange - "Les trois lunes de Tanjor 3 La mort d'Ayesha" (14686) : – [Je] [suis] [heureux] [de] vous [revoir], Marikani, dit le Maître des Exilés.
 Fetjaine Jean-Louis - "Chroniques des elfes 3 Le sang des elfes" (5856) : – [Je] [suis] [heureux] [de] te [revoir], dit-il.
 Grimbert Pierre - "Le secret de Ji 1" (27604) : [Je] [suis] [heureuse] [de] vous [revoir].\n
 Loevenbruck Henri - "Gallica L'intégrale de la trilogie" (20899) : – [Je] [serai] [heureux] [de] vous [revoir] alors, Bernard, car... Car j'aurai besoin des conseils d'hommes de bonne volonté pour ce que je veux faire quand nous aurons sauvé les Brumes...\n
 Peru Olivier - "Les hauts conteurs 5 La mort noire" (32288) : [Je] [suis] [heureux] [de] te [revoir] saine et sauve.\n
 Peru Olivier - "Les hauts conteurs 5 La mort noire" (32955) : – [Je] [suis] [heureux] [de] te [revoir], Asa, mentit-il effrontément.
 Robert Michel - "1 L'agent des ombres 1 L'ange du chaos" (24071) : [Je] [suis] si [heureuse] [de] te [revoir].
 Robert Michel - "8 L'agent des ombres 8 Ange et Loki" (74058) : – Ah, si tu savais comme [je] [suis] [heureux] [de] te [revoir], Petit-Homme !
 Robillard Anne - "Les dragons de l'Empereur Noir" (62478) : [Je] [suis] fort [heureux] [de] vous [revoir].\n

- 45 Outre les éléments d'ordre qualitatif (catégorie sémantique, schéma syntaxique, exemples d'usage), les entrées consacrées aux ALR fournissent également des indications d'ordre quantitatif (fréquence d'apparition, score de spécificité LLR, dispersion) provenant des calculs effectués au moment de l'extraction automatique. Quant aux entrées consacrées aux motifs stylistiques, elles comprennent des exemples d'usage et indiquent la fonction discursive, d'éventuelles composantes facultatives et la distribution syntaxique, à moins que l'un de ces éléments ne soit utilisé comme paramètre de regroupement macrostructurel. La Figure 9 illustre l'entrée se rapportant au motif <apparaître sur écran>, qui est associée à la FD narrative dans les romans de SF :

Figure 9. Entrée lexicographique se rapportant au motif stylistique <apparaître sur écran> dans les romans de science-fiction (SF)

apparaître_sur_écran

SF

narratif

apparaître_sur_écran

Fonctions et exemples

Patron syntaxique : N+DET+V+SPrép

Position : DD

Distribution : --

Composante facultative : --

Exemples :

[s18389; Guieu J., *Hantise sur le monde*, 1953] : Un Mokranien casqué apparut sur l'écran.\n – Pourquoi nous avoir stoppés ? s'indigna-t-il.\n – Depuis quand, Boukoniev, discutes-tu les ordres du Chef Letvinof ?

[s26866; Walther D., *Mais l'espace... Mais le temps...*, 1998] : Quand il se fut installé dans une chambre à peu près climatisée, il appuya sur la touche de l'interphone et le visage fatigué du portier s'inscrivit sur l'écran : – Savez-vous si dans cette ville réside encore un certain Félide Gershah ?\n

[s4736; Barbéri J., *Le crépuscule des chimères*, 2002] : Il indiqua négligemment du doigt les suites ininterrompues de chiffres qui défilaient sur les écrans. « Spectrographie vibrationnelle... Quasiment identique à celle que nous avions relevée près des brèches. »\n Anjel repensa aussitôt aux dauphins.

- 46 Au-delà de l'affichage d'entrées lexicographiques, les requêtes portant sur les données syntactico-sémantiques enchaînent sur différents traitements statistiques¹⁸ dont l'objectif est la caractérisation et la classification des genres littéraires en fonction des ALR ou des catégories sémantiques répertoriés par une requête donnée. Les procédés mis en place¹⁹ s'inscrivent dans la lignée de la textométrie (AFC, calcul de spécificités ; cf. Lebart & Salem 1994) et de la stylométrie contemporaine (*clustering*, ACP, échelonnement multidimensionnel ; cf. Eder, Rybicki & Kestemont 2016). En guise d'exemple, le graphique de la Figure 10 donne les résultats d'une AFC basée sur la distribution des ALR relevant des catégories sémantiques *communication:physique*, *communication:verbale* et *communication:indéfini* par genre littéraire.

Figure 10. Classification par AFC des genres littéraires en fonction des ALR associés aux catégories *communication:physique*, *communication:verbale* et *communication:indéfini*

- 47 L'AFC montre que, par rapport aux ALR sélectionnés, les romans de GEN s'opposent à ceux de FY sur l'axe principal (en horizontal) et que le deuxième axe (en vertical) met en opposition les romans historiques (HIST) avec les romans policiers (POL) et sentimentaux (SENT).
- 48 Parmi les procédés implémentés à travers la plateforme, on trouve également la classification ascendante hiérarchique (CAH). Une CAH appliquée aux genres romanesques à partir des ALR associés aux catégories sémantiques *action:déplacement* et *action:mouvement*, et effectuée avec la mesure de distance cosinus (Jannidis *et al.* 2015) ainsi que la méthode d'agrégation de Ward, permet d'obtenir la matrice de distances (représentée sous forme de heatmap, Figure 11) et le dendrogramme (Figure 12).

Figure 11. Heatmap représentant la matrice de distances entre genres de romans, obtenue à partir des ALR relevant des catégories sémantiques *action:déplacement* et *action:mouvement*

Figure 12. Dendrogramme résultant de la classification ascendante hiérarchique des genres romanesques en fonction des ALR associés aux catégories sémantiques *action:déplacement* et *action:mouvement*

Cluster Dendrogram

- 49 Comme le fait apparaître le dendrogramme (Figure 12), les genres romanesques se regroupent, sur la base des ALR en question, en deux classes : les romans sentimentaux (SENT), policiers (POL) et de littérature générale (GEN) d'un côté, et les romans historiques (HIST), de fantasy (FY) et de science-fiction (SF) de l'autre.

5. Conclusion

- 50 Fondé sur une approche originale en linguistique et en stylistique outillées, le projet PhraseoRom a produit des résultats aussi bien sur le plan théorique qu'appliqué. Il a tout d'abord révélé la pertinence de la notion de motif textuel – inconnue jusqu'alors dans la tradition linguistique et stylistique anglosaxonne – pour caractériser et distinguer les genres littéraires. Il contribue ainsi à l'élaboration d'une « théorie opératoire des genres » (Rastier 2011 : 72). Ensuite, les outils créés dans le cadre du projet (Lexicoscope 2.0 et PhraseoLing) auront un impact dans le domaine de l'apprentissage automatique des textes (*deep learning*) et des nouvelles technologies pour la fouille des textes. Enfin, ses résultats, disponibles dans la base de données PhraseoBase, pourront trouver de multiples applications dans d'autres domaines, notamment en narratologie cognitive, en traductologie et dans des ateliers d'écriture créative.

BIBLIOGRAPHIE

- Adam J.-M. (2011). *Les Textes : types et prototypes : récit, description, argumentation, explication et dialogue*. Paris : Armand Colin.
- Augustinus L., Vandeghinste V. & Vanallemeersch T. (2016). « Poly-GrETEL : Cross-Lingual Example-based Querying of Syntactic Constructions ». In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (éd.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, 3549-3554.
- Barnier J. (2019). *Explor : Interactive Interfaces for Results Exploration*. R package version 0.3.5. <https://CRAN.R-project.org/package=explor>.
- Baroni R. (2015). « Temps, mode et intrigue : de la forme verbale à la fonction narrative ». *Modèles linguistiques* 71 : 125-142.
- Baroni R. (2007). *La tension narrative : suspense, curiosité et surprise*. Paris : Seuil.
- Diwersy S., Gonon L., Goossens V., Gymnich M. & Tutin A. (2020). « Speech Verbs in French and English Novels ». In I. Novakova & D. Siepmann (éd.), *Phraseology and Style in Subgenres of the Novel : a Synthesis of Corpus and Literary Perspectives*. Cham : Palgrave MacMillan, 83-113.
- Dunning T. (1993). « Accurate methods for the statistics of surprise and coincidence », *Computational Linguistics* 19(1) : 61-74.
- Fellbaum C. (éd.) (1998). « A Semantic Network of English : The Mother of All WordNets ». *Computers and the Humanities* 32 : 209-222.
- Gonon L., Goossens V. & Novakova I. (2020). « Les phraséologismes spécifiques à deux sous-genres de la paralittérature : le roman sentimental et le roman policier ». In S. Mejri, L. Meneses-Lerin & B. Buffard-Moret, *La phraséologie française en questions*. Paris : Hermann, 105-111.

- Goossens V., Jacquot C. & Dyka S. (2020). « Science Fiction versus Fantasy : A Semantic Categorization and its Contribution to Distinguishing Two Literary Genres ». In I. Novakova & D. Siepmann (éd.), *Phraseology and Style in Subgenres of the Novel : a Synthesis of Corpus and Literary Perspectives*. Cham : Palgrave MacMillan, 189-221.
- Grossmann F., Gymnich M. & Siepmann D. (2020). « Alcohol and Tobacco Consumption in English and French Novels since the 1950s : A Corpus-stylistic Analysis ». In I. Novakova & D. Siepmann (éd.), *Phraseology and Style in Subgenres of the Novel : a Synthesis of Corpus and Literary Perspectives*. Cham : Palgrave MacMillan, 114-150.
- Hoey M. (2005). *Lexical priming. A New Theory of Words and Language*. London/New York : Routledge.
- Jannidis F., Pielström S., Schöch C. & Vitt T. (2015). « Improving Burrows' Delta – An empirical evaluation of text distance measures ». In *Digital Humanities 2015 : Conference Abstracts*. Sydney : University of Western Sydney.
- Kassambara A. & Mundt F. (2020). *Factoextra : Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7. <https://CRAN.R-project.org/package=factoextra>.
- Kilgariff A., Rychly P., Smrz P. & Tugwell D. (2004). The Sketch Engine. *Proceedings of the Eleventh EURALEX International Congress*. Lorient, 105-116.
- 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 : 67-82.
- Kraif O. & Diwersy S. (2012). « Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques », *Actes de la conférence TALN 2012*. Grenoble, 399-406.
- Eder M., Rybicki J. & Kestemont M. (2016). « Stylometry with R : A Package for Computational Text Analysis ». *The R Journal* 8(1) : 107-121. doi : 10.32614/RJ-2016-007.
- Lê S., Josse J. & Husson F. (2008). « FactoMineR : An R Package for Multivariate Analysis ». *Journal of Statistical Software* 25(1) : 1-18.
- Lebart L. & Salem A. (1994). *Statistique textuelle*. Paris : Dunod.
- Legallois D. & Koch S. (2020). « The Notion of Motif Where Disciplines Intersect : Folkloristics, Narrativity, Bioinformatics, Automatic Text Processing and Linguistics ». In I. Novakova & D. Siepmann (éd.), *Phraseology and Style in Subgenres of the Novel : a Synthesis of Corpus and Literary Perspectives*. Cham : Palgrave MacMillan, 17-46.
- Legallois D. (2006). « Des phrases entre elles à l'unité réticulaire du texte ». *Langages* 163 : 56-70.
- Longrée D., Luong X. & Mellet S. (2008). « Les motifs : un outil pour la caractérisation topologique des textes ». In S. Heiden, B. Pincemin & L. Vosghanian, *Actes des JADT 2008 : 9^{es} journées internationales d'analyse statistique des données textuelles*, 733-744. Disponible en ligne [<http://www.jadt.org/>].
- 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.
- Martin R. (1983). *Pour une logique du sens*. Paris : PUF.
- Mahlberg M. (2007). « Clusters, Key Clusters and Local Textual Functions in Dickens ». *Corpora* 2(1) : 1-31.

- Mayaffre D., Pincemin B & Poudat C. (2019). « Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse du discours ». *Langue française* 203 : 101-115.
- McCarthy D., Kilgarriff A., Jakubicek M. & Reddy S. (2015). « Semantic word sketches ». Communication présentée à *Corpus Linguistics 2015*, Lancaster University, UK – 21st to 24th July 2015.
- Mellet S. & Longrée D. (2012). « Légitimité d'une unite textométrique : le motif ». In A. Dister, G. Purnelle & D. Longrée, *Actes des JADT 2012 : 11^{es} journées internationales d'analyse statistique des données textuelles*, 715-728. Disponible en ligne [<http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm>].
- Novakova I. & Siepmann D. (2020). « Literary Style, Corpus stylistic and Lexico-grammatical narrative patterns. Toward the concept of literary motifs ». In I. Novakova & D. Siepmann (éd.), *Phraseology and Style in Subgenres of the Novel : a Synthesis of Corpus and Literary Perspectives*. Cham : Palgrave MacMillan, 1-15.
- Quiniou S., Cellier P., Charnois T. & Legallois D. (2012). « Fouille de données pour la stylistique : cas des motifs séquentiels émergents ». In A. Dister, D. Longrée & G. Purnelle (éd.), *Actes des JADT 2012 : 11^{es} journées internationales d'analyse statistique des données textuelles* [<http://www.jadt.org/>, consulté le 04/10/2017].
- Rastier F. (2011). *La mesure et le grain. Sémantique de corpus*. Paris : Honoré Champion.
- Sievert C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.
- Sinclair J. (2004). *Trust the Text : Language, Corpus and Discourse*. London : Routledge.
- Wang I. (2017). *Syntactic Similarity Measures in Annotated Corpora for Language Learning : application to Korean Grammar*, Thèse de doctorat, sous la dir. de S. Kahane & I. Tellier, Université Paris 10.

NOTES

1. Le projet est consultable en ligne : <https://phraseorom.univ-grenoble-alpes.fr/>.
2. Pour un historique sur ce point, voir Mayaffre, Pincemin & Poudat (2019).
3. Le Lexicoscope est consultable en ligne : http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/.
4. La PhraseoBase est consultable en ligne : <http://phraseotext.univ-grenoble-alpes.fr/phraseobase/index.html>.
5. Pour le détail du nombre d'auteurs et d'œuvres dans chaque sous-genre, voir Novakova & Siepmann 2020 : 288.
6. C'est le seuil à partir duquel la surreprésentation de l'ALR dans un corpus peut être considérée comme statistiquement significative. Il s'agit d'un rapport de vraisemblance permettant d'évaluer objectivement si la répartition des unités linguistiques au sein d'un corpus est aléatoire ou non.
7. Cette grille a été créée dans la continuité des ontologies WordNet (Fellbaum 1998), des *functional groups clusters* de Mahlberg (2007) ou encore de McCarthy *et al.* (2015).
8. Pour une définition exhaustive des différentes dimensions et valeurs sémantiques retenues dans la grille sémantique, cf. le manuel d'annotation sémantique (en français et en anglais), consultable sur la PhraseoBase.
9. Pour consulter les fichiers complets des ALR annotés en français et en anglais, cf. <http://phraseotext.univ-grenoble-alpes.fr/phraseobase/ressources-fr.html>.

10. Par exemple, la requête pour l'ALR *<ils passèrent dans la rue>* se présente sous cette forme
`<l=passer,c=VERB,#1>&&<l=dans,c=PREP,#2>&&<l=le,c=DET,#3>&&<l=rue,c=NOUN,#4>::`
 (DETERM_DEF,4,3) (PREPOBJ,4,2) (VMOD_POSIT1,1,4).
11. Le regroupement automatique des 8415 ALR pour le français a été effectué à l'aide du logiciel Word2vec.
12. Le motif est une notion transversale utilisée dans de nombreux domaines comme la narratologie, les études folkloriques, la bio-informatique, le TAL, la stylistique, la linguistique (Legallois & Koch 2020).
13. Les fichiers complets des annotations stylistiques des données françaises et anglaises sont disponibles sur <http://phraseotext.univ-grenoble-alpes.fr/phraseobase/ressources-fr.html>.
14. Sur ce point, voir Grossmann, Gymnich & Siepmann (2020 : 127).
15. Ces fonctionnalités sont illustrées par des vidéos explicatives accessibles en ligne : http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/help.
16. Ces fonctionnalités étant coûteuses en ressources, elles ne sont accessibles qu'aux utilisateurs en ayant fait la demande explicite auprès de l'administrateur.
17. Le système de gestion choisi pour la base de données est MySQL et l'interface est programmée essentiellement en php.
18. Les calculs sont mis en œuvre à partir des valeurs de fréquence d'apparition par sous-corpus de genre littéraire enregistrées dans la base de données pour les ALR à la suite de leur extraction automatique.
19. Les calculs en question sont effectués au moyen de différents scripts R faisant appel aux packages FactoMineR (Lê, Josse & Husson 2008), explor (Barnier 2019), stylo (Eder, Rybicki & Kestemont 2016), factoextra (Kassambara & Mundt 2020) et plotly (Sievert 2020).

RÉSUMÉS

Cet article présente les ressources développées dans le cadre du projet franco-allemand ANR DFG PhraseoRom (2016-2020). Son objectif était de caractériser la phraséologie du genre romanesque contemporain dans trois langues (français, anglais, allemand). Nous détaillons ici les choix ayant présidé à la constitution des corpus littéraires trilingues, ainsi qu'à la méthodologie inédite d'extraction des données et de leur analyse linguistique et stylistique. Nous présentons enfin les fonctionnalités de la PhraseoBase, interrogeable avec l'outil Lexicoscope 2.0, et celles de l'interface PhraseoLing. Ces ressources sont librement consultables sur <http://phraseotext.univ-grenoble-alpes.fr/phraseobase/index-en.html>

This article presents the resources developed by the Franco-German project ANR DFG PhraseoRom (2016-2020). The main goal of the project was to characterize the phraseology of the contemporary fiction in French, English, and German. We detail here the choices that led to the constitution of a trilingual literary corpora, as well as the new methodology of the data extraction and of the linguistic and stylistic analysis. Finally, we present the functionalities of the PhraseoBase, with the Lexicoscope 2.0, and the PhraseoLing interface. These resources are freely available online <http://phraseotext.univ-grenoble-alpes.fr/phraseobase/index-en.html>

INDEX

Keywords : corpus linguistics, NLP, digital humanities, corpus stylistics, contemporary literary genres, extended phraseological units, motifs

Mots-clés : linguistique de corpus, TAL, stylistique outillée, genres romanesques contemporains, unités phraséologiques étendues, motifs

AUTEURS

SASCHA DIWERSY

Université Montpellier 3 Paul Valéry, Praxiling UMR 5267

LAETITIA GONON

Université de Rouen Normandie, CÉRÉdI

VANNINA GOOSSENS

Université Grenoble Alpes, LIDILEM

OLIVIER KRAIF

Université Grenoble Alpes, LIDILEM

IVA NOVAKOVA

Université Grenoble Alpes, LIDILEM

JULIE SORBA

Université Grenoble Alpes, LIDILEM & Litt&Arts UMR 5316

ILARIA VIDOTTO

Université de Lausanne, Section de français