Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction

Olivier Kraif, Iva Novakova, Julie Sorba

Lidil
Revue de linguistique et de didactique des langues

53 | 2016
Phraséologie et genres de discours

Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction

Specific Phraseological Units in Mystery Novels and Science-Fiction

Olivier Kraif, Iva Novakova et Julie Sorba

Édition électronique
URL : http://journals.openedition.org/lidil/3976
DOI : 10.4000/lidil.3976
ISSN : 1960-6052

Éditeur
UGA Éditions/Université Grenoble Alpes

Édition imprimée
Date de publication : 30 mai 2016
Pagination : 143-159
ISBN : 978-2-84310-326-1
ISSN : 1146-6480

Référence électronique
Olivier Kraif, Iva Novakova et Julie Sorba, « Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction », *Lidil* [En ligne], 53 | 2016, mis en ligne le 01 janvier 2017, consulté le 01 mai 2019. URL : http://journals.openedition.org/lidil/3976 ; DOI : 10.4000/lidil.3976

Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction

*Olivier Kraif, Iva Novakova et Julie Sorba**

Résumé

Cet article expose les résultats d'une étude préliminaire menée dans le cadre du projet PHRASEOTEXT qui vise à articuler les niveaux phraséologiques, stylistiques et discursifs à travers l'analyse comparée de corpus représentants différents sous-genres littéraires. Son objectif est à la fois méthodologique, avec une approche originale s'appuyant sur les principes de la linguistique de corpus et sur les méthodes quantitatives issus du TAL, et descriptif, pour identifier et interpréter les caractéristiques phraséologiques spécifiques à chaque sous-genre étudié. Dans cette étude exploratoire, nous nous appuyons sur un corpus de romans policiers et de textes de science-fiction.

Abstract

In this paper, we present the preliminary results of a research conducted as part of the PHRASEOTEXT project. The objectives of this project are to address the interrelations of phraseological, stylistic and discursive levels through the comparative analysis of corpora that represent various literary subgenres. Our goal is twofold, both methodological and descriptive: first, by developing an original approach combining corpus linguistics and quantitative methods from NLP; secondly, by spotting and interpreting, from a linguistic and literary point of view, the phraseological peculiarities of each subcorpus. In this exploratory study, our observations are based on two corpora of mystery novels and science fiction texts.

* LIDILEM, Université Grenoble Alpes.

1. Introduction

Nous livrons ici les premiers résultats d'une étude pilote menée dans le cadre du projet PHRASEOTEXT[1], lequel a pour objectif l'étude comparative de la phraséologie stéréotypée à partir d'un corpus de textes littéraires contemporains en langue française. Notre travail vise à mettre en lumière la pertinence des phénomènes lexico-grammaticaux pour l'identification des sous-genres littéraires[2]. Pour cela, nous proposons d'analyser les constructions lexico-syntaxiques (CLS) spécifiques dans le roman policier (POL) et dans la science-fiction (SF) en appliquant une méthodologie innovante, élaborée pour extraire et analyser les objets phraséologiques afin de mieux caractériser leur structure et leur statut. Plus généralement, nous proposons d'articuler des critères locaux (lexique, syntaxe) et des critères globaux « portant sur le genre du texte, le discours dont il relève, le corpus où il prend sens » (Rastier, 2011, p. 32). En suivant l'hypothèse de Siepmann (2015 et à paraitre) selon laquelle le langage littéraire se fonde, au moins en partie, sur l'emploi statistiquement significatif de mots et de phraséologismes hautement spécifiques[3], nous avons supposé qu'il existe des CLS récurrentes et donc spécifiques aux sous-genres étudiés.

Notre étude s'inscrit dans le sillage de Longrée et Mellet (2013) qui s'intéressent à des séquences récurrentes, nommées *motifs* : « Sur le plan fonctionnel, le motif est un "cadre collocationnel" accueillant un ensemble d'éléments fixes et variables susceptibles d'accompagner la structuration textuelle, et simultanément, de caractériser des textes de genres divers. » (p. 66) Ces motifs peuvent être également considérés comme des « unités multidimensionnelles », constituées à la fois d'associations lexicales et grammaticales, d'appariements entre forme et sens, ou entre fonction pragmatique et discursive (Legallois, 2012, p. 45). Néanmoins, à la différence des auteurs cités, nous nous intéressons à des structures hiérarchiques (ou arbres de dépendance), lesquelles ne correspondent pas forcément à des séquences linéaires (voir Tutin & Kraif, dans ce volume).

---

1. Voir <http://phraseotext.u-grenoble3.fr/lexicoscope/> (projet AGIR POLE 2015-2016).
2. Pour les critères servant à catégoriser les sous-genres, voir section 2.
3. Siepmann (à paraitre) montre que certains types de séquence comme « Il en était là de ses réflexions quand… » sont spécifiques à la composition littéraire dans le roman contemporain.

Nous proposons une analyse des données fondée sur des modèles fonctionnels et contextualistes (voir Sinclair, 2004 ; Hoey, 2005) qui explorent systématiquement quatre niveaux d'analyse (lexical, sémantique, syntaxique et discursif). Notre approche globale permet ainsi de réunir un faisceau de traits lexico-syntaxiques susceptible de caractériser les genres textuels. Après avoir explicité nos principes méthodologiques (section 2) et décrit le corpus de l'étude ainsi que les extractions automatiques d'unités phraséologiques effectuées (section 3), nous procéderons à l'analyse des premiers résultats obtenus dans le sous-genre de la science-fiction (section 4) et, ensuite, de ceux recueillis dans le sous-genre du roman policier (section 5).

2. Principes méthodologiques

Notre méthodologie est d'abord inductive : nos observations ont été essentiellement guidées par les données, sans partir d'un cadre théorique nous permettant de définir et de caractériser les genres littéraires étudiés à priori. Ainsi, la classification des œuvres en sous-genre a été guidée par des considérations éditoriales. En effet, lors de la constitution du corpus, la distinction entre les deux sous-genres POL et SF s'est appuyée sur la classification adoptée dans les catalogues des éditeurs et dans les rayons des librairies : les œuvres sont cataloguées dans tel ou tel genre, avec un certain consensus, en général corroboré par leur édition au sein d'une collection spécialisée.

Pour l'étude linguistique, nous nous situons donc dans le cadre d'une approche *corpus-driven* (Biber, 2009 ; Sinclair, 1991), plutôt que *corpus-based* (Tognini Bonelli, 2001) : ce choix méthodologique ne signifie pas que nous rejetons une approche hypothético-déductive où le corpus servirait à étayer ou à rejeter des hypothèses émises dans le cadre d'une théorie linguistique ou herméneutique. D'un point de vue heuristique, les outils numériques originaux que nous mettons en œuvre sont à même de faire surgir des régularités et des phénomènes inattendus. Une fois ces derniers mis au jour, nous avons besoin d'un cadre interprétatif pour tenter de les comprendre et de les expliquer. Procéder en sens inverse risquerait de nous faire passer à côté de certaines observations ne cadrant pas avec les hypothèses de départ.

Un second choix méthodologique consiste à s'appuyer principalement sur la syntaxe pour l'identification des CLS récurrentes. Les analyseurs syntaxiques en dépendance sont, selon nous, arrivés à un degré de précision suffisant pour autoriser des études sur corpus à grande

échelle. Certes, même les meilleurs parseurs font des erreurs, qu'il s'agisse d'erreurs de rattachement ou d'erreurs d'étiquette de relation[4], lesquelles sont susceptibles d'affecter les analyses. D'autre part, même avec des résultats fiables à 100 %, on pourrait reprocher à un analyseur de biaiser nécessairement les observations. Comme le notent Hunston et Francis (2000), « *If a corpus is annotated in any way, the annotation will reflect a particular theory of grammar, and the results will naturally be cast in terms of that theory* » (p. 19) ; ce constat rappelle une remarque en forme d'adage de Sinclair (1987) : « *The more superficial, the better* » (p. 107).

Pour répondre à ces objections, nous considérons qu'un parseur fournit certes un certain pourcentage de bruit dans ces analyses, mais que ce bruit se dissipe lorsqu'on s'intéresse à des phénomènes récurrents correspondant à des seuils significatifs d'association statistique. En effet, lorsqu'on examine par exemple des cooccurrents de surface[5], dans une fenêtre de plus ou moins 3 mots, on a aussi affaire à de très nombreuses cooccurrences contingentes, fortuites, et assimilables à du bruit. Enfin, il est évident que les relations de dépendance les plus fréquentes observées dans le corpus sont intrinsèquement liées au modèle syntaxique employé : certains analyseurs s'attachent à identifier des dépendances plus profondes au plan sémantique (p. ex. l'objet profond d'un verbe au passif), tandis que d'autres extraient des dépendances de surface (p. ex. le sujet rattaché à l'auxiliaire *avoir* plutôt qu'au participe). Les relations statistiquement pertinentes dépendront logiquement du modèle appliqué. L'important est de ne jamais le perdre de vue et de faire les contrôles nécessaires afin de s'assurer que tel phénomène saillant n'est pas un artefact lié au modèle employé. Un outil d'annotation syntaxique joue pour nous le rôle d'un microscope : bien qu'il ne soit pas exempt de biais, nous restons convaincus qu'il nous aide à identifier des phénomènes qui resteraient invisibles autrement.

---

4. Par exemple, sont recensés environ 10 % de rattachements erronés pour des langues comme l'anglais ou le catalan, lors de la campagne CoNLL 2007, voir Nivre et coll. (2007).
5. Une cooccurrence de surface désigne, pour des mots, le fait d'apparaitre dans le même voisinage, au sein d'une fenêtre de largeur fixe ; cela n'impose pas qu'il y ait une relation fonctionnelle entre les mots cooccurrents : par exemple, dans « des poissons aux longues nageoires de cristal nous sourient » (Werber), les mots *longues* et *cristal* sont des cooccurrents de surface, mais sans lien syntaxique direct.

Par ailleurs, quand c'est possible, on pourra recourir à des analyseurs différents pour trianguler les observations et consolider les hypothèses émises.

Pour l'identification des CLS, nous avons utilisé la technique d'extraction d'arbres lexico-syntaxiques récurrents (ALR)[6], qui permet d'extraire des arbres dont la récurrence est significative sur un plan statistique. De la sorte, on extrait des ALR tels que celui de la figure 1, correspondant à la CLS *se passer la langue sur les lèvres* :

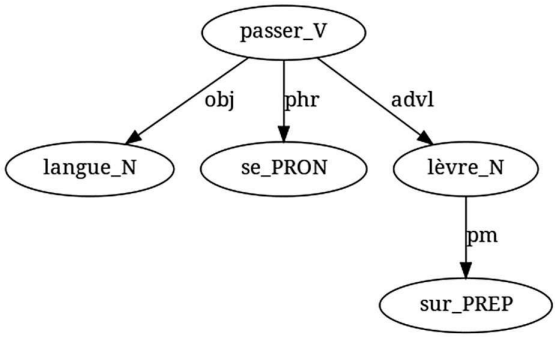

Figure 1. – Arbre lexico-syntaxique récurrent correspondant à l'expression *se passer la langue sur les lèvres*.

Une fois qu'un tel ALR a été identifié, on cherche à déterminer s'il peut être considéré comme spécifique de tel ou tel sous-corpus. Par exemple, l'expression *se passer la langue sur les lèvres* apparait 12 fois dans le POL, 1 fois dans la SF et 3 fois dans les autres sous-corpus. Il semble donc qu'elle soit spécifique au sous-genre POL. Comment s'assurer que cette répartition déséquilibrée n'est pas due au hasard ? Le calcul du rapport de vraisemblance (noté LLR pour *Log Likelihood Ratio*) permet d'évaluer le caractère improbable d'une telle répartition aléatoire : on l'utilise couramment comme mesure de spécificité — c'est le cas par exemple du logiciel *WordSmith*, avec des performances comparables au calcul plus coûteux issu du modèle hypergéométrique (Bertels & Speelman, 2013). Ce calcul se fonde sur un tableau de contingence faisant intervenir les 4 grandeurs suivantes (Dunning,

6. Pour plus de détails, voir l'article de Tutin & Kraif dans ce volume.

2013) : *f*, la fréquence dans le sous-corpus *c* (POL ou SF) ; *F*, la fréquence dans l'ensemble du corpus *C* (LIT_MOD) ; *t*, le nombre total de mots du sous-corpus (POL ou SF) ; et *T*, le nombre total de mots du corpus (LIT_MOD)[7].

À l'instar de Quiniou et coll. (2012), on retient donc des structures « émergentes », dont la fréquence relative dans un sous-ensemble du corpus est significativement supérieure à la fréquence dans un autre sous-ensemble. Mais à la différence de ces auteurs, les objets extraits correspondent à des structures hiérarchiques et non à des motifs séquentiels. Ainsi, même si la plupart des motifs observés dans notre étude sont séquentiels, ce n'est pas pour autant une contrainte préalable dans notre méthodologie.

3. Corpus d'étude et extractions automatiques

Les données sont issues d'un corpus constitué dans le cadre du projet Emolex[8]. Pour cette étude, nous avons seulement retenu les textes littéraires français contemporains (LIT_MOD, 16 millions de mots) que nous avons répartis en six sous-ensembles définis par les sous-genres littéraires auxquels chaque texte appartient. Nous avons effectué une extraction des ALR sur l'ensemble du corpus LIT_MOD, puis sur les deux sous-corpus retenus pour notre étude : le POL (4 millions de mots, 14 auteurs[9], 35 ouvrages) et la SF (2,4 millions de mots, 13 auteurs[10],

7. Son calcul intègre quatre composantes pour les valeurs du tableau de contingence correspondant au nombre de fois que la forme est tirée (ou qu'une autre forme est tirée) dans le sous-corpus *c* et dans son complémentaire dans *C* :

$$\log Like = 2\cdot(f\cdot\log(f\cdot\frac{T}{t\cdot F})+(F-f)\cdot\log((F-f)\cdot\frac{T}{(T-t)\cdot F})+(t-f)\cdot\log((t-f)\cdot\frac{T}{t\cdot(T-F)})+(T-t-(F-f))\cdot\log((T-t-(F-f))\cdot\frac{T}{(T-t)\cdot(T-F)}))$$

8. Le corpus a été annoté syntaxiquement avec le parseur *Connexor* (Tapanainen & Järvinen, 1997). Pour plus de détails, voir la présentation méthodologique de l'*EmoBase* (<http://emolex.u-grenoble3.fr/emoBase>) dans Diwersy et coll. (2014).
9. Les auteurs du sous-corpus POL sont, par ordre alphabétique, Brigitte Aubert, Cédric Bannel, Serge Brussolo, Maxime Chattam, Frédéric Dard, Ken Follett, Jean-Christophe Grangé, John Grisham, Marek Halter, Pierre Lemaître, Daniel Macouin, Jean-François Parot, Georges Simenon, Fred Vargas.
10. Les auteurs du sous-corpus SF sont, par ordre alphabétique, Alcide Dario, René Barjavel, Pierre Boulle, B. R. Bruss, Mathieu Gaborit, Paul-Jean

27 ouvrages). Le tableau 1 montre un échantillon des ALR spécifiques au POL, à titre d'illustration de notre méthodologie. Le calcul des spécificités fait donc intervenir les fréquences des ALR dans les sous-corpus (notée *f1*), comparées aux fréquences dans le corpus pris dans son intégralité (notée *f2*). Les arbres retenus comportant 8 nœuds au maximum ont une fréquence au moins égale à 3 et un score de spécificité LLR ≥ 3,84 (ce qui correspond à une probabilité inférieure à 0,05 d'avoir une distribution aléatoire ; le LLR est indiqué dans la colonne 5). De plus, le tableau 1 inclut le critère de *dispersion* (noté *Disp*.) qui indique le nombre de sous-corpus dans lesquels l'ALR est présent. Nous avons retenu les ALR qui présentaient une dispersion égale ou supérieure à 3[11]. L'extraction automatique nous a permis ainsi de relever 3 186 ALR spécifiques au sous-corpus POL, et 3 496 ALR spécifiques au sous-corpus SF[12].

Dans ces extractions, la tendance générale, aussi bien dans le POL que dans la SF, est la surreprésentation de certains pivots nominaux, qui agrègent autour d'eux de nombreux ALR. Néanmoins, la proportion entre pivots nominaux et pivots verbaux est différente dans les deux sous-corpus. En effet, sur les 50 premières constructions les plus spécifiques (*LLR* > 100), on dénombre seulement 2 ALR avec pivot verbal dans la SF, alors que dans le POL, on retrouve 8 CLS avec ce même type de pivot (voir tableau 2).

Hérault, Robin Hobb, Michel Jeury, Pierre Pelot, Anne Robillard, Magali Ségura, Roland C. Wagner, Bernard Werber.

11. Quand la *dispersion* est au moins égale à 3, cela signifie que l'ALR apparait dans au moins 3 sous-corpus sur 6.
12. En examinant les premiers ALR pour le sous-corpus SF, nous avons dû écarter certains d'entre eux. En effet, malgré une dispersion égale ou supérieure à 3 (c'est-à-dire que l'ALR est attesté dans au moins 3 sous-corpus), certains ALR étaient propres aux œuvres d'un auteur (par exemple dans la SF, *fourmi* et de *horde* sont propres aux œuvres de B. Werber, *elfes* est uniquement attesté dans celles de A. Robillard, tandis que *charogne* est exclusivement utilisé comme nom propre pour désigner le domaine des morts-vivants dans les *Chroniques des Feals* de M. Gaborit). L'accroissement du sous-corpus SF et l'ajout d'auteurs différents permettront d'atténuer ce biais de la surreprésentation d'un auteur.

Arbres lexico-syntaxiques récurrents (ALR)	f1	f2	Disp.	LLR
<l=crime,c=N,#1>&&<l=scène,c=N,#2>&&<l=de,c=PREP,#3>::(mod,2,1) (pm,2,3)	161	165	3	408,05
<l=un,c=DET,#1>&&<l=regard,c=N,#2>&&<l=lancer,c=V,#3>::(det,2,1) (obj,3,2)	121	158	6	182,75
<l=un,c=DET,#1>&&<l=coup,c=N,#2>&&<l=jeter,c=V,#3>::(det,2,1) (obj,3,2)	207	371	6	156,13
<l=mot,c=N,#1>&&<l=sans,c=PREP,#2>&&<l=(?:un|une|des|d|de),c=DET,#3>::(det,1,3) (pm,1,2)	142	248	6	114,15
<l=jeune,c=A,#1>&&<l=inspecteur,c=N,#2>::(attr,2,1)	42	47	4	86,72
<l=quai,c=N,#1>&&<l=orfèvre,c=N,#2>::(mod,1,2)	38	47	5	63,98
<l=tasse,c=N,#1>&&<l=café,c=N,#2>::(mod,1,2)	65	112	5	53,98
<l=crâne,c=N,#1>&&<l=son,c=PRON,#2>::(det,1,2)	135	304	6	52,45
<l=vitesse,c=N,#1>&&<l=à,c=PREP,#2>&&<l=tout,c=PRON,#3>::(det,1,3) (pm,1,2)	78	151	6	48,1
<l=meurtre,c=N,#1>&&<l=série,c=N,#2>::(mod,2,1)	22	24	3	47,98
<l=cadrer,c=V,#1>&&<l=pas,c=ADV,#2>&&<l=ne,c=ADV,#3>::(neg,1,2) (ad,2,3)	25	30	3	44,72
<l=mise,c=N,#1>&&<l=scène,c=N,#2>::(mod,1,2)	70	135	6	43,61
<l=chance,c=N,#1>&&<l=tomber,c=V,#2>&&<l=de,c=PREP,#3>::(pm,1,3) (mod,1,2)	10	5	3	41,4
<l=point,c=N,#1>&&<l=suspension,c=N,#2>&&<l=de,c=PREP,#3>::(pm,1,3) (mod,1,2)	22	30	3	30,44
<l=instant,c=N,#1>&&<l=à,c=PREP,#2>&&<l=même,c=A,#3>::(attr,1,3) (pm,1,2)	42	77	5	29,91
…	…	…	…	…

Tableau 1. – ALR spécifiques au sous-corpus POL (LLR >= 3,84).

CLS	POL	SF
à pivot nominal	34/50 (*crime*, *meurtre*, *téléphone*, *sang*, *coup*, *inspecteur*, *lieutenant*, *cadavre*, etc.)	43/50 (*fourmis*, *sol*, *horde*, *charogne*, *onde*, *combat*, *adversaire*, *cristal*, etc.)
à pivot verbal	8/50 (*lancer*, *jeter*, *comprendre*, *hocher*, *se pencher*, *border*, *devoir*, *coller*)	2/50 (*s'enfoncer*, *se retrouver*)

Tableau 2. – Les pivots les plus spécifiques dans les sous-corpus POL et SF (LLR > 100).

De prime abord, on peut supposer que la surreprésentation de certains pivots nominaux est liée à des effets de saillance thématique. De manière attendue, certains thèmes sont caractéristiques des sous-genres. Dans le sous-corpus POL, les ALR les plus fréquents gravitent autour de personnages (*inspecteur de police*, *flic*, *légiste*…), de lieux (*quai des Orfèvres*, *laboratoire de police*…), ou d'actes déclencheurs de la diégèse dans le roman policier (*crime*, *meurtre*, *violence*…). Dans le sous-corpus SF, les ALR les plus fréquents relèvent de quatre champs lexicaux distincts, comme les éléments naturels (*sol*, *soleil*, *ciel*, *cristal*…), la guerre (*adversaire*, *combat*…), les créatures (*fourmi*, *ange*, *singe*, *humain*, *elfe*, *fée*…) et les lieux (*continent*, *monde*, *royaume*, *cité*, *forêt*…), qui posent les ingrédients du décor et des univers créés. Ces éléments constituent des *récurrences thématiques*.

Lors de l'observation des données, un autre résultat est apparu. En effet, certains ALR ne reposent pas sur des récurrences de mots en lien avec des thématiques propres à ces sous-genres : par exemple, l'expression *lancer un regard* ne ressortit pas, de prime abord, à la thématique du roman policier. Or, elle n'en demeure pas moins statistiquement spécifique. Les séquences de ce type constituent des *récurrences stylistiques*. Dans les sections qui suivent, nous présenterons les premiers résultats pour chacun des deux sous-corpus SF (section 4) et POL (section 5) en sélectionnant une récurrence thématique et une récurrence stylistique.

4. Premiers résultats dans le sous-genre science-fiction

4.1. Récurrence thématique dans SF : de cristal

L'extraction automatique a fourni une première séquence *de cristal*[13] qui présente un LLR élevé (173,9). Sur les 165 occurrences du corpus LIT_MOD (soit 61 %, réparties chez 9 auteurs), 101 apparaissent dans le sous-corpus SF. En procédant à une extension syntagmatique à gauche,

13. L'association préférentielle de *cristal* avec la préposition *de* correspond à une *collocation grammaticale*. Comme l'indique Bolly (2010) la « collocation grammaticale (vs lexicale) désigne en particulier une combinaison lexicalement contrainte constituée d'un mot lexical (de contenu) qui sélectionne de manière arbitraire un mot grammatical (mot outil), généralement une préposition […] » (p. 16-17). Voir aussi, à ce sujet, Legallois (2012, p. 42).

nous avons trouvé une structure récurrente SN + *de cristal*[14]. Dans le sous-corpus SF, ce SN précédant *de cristal* présente quelques particularités par rapport aux autres sous-corpus. Le tableau 2 ci-dessous répertorie les SN à gauche *de cristal* classés selon leur contenu référentiel :

	Dans le sous-corpus SF	Dans les 5 autres sous-corpus
objets matériels	*barres*, *rideau*, *colonnes*, *pendule*, *objets*, *table*, *anneau*, *murs*, *filin*, *morfil*, *poignard*	*camées*, *sculpture*, *voile*, *vase*, *flacon*, *pendeloques*, *boite*, *gouttes*, *lustre*, *parois*, *seau*, *coffret*, *cruche*, *loupe*, *carafon*, *compotier*, *ballon*, *plaque*…
parties du corps	*nageoires*	*cicatrices*, *cœur*, *œil*, *iris*
lieu	*prison*, *caverne*, *sépulcre*, *montagne*, *royaume*	*forêt*, *buissons*, *palais*, *émergence*, *archives*
appel aux sens	*bruit*, *points*, *souffle*, *arpèges*	*tintement*.s
propriétés	ø	*pureté*, *clarté*, *netteté*

Tableau 3. – SN + *de cristal* : associations lexicales spécifiques.

Dans la SF, le SN à gauche appartient à un fonds lexical connu du lecteur : il s'agit d'un vocabulaire courant, ne présentant pas de difficulté de compréhension. Mais sa combinaison avec le complément *de cristal* crée un syntagme original provoquant une impression d'étrangeté dans l'ensemble des associations lexicales (hormis peut-être une *table* ou *objets de cristal* qui peuvent exister dans le monde réel) :

(1) Dans les lacs transparents, des poissons aux longues *nageoires de cristal* nous sourient. (Werber, SF)

(2) En guise de réponse, les Pégasins dégainèrent leurs *poignards de cristal* et vinrent encadrer l'échevin. (Gaborit, SF)

En revanche, dans les autres sous-corpus, c'est la séquence toute entière SN + *de cristal* qui appartient au fonds lexical connu du lecteur et désigne des objets du quotidien (3, *compotier de cristal*), même de manière poétique (4, *palais de cristal* comme métaphore des bâtiments recouverts de givre ou encore *cicatrices de cristal* comme métaphore des traces laissées par différentes sécrétions gelées sur un cadavre) :

14. Nous avons écarté les deux séquences *Royaume de Cristal* et *Magicien de Cristal* dans lesquelles *Cristal* est employé comme un toponyme.

(3) L'oncle Jules qui venait d'entrer était si rouge de fierté, que Paul, écœuré, m'entraîna dans la salle à manger, où nous dégustâmes les quatre bananes qu'il avait repérées au passage dans le *compotier de cristal*. (Pagnol, AUT)

(4) Vers la ville, deux immenses bâtiments symétriques sortaient de terre. Les échafaudages de bois, couverts de givre, leur donnaient l'aspect d'éphémères *palais de cristal*. (Parot, POL)

Ainsi, nous pouvons dégager une tendance spécifique à la séquence SN + *de cristal* dans le sous-genre SF : elle permet de fournir les ingrédients de la « xéno-encyclopédie[15] » propre à la SF : « 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. » (Bozzeto, 2007, p. 60-61) Il s'agit alors d'une caractéristique interne au genre SF.

4.2. Récurrence stylistique dans SF : SN + trop + *ADJ*

L'extraction automatique a fourni la séquence SN + *trop* + ADJ (principalement de *couleur*) *ciel trop bas*/*bleu*/*noir* ou *soleil trop blanc*/*chaud*/*âgé*/*étroit*, spécifique au sous-corpus SF (6 occ. SF / 9 occ. LIT_MOD) :

(5) Les *soleils trop âgés* ne sont pas assez puissants. (Werber, SF)

Dans le cadre de notre démarche *corpus-driven*, nous avons procédé à une expansion paradigmatique de ce patron lexico-syntaxique pour extraire les occurrences correspondant à SN + *trop* + ADJ *couleur*. Il s'est avéré que dans le POL, le SN désigne majoritairement une partie du visage (*peau*, *yeux*, *bouche*, *lèvre*, *paupière*, *cheveux*, *mains*, *dents*, *joues*) (11/17 occ.), alors que dans la SF, il désigne systématiquement un élément ou un objet naturel (6/6 occ.) :

(6) Fascinant, dit-il, en léchant ses *lèvres trop rouges*. (Aubert, POL)

(7) Ce paysage trop calme et trop accueillant suintait l'angoisse par ses *pierres trop blanches*, ses *feuilles trop vertes*, son *ciel trop bleu*. (Jeury, SF)

---

15. Voir Bozzetto (2007) : « La SF traite d'objets, de personnages, de lieux, de moyens d'action, d'une xéno-encyclopédie, qui ne font pas partie de l'encyclopédie constituant la représentation du monde du lecteur. » (p. 56)

Il s'agit ici de *récurrences stylistiques* aussi bien dans la SF que dans le POL, avec toutefois des variations sur le SN, spécifiques à chaque sous-genre. On pourrait associer ces récurrences à des *patrons* narratifs, définis comme des cooccurrences récurrentes (collocation, séquence figée, motif) dans un corpus de textes narratifs, qui peuvent avoir une fonction narrative et/ou descriptive spécifique (cf. Siepmann 2015 et à paraitre). Dans notre cas, ces patrons correspondent à des schémas descriptifs. Ils sont indispensables à la participation passive et automatique du lecteur à l'univers référentiel du texte (Dufays, 2010 ; Siepmann, 2015).

Pour résumer, les résultats obtenus permettent de faire émerger deux CLS caractéristiques du sous-corpus SF qui connaissent toutes deux des variations paradigmatiques autour d'une structure syntaxique fixe : SN + *de cristal* correspond à une *récurrence thématique* propre à la SF en participant à la construction de la « xéno-encyclopédie » de cet univers si spécifique, tandis que la séquence SN + *trop* + ADJ *couleur* représente une *récurrence stylistique* (*motif* de l'excès exprimant l'a-normal), ayant une fonction descriptive dans les deux sous-corpus SF et POL au moyen de choix lexicaux distincts.

5. Premiers résultats dans le sous-genre roman policier

Les résultats obtenus pour le sous-corpus POL ont fait apparaitre des récurrences aussi bien thématiques que stylistiques.

5.1. Récurrence thématique dans POL : scène de crime

De manière peu surprenante, la séquence *scène de crime* apparait comme la plus spécifique au sous-corpus POL (LLR 1155 ; 116 occ. sur les 119 de LIT_MOD). Elle correspond à une *récurrence thématique* propre à ce sous-genre[16]. Ce degré élevé de spécificité révèle l'attraction mutuelle préférentielle entre *scène* et *crime*, ce qui n'est pas le cas de *meurtre*, synonyme de *crime* ; en effet, *meurtre* a une aversion (Hoey, 2005) pour le collocatif *scène* (qui n'apparait que 3 fois dans le corpus) et se combine préférentiellement avec *série* (37 occ. pour *série de meurtres* contre 7 occ. uniquement pour *série de crimes*).

16. Sur l'histoire de l'apparition de la séquence *scène de crime* dans le thriller contemporain, voir Gonon et coll. (à paraitre).

La séquence *scène de crime* correspond ici à une collocation[17]. L'observation des contextes révèle aussi l'existence d'une collocation grammaticale (voir la note 13) spécifique qui associe, de manière privilégiée, sur l'axe syntagmatique la séquence *scène de crime* précédée de son déterminant (*une*, *la*, *les*) à la préposition *sur* (36 occurrences sur les 116 de *scène de crime*, LLR 179) :

(8) Et *sur la scène de crime* de l'Indien, boulevard Malesherbes, tu en as trouvé aussi ? (Chattam, POL).

À partir de ce patron lexico-syntaxique PREP *sur* + DET *une*, *la*, *les* + *scène.s de crime*, nous avons procédé à une extension syntagmatique à gauche. Nous avons pu alors remarquer une préférence pour une association régulière de ce patron lexico-syntaxique avec des verbes de mouvement (*aller*, *se rapprocher*, *partir*, *arriver*, *revenir*) :

(9) Si vous faites appel à quelqu'un pour *aller sur les scènes de crime*, je voudrais que ce soit moi. (Chattam, POL)

Cette séquence ainsi élargie [Vmouv + PREP (*sur*) + DET (*une*, *la*, *les*) + *scène.s de crime*] représente le *motif thématique de la localisation*. Dans ce cas, *scène de crime* apparait en fonction de complément du Vmouv, ce qui correspond structurellement à une colligation[18]. La fonction discursive de ce motif de la localisation (*aller sur la scène de crime*) est d'introduire l'élément initial de la diégèse : en effet, se rendre sur la scène de crime constitue le premier jalon vers la résolution de l'énigme (voir Gonon et coll., 2016).

5.2. Récurrence stylistique dans le POL : lancer un regard + *ADJ* + à

De manière plus surprenante, nous avons relevé la grande spécificité statistique de la structure *lancer un regard* + ADJ (*circulaire*, *incertain*, *interrogatif*, *mauvais*, *noir*) + PREP (*vers*, *autour de*). En regardant de près la dispersion de ce patron lexico-syntaxique par sous-genres, il

17. La *collocation*, au sens de Stubbs (1995) ou de Sinclair (2004), est conçue comme une cooccurrence statistiquement significative d'items lexicaux.
18. À la suite de Hoey (2005, p. 44), nous définissons la *colligation* comme une cooccurrence de phénomènes grammaticaux sur l'axe syntagmatique qui révèlent la préférence des lexies pour certains environnements grammaticaux et, de là, pour certaines fonctions grammaticales.

s'avère que sur les 93 occurrences relevées dans l'ensemble du corpus littéraire (LIT_MOD), 68 apparaissent dans le POL[19] :

(10) Barnes *lança un regard incertain* autour de lui. (Grangé, POL)

On trouve aussi de manière récurrente une expansion syntagmatique de ce patron à un tiers-actant *lancer un regard* ADJ *à* N :

(11) Il *lance à Oscar Avane un long regard* d'épagneul pour le remercier de son pieux mensonge. (Dard, POL)

Cette variante avec le tiers actant se révèle d'une grande spécificité générique car sur les 20 occurrences de ce patron dans LIT_MOD, 16 se rencontrent dans le POL contre 4 seulement dans la SF. Il semblerait alors que cette récurrence stylistique puisse s'apparenter à un *motif* lié au comportement des personnages (celui du « regard lancé »), propre au sous-corpus POL. D'un point de vue discursif, à la différence du motif de l'excès exprimant de l'a-normal (voir section 4.2) qui entre dans un schéma descriptif, le motif du « regard lancé » renvoie aux comportements courants dans le cadre de l'enquête, induits par l'atmosphère de tension qui règne dans le roman policier (le criminel nerveux, l'inspecteur procédant à un interrogatoire, etc.).

6. Conclusion et perspectives

Ainsi, notre méthodologie a permis de faire émerger des phénomènes de deux ordres. Dans la SF, les ALR spécifiques représentant des *récurrences thématiques* sont principalement des syntagmes nominaux à fonction référentielle pure (*de cristal*), alors que dans le POL, les séquences récurrentes semblent davantage avoir une fonction discursive et produire, de ce fait, des motifs (par exemple, le motif de la localisation). Quant aux *récurrences stylistiques*, les ALR du type SN + *trop* + ADJ *couleur* sont spécifiques aux deux sous-corpus POL et SF (*motif* de l'excès exprimant l'a-normal), mais dans la SF, le SN renvoie à des phénomènes naturels, tandis que dans le POL, le SN dénote des parties du corps. Enfin, la *récurrence stylistique* hautement spécifique relevée dans le POL (*lancer un regard* + ADJ + *à*) constitue un motif lié au comportement des personnages. Toutes ces observations tendent à

19. Les autres occurrences se répartissent entre les sous-corpus SF (11 occ.), roman sentimental (2 occ.), roman historique (2 occ.) et autres (10 occ.).

confirmer notre hypothèse de départ de l'affinité de certaines CLS pour les sous-genres littéraires étudiés. Grâce aux méthodes textométriques, les expressions récurrentes relevées dans le POL et la SF participent donc à la caractérisation linguistique des genres, classés au départ selon des critères éditoriaux.

Conformément aux objectifs de ce numéro thématique visant à clarifier les statuts des unités phraséologiques, cette étude permet ainsi d'éclairer la variété structurelle des motifs (collocation lexicale, collocation grammaticale, patron lexico-syntaxique). Nous définissons donc le *motif*, à la suite de Longrée et Mellet (2013), selon plusieurs critères : l'attraction mutuelle entre des termes formant une cooccurrence, la récurrence de cette cooccurrence mesurée par des outils lexicométriques, et la présence de variations syntagmatiques et paradigmatiques au sein de cette séquence qui a également une fonction discursive.

Dans cette étude, nous avons relevé surtout des motifs séquentiels, mais on pourrait explorer quel est l'apport des ALR — propres à notre méthodologie d'observation — pour identifier des motifs hiérarchiques présentant une variabilité séquentielle (cf. Tutin & Kraif, dans ce volume) : pour cela, un corpus plus volumineux et plus diversifié en termes de genres et d'auteurs serait nécessaire. Le prolongement de cette étude pilote pourrait également consister à analyser comment la cellule organisée et fonctionnelle récurrente qu'est le motif occupe une fonction de « marqueur discursif structurant » (Mellet & Longrée, 2012, p. 718). Ainsi, par une étude stylistique et narratologique plus approfondie, il serait intéressant d'identifier les axes fonctionnels affectés par ces marqueurs (p. ex. diégèse, cadre thématique, dialogues, descriptions).

RÉFÉRENCES BIBLIOGRAPHIQUES

BERTELS, Ann & SPEELMANN, Dirk. (2013). 'Keywords Method' versus 'Calcul des Spécificités. A Comparison of Tools and Methods. *International Journal of Corpus Linguistics*, *18*(4), 536-560. <http://dx.doi.org/10.1075/ijcl.18.4.04ber>.

BIBER, Douglas. (2009). A Corpus-Driven Approach to Formulaic Language in English. Multi-Word Patterns in Speech and Writing. *International Journal of Corpus Linguistics*, *14*(3), 275-311. <http://dx.doi.org/10.1075/ijcl.14.3.08bib>

BOLLY, Catherine. (2010). Flou phraséologique, quasi-grammaticalisation et pseudo marqueurs de discours : un no man's land entre syntaxe et

discours ? *Linx*, *62-63*, 11-38. Disponible en ligne sur <http://linx.revues.org/1356> (consulté le 1er décembre 2015).

Bozzetto, Roger. (2007). *La science-fiction*. Paris : A. Colin.

Diwersy, Sascha, Goossens, Vannina, Grutschus, Anke, Kerne, Beate, Kraif, Olivier, Melnikova, Elena & Novakova, Iva. (2014). Traitement des lexies d'émotion dans les corpus et les applications d'*EmoBase*. *Corpus*, *13*, 269-293. Disponible en ligne sur <http://corpus.revues.org/2537> (consulté le 26 mars 2016).

Dufays, Jean-Louis. (2010). *Stéréotype et lecture. Essai sur la réception littéraire*. Bern : Peter Lang.

Dunning, Ted. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61-74. Disponible en ligne sur <www.aclweb.org/website/old_anthology/J/J93/J93-1003.pdf> (consulté le 1er décembre 2015).

Gonon, Laetitia, Kraif, Olivier, Novakova, Iva, Piat, Julien & Sorba, Julie. (À paraitre). *Sur la scène de crime…* Enquête sur les enjeux linguistiques et stylistiques de motifs récurrents dans le *thriller* contemporain. Dans *Actes du 5e Congrès mondial de linguistique française – CMLF 2016*. EDP Sciences. <www.linguistiquefrancaise.org>.

Hoey, Michael. (2005). *Lexical Priming: A New Theory of Words and Language*. Londres / New York : Routledge.

Hunston, Susan & Francis, Gill. (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam / Philadelphie : John Benjamins Publishing Company.

Kraif, Olivier & Diwersy, Sascha. (2012). Le Lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques. Dans G. Antoniadis, H. Blanchon & G. Sérasset (dir.), *Actes de la conférence conjointe JEP-TALN-RECITAL 2012* (vol. 2, p. 399-406). Grenoble : ATALA & AFCP. Disponible en ligne sur <www.jeptaln2012.org/actes/TALN2012/pdf/TALN2012033.pdf> (consulté le 1er décembre 2015).

Legallois, Dominique. (2012). La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique ? *Corpus*, *11*, 31-54. Disponible en ligne sur <https://corpus.revues.org/2202> (consulté le 1er décembre 2015).

Longrée, Dominique & Mellet, Sylvie. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours. *Langages*, *189*, 68-80. <http://dx.doi.org/10.3917/lang.189.0065>.

Mellet, Sylvie & Longree, Dominique. (2012). Légitimité d'une unité textométrique : le motif. Dans A. Dister, D. Longrée & G. Purnelle (dir.),

Actes des 11e Journées internationales d'analyse statistiques des données textuelles – JADT 2012 (p. 715-728). Liège. Disponible en ligne sur <http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Mellet,%20Sylvie%20et%20al.%20-%20Legitimite%20d'une%20unite%20textometrique.pdf> (consulté le 1er décembre 2015).

Nivre, Joakim, Hall, Johan, Kübler, Sandra, McDonald, Ryan, Nilsson, Jens, Riedel, Sebastian & Yuret, Deniz. (2007). The CoNLL 2007 Shared Task on Dependency Parsing. Dans J. Eisner (dir.), *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007* (p. 915-932). Prague : Association for Computational Linguistics. Disponible en ligne sur <http://www.aclweb.org/anthology/D/D07/D07-1096.pdf> (consulté le 1er décembre 2015).

Quiniou, Solen, Cellier, Peggy, Charnois, Thierry & Legallois, Dominique. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. Dans A. Dister, D. Longrée & G. Purnelle (dir.), *Actes des 11e Journées internationales d'analyse statistique des données textuelles – JADT 2012* (p. 821-833). Liège. Disponible en ligne sur <https://hal.archives-ouvertes.fr/hal-00675586/document> (consulté le 1er décembre 2015).

Rastier, François. (2011). *La mesure et le grain. Sémantique de corpus*. Paris : Honoré Champion.

Siepmann, Dirk. (2015). A Corpus-Based Investigations into Key Words and Key Patterns in Post-War Fiction. *Functions of Language*, *22*(3), 362-399. <http://dx.doi.org/10.1075/fol.22.3.03sie>.

Siepmann, Dirk. (À paraitre). Lexicologie et phraséologie du roman contemporain : quelques pistes pour le français et l'anglais. *Cahiers de lexicologie*.

Sinclair, John M. (1991). *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

Sinclair, John M. (2004). *Trust the Text: Language, Corpus and Discourse*. Londres : Routledge.

Stubbs, Michael. (1995). Collocations and Semantic profiles : on the cause of the trouble with quantitative studies. *Functions of Language*, *2*(1), 23-55. <http://dx.doi.org/10.1075/fol.2.1.03stu>.

Tapanainen, Pasi & Järvinen, Timo. (1997). A Non-Projective Dependency Parser. Dans *Proceedings of the 5th Conference on Applied Natural Language Processing* (p. 64-71). Stroudsburg, PA : Association for Computational Linguistics. Disponible en ligne sur <http://dx.doi.org/10.3115/974557.974568>.