Titre : Phraséologismes spécifiques dans les romans historiques et les romans de littérature blanche

Auteurs : Laetitia Gonon (Univ. Grenoble Alpes - Litt&Arts UMR 5316) & Julie Sorba (Univ. Grenoble Alpes - LIDILEM)

Notre étude lexicologique s'inscrit dans le cadre de la linguistique de corpus. Elle propose d'analyser plusieurs unités phraséologiques saillantes dans deux sous-corpus romanesques français (la littérature blanche ou générale et le roman historique). À la suite de Siepmann 2015, nous considérons que la langue littéraire se caractérise par la surreprésentation significative de phraséologismes. Notre objectif est donc de vérifier que les unités phraséologiques extraites par des méthodes statistiques permettent de contraster ces deux sous-genres littéraires. Le corpus est interrogé au moyen de l'interface Lexicoscope (Kraif 2016).

De récentes études, réalisées dans le cadre du projet ANR-DFG PhraseoRom (https://phraseorom.univ-grenoble-alpes.fr/accueil), ont montré que les phénomènes phraséologiques permettaient de distinguer deux sous-genres romanesques l'un de l'autre : le roman policier et le roman sentimental (Gonon, Goossens & Novakova, 2018 sous presse) ou le roman policier et la littérature blanche (Gonon, Goossens, Kraif, Novakova & Sorba, 2018 sous presse). L'objectif final du projet est de mettre au jour des critères aidant à l'identification et à la définition des genres romanesques en particulier et de la langue littéraire en général. Nous poursuivons ici ces analyses en contrastant les romans de littérature blanche et les romans historiques.

Notre étude s'inspire des modèles fonctionnels et contextualistes (Sinclair, 2004), qui explorent systématiquement quatre niveaux (lexical, sémantique, syntaxique et discursif) pour l'analyse des unités linguistiques. Ces analyses permettent de faire émerger des motifs que nous définissons à la suite de Longrée & Mellet (2013 : 66) comme « 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 ». Ces derniers peuvent être 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 : 45). Notre approche est essentiellement inductive (*corpus driven*).

Le corpus de notre étude est constitué de romans français contemporains postérieurs à 1950. Le sous-corpus de littérature blanche (GEN), d'environ 34 millions de mots, est constitué de 445 textes écrits par 170 auteurs (par ex. Tahar BenJelloun, Frédéric Beigbeder, Jean d'ormesson, Marie Darrieusecq, Didier Decoin, Jean Echenoz, David Foenkinos, Patrick Modiano etc.). Dans le sous-corpus de romans historiques (HIST), constitué d'environ 15 millions de mots, on trouve 114 textes répartis entre 39 auteurs dont David Camus, Maurice Druon, Patrick Girard, Christian Jacq, Robert Merle.

Le corpus a été annoté syntaxiquement au moyen de l'analyseur Xip (Aït Mokhtar, Chanod & Roux, 2001), ce qui permet d'en extraire automatiquement des arbres lexico-syntaxiques récurrents (ALR, Tutin & Kraif, 2016). Ces ALR regroupent des unités lexicales reliées par des dépendances syntaxiques et sont construits à partir de séries de cooccurrences statistiquement significatives (en fonction d'une mesure d'association statistique). Les ALR sont extraits des deux corpus à partir des pivots nominaux et verbaux dont la fréquence est supérieure à 5, puis leurs fréquences respectives sont comparées afin de mesurer leur spécificité dans chaque corpus. Suivant la méthode *Keywords* (Bertels & Speelmann, 2013), nous utilisons le calcul du rapport de vraisemblance ou *log-likelihood ratio* (LLR) afin de

déterminer si une répartition s'écarte significativement d'une distribution aléatoire ou non. De la sorte, on peut mettre en évidence les ALR dont la fréquence relative dans l'un de nos deux corpus est significativement supérieure à la fréquence dans l'autre corpus. Les critères retenus pour la sélection des ALR représentatifs sont les suivants :

a) LLR supérieur ou égal à 10,83, seuil à partir duquel la surreprésentation de l'ALR dans un corpus peut être considérée comme statistiquement significative ;
b) Critère morpho-syntaxique : on s'intéresse aux ALR contenant des verbes, ce qui permet d'exclure les expressions exclusivement référentielles.

Suite à l'application de ces critères, nous obtenons 2 589 ALR spécifiques à HIST et 3 828 ALR spécifiques à GEN. Nous avons ensuite opéré un second filtrage avec un LLR supérieur à 20, un nombre d'occurrences supérieur à 100 en maintenant une dispersion couvrant au minimum 20% des auteurs dans chaque sous-corpus. À l'issue de cette opération, nous remarquons dans HIST des ALR présentant des archaïsmes stylistiques propres au roman historique (par ex. : *dire vrai, ne … point, avoir ouï*) ainsi que des ALR au contenu plus référentiel traduisant une société hiérarchisée (par ex. : *monter sur le trône, ils s'inclinent, baiser les mains, être le chef*) et conquérante (par ex. : *venir à bout, fait la guerre, se faire égorger*). À côté de ces thématiques et traits stylistiques attendus, apparaissent des ALR *a priori* moins spécifiques de HIST : ce sont des ALR qui interviennent dans (ou initient) une interaction verbale à voix haute entre deux personnages (par ex. : *le roi dit, je dis monsieur, j'ai ouï, tu as raison, prend la parole, nous avons besoin, je vous assure*). Dans le sous-corpus GEN, les ALR relevant du domaine de l'interaction verbale sont aussi très présents ; néanmoins, celle-ci est nettement plus variée, et empiète sur le domaine de la cognition, l'interlocution étant davantage tournée vers l'intériorité (discours intérieur, discussion avec un interlocuteur *tu* intime de façon peu formelle, mais aussi avec soi-même) : par ex. *j'ai envie, j'ai peur, disant que c'était, je voulais pas, je me rappelais, j'ai besoin, il a peur, je lui dis, c'est moi, je crois que c'est, il dit que, je le savais*. Dans cet ensemble d'ALR de parole, nous avons choisi, pour la présente étude, de décrire 4 constructions lexico-syntaxiques (CLS) dans chaque sous-corpus présentés ci-dessous avec leur LLR entre parenthèses :

HIST	*donner l'ordre* (307)	*dit d'une voix* (191)	*je vous prie* (132)	*je vous remercie* (30)
GEN	*je sais pas* (1078)	*j'ai l'impression* (311)	*j'ai oublié* (376)	*je me dis* (250)

L'analyse linguistiques de ces 8 CLS permet d'identifier et d'analyser les motifs spécifiques à chaque sous-genre au sein desquels ces CLS se retrouvent. Sur le plan stylistique, nous étudions les fonctions textuelles (descriptives et narratives, Adam 2005) de ces motifs. Plus généralement, nous nous interrogerons sur la façon dont la phraséologie étendue permet de définir autrement les genres paralittéraires.

Références bibliographiques

Adam J.-M. (2005). *Les Textes : types et prototypes. Récit, description, argumentation, explication et dialogue*. Paris, Armand Colin.
Aït Mokhtar S., Chanod J.-P. & Roux C. (2001). Robustness beyond Shallowness: Incremental Deep Parsing, *Natural Language Engineering*, 8, p.121-144.
Gonon L., Goossens V., Kraif O., Novakova I. & Sorba J. (2018, sous presse). Motifs textuels spécifiques au genre policier et à la littérature « blanche ». *6e Congrès Mondial de Linguistique Française*, Université de Mons (Belgique).
Gonon L., Goossens V. & Novakova I. (2018, sous presse). Les phraséologismes spécifiques à deux sous-genres de la paralittérature : le roman sentimental et le roman policier. *Actes du colloque Phraséologie française*. Paris, Hermann.
Kraif, O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de Lexicologie* 108, p. 91-106
Siepmann, D. (2015). A corpus-based investigation into key words and key patterns in post-war fiction, *Functions of Language*, 22/3, p. 362-399.
Tutin A. & Kraif O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents, *Lidil*, 53, p. 119-141.
Bertels A. & Speelmann D. (2013). 'Keywords Method' versus 'Calcul des Spécificités'. A comparison of tools and methods, *International Journal of Corpus Linguistics*, 18/4, p. 536-560.
Sinclair J. (2004). *Trust the Text: Language, Corpus and Discourse*. Londres, Routledge.
Longrée D. & Mellet S. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours, *Langages*, 189, p. 68-80.
Legallois D. (2012). La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique ?, *Corpus*, 11, p. 31-54.

Mots clefs : phraséologie, corpus, constructions lexico-syntaxiques, motif textuel, genres romanesques, français