Spécificités des expressions spatiales et temporelles dans quatre sous-genres romanesques (policier, science-fiction, historique et littérature générale)

Olivier Kraif, Julie Sorba

Spécificités des expressions spatiales et temporelles dans quatre sous-genres romanesques (policier, science-fiction, historique et littérature générale)

Olivier Kraif[1], Julie Sorba[2]

[1]Univ. Grenoble Alpes, LIDILEM – olivier.kraif@univ-grenoble-alpes.fr

[2]Univ. Grenoble Alpes, LIDILEM – julie.sorba@univ-grenoble-alpes.fr

Abstract

In this paper, we aim to test if the classifications of the phraseological units based on recurring trees and ngram methods are functional in order to separate novel genres one from another. Our results confirm that these two methods are relevant for the expressions relative to space and time into our corpora.

Résumé

Notre objectif est de tester les classifications des phraséologismes, opérées par les méthodes des ALR et des SR, dans le but de distinguer des sous-genres romanesques les uns des autres. Dans nos corpus, nos résultats confirment la pertinence de ces classifications pour les deux champs de l'espace et du temps.

Keywords: ngram, recurring trees, novel genres, phraseology

1. Introduction

Notre étude, qui s'inscrit dans le cadre de l'analyse exploratoire des données textuelles, concerne des romans français contemporains rassemblés dans le cadre du projet ANR-DFG PhraseoRom. Ce corpus (plus de 110 millions de mots pour le français) est partitionné en plusieurs sous-corpus correspondant à différents sous-genres littéraires (policier, science-fiction, fantasy, roman historique, roman sentimental, littérature générale). Notre objectif est de caractériser ces genres et sous-genres textuels par les unités phraséologiques spécifiques qu'ils contiennent. À l'instar de Boyer, nous postulons que « chaque genre comprend un certain nombre de sous-ensembles, des séries fondées sur la réutilisation de composantes identiques » (1992, p.91). Dans la mesure où la phraséologie étendue s'intéresse à tout ce qui est « préfabriqué » dans les séquences lexicales, elle constitue donc un point d'entrée privilégié pour mettre en évidence ces « séries ».

Pour cette étude, nous retenons spécifiquement 4 sous-genres : les romans de science-fiction (SF), les romans policiers (POL), les romans historiques (HIST) et les romans de littérature dite blanche ou générale (GEN). La fouille des textes utilise la technique de repérage des Arbres Lexicosyntaxiques Récurrents (ou ALR, Kraif & Diwersy, 2012 ; Kraif, 2016) dont la validité a déjà été montrée par le repérage d'unités phraséologiques spécifiques dans les textes scientifiques (Tutin & Kraif, 2016). Nous proposons en outre de comparer ici cette technique d'extraction avec celle des segments répétés (Salem, 1987), les ALR ayant montré une meilleure prise en compte de la variabilité syntaxique pour le repérage des routines, mais s'avérant parfois défaillants pour identifier des segments figés en surface, du fait du modèle dépendanciel employé.

Dans des travaux antérieurs, nous avons montré comment les ALR permettaient de repérer des motifs récurrents construits autour d'expressions spécifiques fortement liées à la composante thématique des sous-genres en question : c'était le cas pour « scène de crime » dans POL (Kraif, Novakova & Sorba, 2016). Ici, nous nous concentrons sur des expressions moins directement liées aux univers de référence des sous-genres (le crime, l'amour, la science, etc.), afin de mettre en évidence des traits moins prévisibles. C'est pourquoi, nous avons choisi de sélectionner les séquences – bien souvent adverbiales – liées à l'expression du temps et de l'espace.

Nous allons désormais présenter les résultats obtenus dans des travaux antérieurs (partie 2), puis décrire notre méthodologie expérimentale (partie 3). Enfin, nous exposerons et discuterons nos observations (partie 4) avant de proposer des conclusions et perspectives à notre étude (partie 5).

2. Travaux antérieurs

Lefer, Bestgen & Grabar (2016) s'appuient sur une extraction de n-grammes de 2 à 4 mots pour caractériser 3 genres textuels : des débats parlementaires européens, des éditoriaux de presse et des articles scientifiques. Ces auteurs utilisent une méthode d'AFC pour identifier les expressions les plus typiques et en tirent des observations contrastives concernant l'expression de la certitude et de l'opinion. De notre côté, nous avons analysé des contrastes génériques sur un plan qualitatif, en identifiant des ALR dans des corpus de romans policiers et de science-fiction, en nous fondant sur des mesures de spécificité (Kraif, Novakova & Sorba, 2016). Nous avons également utilisé l'extraction des ALR pour classer automatiquement, dans une approche supervisée, des sous-corpus POL, SF et GEN (Chambre & Kraif, 2017). Ces travaux préliminaires ont montré que les ALR donnaient de meilleurs résultats que les autres catégories de traits (ponctuation, morphosyntaxe, lexique), et permettaient de classer correctement 98% des textes du corpus à partir d'une sélection de traits discriminants. La plupart de ces traits appartenaient à des champs lexicaux précis, liés aux univers de référence propres à chaque sous-genre, comme ceux du 'téléphone' (*le numéro de portable*, *passer un coup de fil, etc.*) ou de la 'voiture' (*à travers le pare-brise*, *démarrer en trombe, etc.*) pour POL. De plus, des expressions temporelles (p.ex. pour POL *à huit heures, vingt et une heure, au bout de X minutes*) et des indications spatiales très variées (p.ex. pour SF *par la voie, dans le territoire, dans la sphère, dans l'espace, la zone de*) ont été mises en évidence.

Nous proposons ici un prolongement de cette expérimentation, d'une part, en étudiant les expressions spatiales et temporelles, et d'autre part, en ajoutant le sous-genre des romans historiques (HIST), afin de déterminer si ces classes d'expression sont suffisantes pour différencier les quatre sous-genres (POL, SF, GEN, HIST).

3. Méthodologie

Pour chaque sous-genre, notre corpus comporte un échantillon d'environ 8 millions de mots, correspondant à environ 70 œuvres d'une quarantaine d'auteurs (cf. Tableau 1). Ces œuvres sont toutes postérieures à 1950, et la majorité d'entre elles ont été publiées pour la première fois après 2000. La classification des œuvres en genre a été effectuée a priori selon des critères éditoriaux, en fonction des collections de publication.

	Auteurs	Romans	Taille
POL	46	69	8 008 395
SF	36	75	8 001 582
HIST	38	70	8 015 933
GEN	46	69	8 008 395

Tableau 1 : Constitution du corpus

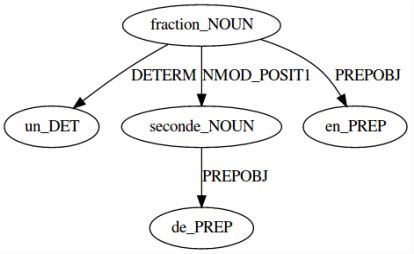

Figure 1 : ALR représentant l'expression en une fraction de seconde

Pour identifier les expressions phraséologiques caractéristiques des différents sous-genres, nous utilisons deux méthodes de repérage :

- la méthode des ALR : nos corpus étant analysés en dépendances avec XIP (Aït-Mokhtar et al., 2002), ces ALR sont des sous-arbres respectant des critères de fréquence (ici ≥ 10 occurrences), de dispersion (ici ≥ 10 auteurs différents, appartenant à au moins 3 sous-genres différents) et de taille (ici ≥ 3 nœuds et ≤ 8 nœuds). En outre, lors de la recherche de ces ALR, une mesure d'association est calculée afin de ne retenir que les nœuds significativement associés avec le reste de l'arbre. La figure 1 montre un exemple d'ALR correspondant à l'expression *en une fraction de seconde*.

- la méthode des segments répétés (ou SR, Salem, 1987) : nous avons appliqué les mêmes critères de dispersion et de taille (≥ 3 et ≤ 8), afin de comparer les deux méthodes *in fine*. Les SR sont constitués de séquences de lemmes (obtenus avec XIP), et non de formes fléchies. Cette dernière méthode est plus simple à mettre en œuvre et nécessite peu de ressources linguistiques, bien qu'elle pose des problèmes d'explosion combinatoire (cf. partie 4).

Dans un second temps, nous appliquons un filtrage par mots-clés afin de ne retenir que les séquences liées aux deux sous-domaines étudiés, à savoir l'expression du temps et de l'espace. Les mots-clés pour l'espace sont des noms de lieux, d'espaces naturels, de description géographique, des mesures de distance, des adverbes de lieu, sélectionnés après un premier sondage des ALR extraits :

- Mots-clés ESPACE : cave, salon, hôpital, immeuble, bâtiment, camp, restaurant, village, route, rue, quai, chaussée, terrasse, ministère, parc, bureau, carlingue, maison, toit, chambre, hôtel, palais, rez-de-chaussée, entrée, pont, escalier, chemin, place, salle, jardin, seuil, cour, couloir, colline, sentier, sol, rive, rivage, plage, rivière, mont, montagne, mer, océan, lac, bois, forêt, espace, endroit, coin, pays, continent, frontière, direction, cap, sud, est, nord, ouest, confins, mètre, kilomètre, année-lumière, hectare, acre, loin, proche, près de, au bord de, orée, distance.

Les mots-clés pour le temps désignent des moments de la journée et de l'année, des unités de mesure et des découpages conventionnels de période (noms, adverbes et locutions adverbiales) :

- Mots-clés TEMPS : matin, soir, soirée, après-midi, nuit, jour, temps, fois, moment, instant, toujours, jamais, parfois, souvent, autrefois, jadis, tôt, tard, longtemps, brièvement, immédiatement, subitement, tout à coup, tout de suite, aujourd'hui, demain, hier, lendemain, maintenant, heure, minute, seconde, journée, semaine, mois, an, année, décennie, siècle, millénaire, printemps, été, automne, hiver.

Ces listes ne prétendent pas être exhaustives et le filtrage opéré produit à la fois du silence et du bruit, du fait des ambiguïtés. Celles-ci demeurent toutefois marginales (d'après un sondage manuel, le bruit est inférieur à 10 %).

Pour identifier les ensembles de traits pertinents du point de vue des sous-genres, nous injectons ces expressions (ALR ou SR) dans un système de classification automatique. De la sorte, nous visons un double objectif : d'une part, vérifier que nos classes constituées a priori

sont cohérentes et corrélées à des critères objectivables ; d'autre part, identifier ces critères sous la forme d'ensemble de traits discriminants pour la classification.

4. Résultats et discussion

Dans une première étape, nous avons extrait les 6000 ALR les plus fréquents sur l'ensemble du corpus. En effectuant une classification sur ces traits, avec un modèle SVM optimisé par SMO (avec la plate-forme Weka, Eide et al. 2016), on obtient, dans une évaluation croisée à 10 plis, une précision de 74 % (123 sur 166), avec un Kappa de 0,65, ce qui correspond à un très bon accord avec la classification de référence. La matrice de confusion (cf. Tableau 2) montre que les deux genres les mieux classés sont SF (93,1 %) et POL (79,5 %). Le genre GEN obtient la précision la plus faible (64%) avec des confusions fréquentes avec POL et HIST ; HIST est de son côté fréquemment confondu avec GEN.

L'examen des ALR les plus discriminants montre, comme on pouvait s'y attendre, la forte présence de certains thèmes dans POL, HIST et SF (la voiture, le crime, le téléphone pour POL ; la guerre, la religion pour HIST ; l'univers spatial et les artefacts technologiques pour SF) et l'absence de traits saillants dans GEN.

4.1 Sélection des traits TEMPS+ESPACE

Lorsqu'on sélectionne les traits liés à l'expression du temps seul (environ un millier), on obtient une dégradation par rapport aux résultats précédents, avec une précision globale de 48,8 % et un Kappa de 0,31 signifiant un accord faible entre la classification a priori et la classification automatique. Les expressions spatiales, de leur côté (on en obtient 1560, mais nous avons retenu les 1000 plus fréquentes afin de disposer de résultats comparables), obtiennent des résultats un peu meilleurs, toutefois moins bons que les traits non filtrés : la précision est de 59,6 %, avec un Kappa de 0,46 correspondant à un accord modéré.

Quand on sélectionne conjointement les ALR de TEMPS et ESPACE, on obtient une légère amélioration par rapport à la classification avec ESPACE seul : 61,4 % (102 instances bien classées sur 166), avec un Kappa assez bon de 0,48. La matrice de confusion (cf. tableau 2) montre que POL obtient la meilleure précision (69%) et GEN la moins bonne (55,9 %).

Si on sélectionne les traits les plus discriminants (attributs *SfcSubsetEval* avec méthode *BestFirst* dans Weka), on obtient un ensemble de 54 attributs. On peut évaluer, de manière indicative, le pouvoir classificateur de ces attributs sur notre corpus en les réinjectant dans une classification par SMO : on obtient alors une précision globale très légèrement supérieure (62 %), mais il est intéressant de noter que les genres marqués POL, SF et HIST sont très bien classés sur la base de ces traits (précision de 85,7% pour HIST, 84 % pour SF, 75,7 % pour POL) avec une dégradation forte pour GEN (43,4%), comme le montre la matrice de confusion ci-dessous (tableau 2).

	(1) Tous les traits (6000 ALR plus fréquents)				(2) TEMPS+ESPACE (2571 traits filtrés)				(3) TEMPS+ESPACE Sélection de 54 traits			
	SF	POL	GEN	HIST	SF	POL	GEN	HIST	SF	POL	GEN	HIST
SF	27	2	2	5	18	5	6	7	21	2	13	0
POL	1	35	9	1	5	29	12	0	3	28	15	0
GEN	1	5	32	8	3	3	33	7	1	6	36	3

HIST	0	2	7	29	3	5	8	22	0	1	19	18

Tableau 2 : Matrices de confusion pour les classifications avec (1) tous les traits, (2) les ALR filtrés (TEMPS+ESPACE) et (3) les ALR sélectionnés

L'examen détaillé des 54 traits sélectionnés révèle plusieurs points saillants :

- d'une manière générale, les ALR relatifs à l'espace sont très largement majoritaires avec 33/54 contre 17/54 pour le temps, après élimination du bruit (4/54).

- si on considère les traits spécifiques à HIST, les expressions spatiales désignent surtout des lieux de pouvoir (*la place forte, de son palais, salle du palais, salle du château, pénétrer dans la grande salle*) et la mer (*sur la mer, de la mer*), tandis que les expressions temporelles font référence à une temporalité longue (*au bout de quelques mois*, *règne de X années*, *avoir le temps*) et à des datations absolues ou relative (*du N^e^ siècle, venir le lendemain, à trois heures de l'après-midi*).

- pour POL, en revanche, les expressions temporelles indiquent des datations horaires (*à 8 heures, 21 heures*) et des durées courtes (*une vingtaine de secondes*). Les expressions spatiales, nombreuses, indiquent des pièces et des espaces intérieurs (*de la salle de bain, vers la salle de bain, entrer dans le bureau, vers le bureau, dans le coin*), des lieux urbains (*aller à l'hôtel, passer à l'hôpital, à l'hôpital*), et des localisations vagues (*dans le coin* au sens de « dans les parages »).

- pour SF, les expressions temporelles sont plus nombreuses (7/18) que dans les autres sous-genres. Elles font référence à des durées extrêmes par leur longueur (*milliers d'années, de mille ans*) ou leur brièveté (*une fraction de seconde, un centième de seconde*). Pour l'espace, on trouve des expressions de distances chiffrées *(dizaines de mètres, centaine de mètres, plusieurs centaines de mètres*), des références attendues à l'espace intersidéral (*dans l'espace, à travers l'espace, être dans l'espace, voyager dans l'espace, flotter dans l'espace*), à l'*espace-temps* et des expressions avec *sol* (*sur le sol, sous-sol*).

- pour GEN : la seule expression spécifique apparaissant dans les traits sélectionnés est *chemin de traverse*.

4.2 Comparaison avec les segments répétés

Nous n'avons pas réussi à extraire la totalité des SR de 3 à 8 mots pour l'ensemble du corpus, du fait des problèmes d'explosion combinatoire (environ 40 000 000 SR générés pour 100 textes du corpus). Nous avons donc retenu les SR contenant les mots-clés sélectionnés pour TEMPS et ESPACE, en conservant les 1000 SR les plus fréquents afin d'avoir des ensembles de traits comparables aux ALR filtrés. On obtient de meilleurs résultats que pour les ALR, avec une précision de 66,7 % pour ESPACE et 58,3 % pour TEMPS contre respectivement 59,6 % et 48,8 %. Pour TEMPS+ESPACE, on constate une certaine dégradation, avec une précision qui tombe à 64,1 %. À ce stade de nos observations, il nous est difficile d'interpréter ces résultats quantitatifs car la sélection du meilleur ensemble de traits pour ESPACE donne peu ou prou les mêmes expressions qu'avec les ALR :

le chambre de, le cour de, à le cour, dans le espace, le salle de bain, de le espace, dans son bureau, de le immeuble, le maison et, à le hôtel de, centaine de mètre, sur le bureau, sur le place de, le palais de, dans le grand salle, de bureau de, de le salle de bain, sur son bureau, cour de France, en route pour, dans mon bureau, dans tout le direction, un dizaine de mètre, de son pays, à le rue, dans le sous-sol, quitter le salle, dans un

restaurant, sur le rivage, mètre plus bas, vers le bureau, route vers le, dizaine de mètre de, un kilomètre de, à ministère de, dans le espace et, de un montagne, le espace et le.

Les deux méthodes donnent donc des résultats convergents en termes qualitatifs en extrayant les mêmes expressions. Néanmoins, des investigations complémentaires seront nécessaires pour interpréter correctement le fait que les SR obtiennent de meilleurs résultats quantitatifs.

5. Conclusion et perspectives

Cette étude confirme que les expressions phraséologiques constituent de bons descripteurs pour la classification en sous-genre (Chambre & Kraif, 2017). En effet, même si les résultats obtenus ici à partir du sous-ensemble constitué des expressions spatiales et temporelles sont sensiblement inférieurs à ceux obtenus à partir de traits plus directement liés aux univers de référence de chaque sous-genre (61.4 % /vs/ 74 %), ces expressions moins riches sur le plan informatif permettent cependant de classer les romans dans les sous-genres marqués POL, SF et HIST de manière satisfaisante. En revanche, pour la catégorie des romans généraux (GEN), elles ne sont pas discriminantes. Notre méthode permet aussi de dégager des spécificités génériques propres à ces deux champs ESPACE et TEMPS (lieux de pouvoir dans HIST /vs/ intérieur et lieux urbains dans POL ; durées et distances extrêmes dans SF). Enfin, à partir de cette sélection d'expressions spatio-temporelles, la méthode des segments répétés produit une classification en sous-genres plus précise que celle des ALR. Ce point, difficile à interpréter à partir de nos premières observations qualitatives, nécessite une étude plus approfondie. Ces résultats nous incitent à poursuivre l'exploration d'autres champs lexicaux en marge des univers de référence de chaque sous-genre, afin, d'une part, d'affiner notre méthodologie et, d'autre part, de cibler les éléments au cœur de la phraséologie.

Références

Aït-Mokhtar S., Chanod J.-P. and Roux C. (2002). Robustness beyond Shallowness: Incremental Deep Parsing. *Natural Language Engineering*, 8:121-144.

Boyer A.-M. (1992). *La paralittérature*. Presses Universitaires de France.

Chambre J. et Kraif O. (2017). Identification de traits spécifiques du roman policier et de science fiction. Communication présentée aux *Journées Internationales de la Linguistique de Corpus - JLC2017*, Grenoble, 05.07.2017.

Eibe F., Hall M. A. and Witten I. H. (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition.

Kraif O., Novakova I. et Sorba J. (2016). Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction. *Lidil*, 53 : 143-159.

Kraif O. et 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*, pp. 399-406.

Lefer M.-A., Bestgen Y. et Grabar N. (2016). Vers une analyse des différences interlinguistiques entre les genres textuels : étude de cas basée sur les n-grammes et l'analyse factorielle des correspondances. *Actes de la conférence conjointe JEP-TALN-RECITAL 2016*, pp. 555-563.

Tutin A. et Kraif O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents. *Lidil*, 53 : 119-141.

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Klincksieck.