Fouille de données pour la stylistique : cas des motifs séquentiels émergents

Solen Quiniou[1,2], Peggy Cellier[3], Thierry Charnois[1], Dominique Legallois[2]

[1] GREYC Université de Caen Basse-Normandie – Campus 2, 14032 Caen cedex – France
[2] CRISCO Université de Caen Basse-Normandie – Campus 1, 14032 Caen cedex – France
[3] IRISA-INSA de Rennes – Campus de Beaulieu, 35042 Rennes cedex – France

Abstract

In this paper, we study the use of data mining techniques for stylistic analysis, from a linguistic point of view, by considering emerging sequential patterns. First, we show that mining sequential patterns of words with *gap* constraints gives new relevant linguistic patterns with respect to patterns built on state-of-the-art *n*-grams. Then, we investigate how sequential patterns of itemsets can provide more generic linguistic patterns. We validate our approach both from a quantitative and a linguistic point of view by conducting experiments on three corpora of various types of French texts (poetry, letters, and fiction, respectively). By considering more particularly poetic texts, we show that characteristic linguistic patterns can be identified using data mining techniques.

Résumé

Dans cet article, nous présentons une étude sur l'utilisation de méthodes de fouille de données pour l'analyse stylistique - d'un point de vue linguistique - en considérant des motifs séquentiels émergents. Nous montrons tout d'abord que la fouille de motifs séquentiels de mots en utilisant la contrainte *gap* permet d'obtenir de nouveaux patrons linguistiques pertinents par rapport aux patrons construits à partir de *n*-grammes. Nous étudions ensuite l'utilisation de motifs séquentiels d'itemsets pour produire des patrons linguistiques plus généraux. Nous validons notre approche d'un point de vue quantitatif et d'un point de vue linguistique, en réalisant des expérimentations sur trois corpus français correspondant à différents genres de texte (la poésie, les correspondances et les romans, respectivement). En considérant plus particulièrement les textes poétiques, nous montrons que les techniques de fouille de données employées permettent d'identifier des patrons linguistiques caractéristiques.

Mots-clés : fouille de données, stylistique, motifs séquentiels émergents, patrons linguistiques.

1. Introduction

Depuis ces 30 dernières années, l'étude linguistique de la phraséologie connaît un intérêt grandissant, et plus particulièrement les travaux en linguistique de corpus. Deux grandes catégories d'approches peuvent être identifiées : les approches *corpus-based* et les approches *corpus-driven*. Les premières supposent l'existence de théories linguistiques et utilisent les corpus pour observer comment elles s'appliquent afin de les valider. Les secondes considèrent que les constructions linguistiques émergent de l'analyse des corpus : cette analyse permet de découvrir des motifs de mots co-occurrents qui serviront ensuite de base pour des analyses

linguistiques à proprement parler. Notre travail se situe dans le cadre des approches *corpus-driven* car notre objectif est d'aider les linguistes à découvrir de nouvelles constructions linguistiques sans utiliser de connaissances *a priori*.

Une des premières approches *corpus-driven* a été proposée dans (Renouf et Sinclair, 1991). Elle concerne l'étude de cadres collocationnels à l'aide de corpus ; les *cadres collocationnels* sont des séquences discontinues de deux mots grammaticaux séparé par un mot lexical (par exemple, *« many +? + of »*[1]). Cependant, cette approche n'est pas entièrement *corpus-driven* car les cadres collocationnels étudiés sont sélectionnés a priori. En fait, la plupart des approches dites *corpus-driven* - dont les grammaires de patterns (Hunston et Francis, 2000) - sont en partie *corpus-based*[2]. Un travail plus récent dans (Biber, 2009) propose une approche intéressante, entièrement *corpus-driven*, pour identifier des motifs fréquents à partir de corpus. Il s'appuie pour cela sur un travail préliminaire sur l'identification de *lexical bundles* - c'est-à-dire des séquences de mots fréquentes, également connues sous le nom de n-grammes - ainsi que sur les cadres collocationnels de Renouf et Sinclair afin d'identifier les éléments fixes et les éléments variables des motifs extraits. De plus, Biber considère deux registres de discours (conversations et écrits académiques) et montre l'intérêt d'une approche *corpus-driven* pour étudier les spécificités des motifs apparaissant dans chaque registre.

Dans cet article, nous présentons une première étude originale visant à montrer l'intérêt des méthodes de fouille de données pour l'analyse stylistique d'importantes collections de textes. L'objectif est de fournir, à des experts linguistes, des motifs pertinents et compréhensibles qui soient caractéristiques de genres de texte afin que ces experts puissent réaliser une analyse stylistique en s'appuyant sur ces motifs. Nous nous plaçons ainsi en quelque sorte dans la continuité des travaux de (Biber, 2009) en considérant ici différents genres de texte, dans le cadre d'une étude stylistique. Pour ce faire, nous mettons en place une méthodologie basée sur la fouille de données séquentielles, allant de l'extraction de motifs à la sélection des motifs les plus pertinents ; nous appliquons cette méthodologie au cadre de la stylistique. À notre connaissance, les méthodes de fouille de données n'ont pas encore été utilisées dans le domaine de la stylistique alors qu'elles ont l'avantage de fournir un résultat interprétable par un utilisateur, contrairement aux méthodes numériques telles que les modèles de Markov cachés ou les champs conditionnels aléatoires. En effet, ces dernières méthodes donnent de bons résultats pour des tâches de catégorisation de textes ou d'extraction d'information mais produisent des résultats difficilement compréhensibles par un humain. Dans cet article, nous nous intéressons à la fouille de *motifs séquentiels fréquents* (Agrawal et Srikant, 1995), une technique bien connue de la fouille de données qui permet de découvrir automatiquement des connaissances en tenant compte de l'ordre séquentiel entre les données. Nous considérons deux types de motifs séquentiels : des motifs d'items (un *item* représente une seule information, *e.g.* la forme d'un mot) et des motifs *d'itemsets*. Dans ce deuxième type de motifs, un mot est alors représenté par un ensemble de traits. Les motifs d'itemsets extraits peuvent ainsi combiner différents niveaux d'abstraction (*e.g.* des formes de mots, des lemmes ou des catégories morpho-syntaxiques) ; par exemple, <(PREP)(DET)(NC)> ou <(pour)(la DET)(NC)>[3]. De plus, dans le cadre de l'étude

1 *« many +? + of »* signifie *many* suivi d'un lexème variable *(symbolisé par ?)* et suivi de *of.*

2 Voir (Biber, 2009) pour un état de l'art plus détaillé sur les approches *corpus-driven* et (Legallois et François, 2006) pour une synthèse sur les grammaires de patterns et de constructions

3 PREP (respectivement DET et NC) correspond à *préposition* (resp. *déterminant* et *nom commun*)

stylistique de textes, l'objectif est d'extraire des patrons linguistiques caractéristiques des genres de texte considérés. C'est pourquoi nous nous focalisons sur un type particulier de motifs séquentiels : les *motifs émergents*. Les motifs émergents permettent de mettre en évidence des caractéristiques propres à des classes ou à des ensembles de données (Dong et Li, 1999). Ces motifs peuvent également être analysés par des experts pour découvrir de nouvelles relations dans un domaine donné afin de mieux le comprendre. Dans le cadre d'études stylistiques, les motifs émergents extraits peuvent ainsi être analysés par des linguistes pour découvrir des patrons linguistiques caractéristiques de genres de texte.

Dans la suite, la section 2 introduit notre méthodologie qui s'appuie sur la fouille de données séquentielles. Les résultats expérimentaux de son application à la stylistique sont ensuite présentés dans la section 3 ; l'analyse est réalisée d'un point de vue quantitatif mais également d'un point de vue linguistique. Pour finir, la section 4 conclut cette présentation et propose également des pistes à suivre pour améliorer cette première étude.

2. Méthodologie

Dans cette section, nous donnons tout d'abord une vue générale de l'approche proposée pour identifier des patrons caractéristiques de genres de texte (section 2.1). Nous présentons ensuite les techniques de fouille de données sur lesquelles s'appuie notre approche : les motifs séquentiels fréquents (section 2.2) et les motifs émergents (section 2.3).

2.1. Présentation générale de l'approche

La figure 1 illustre les différentes étapes de notre approche.

Figure 1 : Vue générale de notre approche

N corpus sont utilisés en entrée du processus, soit un corpus par genre littéraire considéré. Chaque corpus est tout d'abord pré-traité puis ses mots sont étiquetés avec leur lemme et leur catégorie morpho-syntaxique (cf. section 3.1.1). Lors de la première étape de l'approche, des motifs séquentiels sont extraits pour chacun des corpus : *N* ensembles de motifs sont ainsi obtenus. Lors de la seconde étape de l'approche, les motifs émergents de chaque genre sont calculés à partir des *N* ensembles de motifs séquentiels extraits précédemment. Pour finir, les *N* ensembles de motifs émergents sont proposés à un linguiste, afin qu'il réalise une interprétation linguistique à l'aide de ceux-ci. Les deux premières étapes sont présentées plus en détail dans les sous-sections suivantes (le pré-traitement des corpus et le choix des paramètres sont décrits dans la section 3.1).

2.2. Fouille de motifs séquentiels

La *fouille de données séquentielles* est une technique de fouille de données bien connue pour trouver des régularités dans des bases de données, en tenant compte de l'ordre temporel entre les

données ; elle a été introduite dans (Agrawal et Srikant, 1995). Un *itemset*, *I*, est un ensemble de littéraux appelés *items* et est représenté par $I=(i_1...i_n)$. Par exemple, *(a b)* est un itemset contenant deux items : *a* et *b*. Une *séquence*, *S*, est une liste ordonnée d'itemsets et est représentée par $S=<I_1...I_m>$. Par exemple, *<(a b)(c)(d)(a)>* est une séquence de quatre itemsets. Il est à noter que de nombreuses applications ne requièrent qu'un seul item dans leurs itemsets (*e.g.* les chaînes d'ADN ou les séquences de protéines). Ces séquences particulières sont appelées des *séquences d'items* ; pour plus de clarté, elles sont représentées par $S=<i_1...i_m>$, où $i_1...i_m$ sont des items. De nombreux algorithmes ont été développés pour fouiller efficacement ce type de séquences particulières, par exemple (Nanni et Rigotti, 2007). Dans la suite de cet article, les deux types de séquences seront considérés : les séquences d'items et les séquences d'itemsets. Une séquence $S1=<I_1...I_n>$ est *contenue* dans une séquence $S2=<I'_1...I'_m>$ s'il existe des entiers $1 \leq j_1 <...< j_n \leq m$ tels que $I_1 \subseteq I'_{j1}$, …, $I_n \subseteq I'_{jn}$. La séquence S1 est ainsi appelée *sous-séquence* de S2, ce qui est noté $S1 \leq S2$. Par exemple, nous avons la relation suivante : $<(c)(a)\rangle \leq \langle(a\ b)(c)(d)(a)>$. Une base de séquences *SDB* est un ensemble de tuples (*sid, S*), où *sid* est un identifiant de séquence et *S* une séquence. Le tableau 1 représente une base de quatre séquences :

Identifiant de séquence	Séquence
1	<(*a b*)(*c*)(*d*)(*a*)>
2	<(*d*)(*a*)(*e*)>
3	<(*d*)(*a b e*)(*c d e*)>
4	<(*c*)(*a*)>

Tableau 1 : Exemple de base de séquences

Un tuple (*sid, S*) *contient* une séquence S_1 si $S_1 \leq S$. Le *support* d'une séquence S_1 dans une base de séquences *SDB*, noté $sup(S_1)$, est le nombre de tuples contenant S_1 dans la base. Par exemple, dans le tableau 1, $sup(<(a)(e)>) = 2$, puisque les séquences 2 et 3 contiennent un itemset avec *a* suivi d'un itemset avec *e*. Le *support relatif* peut aussi être utilisé, comme défini par l'équation suivante :

$$sup(S_1) = \frac{|\{(sid, S) \mid (sid, S) \in SDB \wedge (S_1 \preceq S)\}|}{|SDB|}$$

Un *motif fréquent* est une séquence dont le support est supérieur ou égal à un seuil fixé : minsup. Les algorithmes de fouille de motifs séquentiels extraient ainsi tous les motifs fréquents apparaissant dans une base de séquences.

L'ensemble des motifs fréquents pouvant être très grand, il existe une représentation condensée permettant d'éliminer les redondances sans perte d'information : les *motifs clos* (Yan *et al.*, 2003). Un motif fréquent *S* est clos s'il n'existe aucun motif fréquent *S'* tel que $S \leq S'$ et $sup(S) = sup(S')$. Par exemple, en fixant *minsup=2*, le motif <(*b*)(*c*)> du tableau 1 n'est pas clos alors que le motif <(*a b*)(*c*)> est clos. De plus, afin de diriger le processus de fouille selon les besoins de l'utilisateur et d'éliminer des motifs non pertinents, des contraintes peuvent être définies (Dong et Pei, 2007). La contrainte la plus couramment utilisée est la contrainte de fréquence (en donnant une valeur à *minsup*). Une autre contrainte couramment employée est la contrainte *gap*. Un motif avec une contrainte *gap[M,N]*, noté $P_{[M,N]}$, est un motif dont chaque couple d'itemsets

est séparé par au moins *M-1* itemsets et au plus *N-1* itemsets. Par exemple, considérons $P_{[1,3]}$ = <(*c*)(*a*)> et $P_{[2,3]}$ = <(*c*)(*a*)> deux motifs avec des contraintes *gap* différentes et considérons les séquences du tableau 1. Les séquences 1 et 4 sont des occurrences du motif $P_{[1,3]}$ (la séquence 1 comporte un itemset entre (*c*) et (*a*) et la séquence 4 ne comporte aucun itemset entre (*c*) et (*a*)) alors que seule la séquence 1 est une occurrence du motif $P_{[2,3]}$ (seules les séquences comportant un ou deux itemsets entre (*c*) et (*a*) sont des occurrences de ce motif).

Dans cet article, les bases de séquences considérées correspondent à des corpus. De plus, deux types de motifs sont considérés : des motifs d'items et des motifs d'itemsets. Dans ce dernier cas, les itemsets peuvent alors comporter trois types d'items : des formes de mots, des lemmes et des catégories morpho-syntaxiques.

2.3. Motifs séquentiels émergents

Les motifs émergents sont des motifs dont le support augmente de manière significative d'un ensemble de données à un autre (Dong et Li, 1999). Les motifs émergents sont ainsi des motifs dont le *taux de croissance* - le rapport des supports dans deux ensembles de données - est supérieur à un seuil fixé : ρ. Un motif *P* d'un ensemble de données D_1 est alors un *motif émergent*, par rapport à un autre ensemble de données D_2, si *TauxCroiss(P)* $\geq \rho$, avec $\rho \geq 1$ et *TauxCroiss(P)* défini par l'équation suivante :

$$TauxCroiss(P) = \begin{cases} \infty, & \text{si } sup_{D_2}(P) = 0 \\ \frac{sup_{D_1}(P)}{sup_{D_2}(P)}, & \text{sinon} \end{cases}$$

où $sup_{D1}(P)$ (respectivement $sup_{D2}(P)$) est le support relatif du motif *P* dans D_1 (respectivement dans D_2). Comme nous nous intéressons uniquement aux motifs émergents dans D_1, nous ne considérons pas les motifs *P* tels que $sup_{D1}(P)=0$.

Dans le cadre de l'analyse stylistique, chaque ensemble de données contient les motifs séquentiels fréquents d'un corpus et donc du genre littéraire correspondant. Il s'agit des motifs extraits pendant la première étape de notre approche (cf. section 2.2). Les motifs émergents sont calculés en utilisant l'équation précédente (D_1 correspond alors à l'ensemble des données du genre et D_2 à l'union des ensembles de données des autres genres).

3. Evaluation expérimentale

Dans cette section, nous présentons les résultats de notre évaluation expérimentale. Nous décrivons tout d'abord, dans la section 3.1, les corpus utilisés ainsi que le réglage des différents paramètres utilisés lors de l'extraction des motifs émergents. Nous présentons ensuite une analyse des motifs extraits en deux parties : d'un point de vue quantitatif (section 3.2) et d'un point de vue linguistique pour la stylistique (section 3.3).

3.1. Protocole expérimentale

3.1.1. Corpus : pré-traitement et étiquetage

Nous avons créé trois corpus correspondant à différents genres de texte : *Poésie*, *Correspondance* et *Roman*. Pour construire chaque corpus, nous avons sélectionné tous les textes de la période

1800-1900 correspondant au genre de texte correspondant ; les textes proviennent de ressources françaises de la base Frantext du CNRTL[4]. Ces trois corpus ont ensuite été pré-traités. Le pré-traitement consiste à mettre les mots en minuscule et à séparer le texte en séquences selon les ponctuations de l'ensemble suivant : {*« . », « ? », « ! », « ... », « ; », « : », « , »*}. Le tableau 2 donnent les informations principales pour chacun des trois corpus pré-traités : le nombre d'auteurs, d'œuvres, de séquences et de mots.

Corpus	Nb. auteurs	Nb. œuvres	Nb. séquences	Nb. mots
Poésie	27	48	151 116	1 167 422
Correspondance	5	9	234 997	1 562 543
Roman	37	52	663 860	5 105 240

Tableau 2 : Caractéristiques des corpus Poésie, Correspondance et Roman

Après avoir été pré-traités, les corpus ont été étiquetés en utilisant *Cordial*[5]. Les mots des corpus sont ainsi étiquetés avec leur forme, leur lemme et leur catégorie morpho-syntaxique. Après des premières expérimentations, il s'est avéré que les catégories morpho-syntaxiques données par Cordial étaient trop détaillées ; nous les avons ainsi post-traitées afin de réduire leur nombre (ce qui a également pour conséquence de réduire le nombre de motifs extraits et donc à analyser). Les catégories trop spécifiques ont été regroupées en des catégories plus générales. Par exemple, la catégorie des adjectifs était initialement décomposée en 16 catégories (selon le genre du mot, son nombre ou encore la présence d'un *h* muet au début de celui-ci). Les catégories suivantes ont ainsi été créées (pour remplacer leurs sous-catégories correspondantes) : adjectifs (ADJ), déterminants (DET), noms communs (NC), noms propres (NP), pronoms démonstratifs (PD), pronoms relatifs (PR), pronoms indéfinis (PI) et participes passés (VPARPRES). Les catégories correspondant aux pronoms personnels ont ensuite été décomposées en deux étiquettes : une pour la catégorie des pronoms personnels (PPER) et une pour la personne du pronom (par exemple, 1S pour la première personne du singulier). De plus, les catégories correspondant aux verbes ont été décomposées en trois étiquettes : une pour la catégorie des verbes (V), une pour le mode du verbe (par exemple, INDP pour le présent de l'indicatif) et une pour la personne (avec les mêmes étiquettes que pour les pronoms personnels). L'ensemble final contient ainsi 35 étiquettes au lieu des 133 étiquettes initiales. En utilisant ces nouvelles étiquettes, la phrase *« une rose qu'on respire »* est alors traduite en <(*une un DET*) (*rose rose NC*) (*qu'que PR*) (*on on PPER 3S*) (*respire respirer V PRES 3S*)>.

3.1.2. Paramètres pour la fouille de données séquentielles d'items

Nous considérons tout d'abord des séquences constituées uniquement des formes des mots. Pour effectuer la tâche de fouille sur les trois corpus, nous utilisons dmt4 (Nanni et Rigotti, 2007) ; cela nous permet de définir plusieurs contraintes sur les motifs d'items extraits : leur longueur, leur fréquence (en fixant le seuil *minsup*) et la contrainte *gap* (en choisissant les valeurs [*M,N*]). Nous avons fixé la longueur des motifs à 2 mots minimum et 20 mots maximum. Nous avons

4 Centre National des Ressources Textuelles et Linguistiques : www.cntrl.fr.

5 L'étiqueteur a été développé par la société Synapse Développement (www.synapse-fr.com).

choisi la valeur de *minsup* empiriquement afin d'avoir un compromis entre extraire des motifs intéressants avec un support faible (et donc donner une valeur faible à *minsup*) et ne pas obtenir trop de motifs (ce qui revient à donner une valeur élevée à *minsup*). Comme les trois corpus considérés ont des tailles différentes (*e.g. Roman* est cinq fois plus gros que *Poésie*), nous avons choisi un seuil relatif et nous l'avons fixé à 0,001 % ; les seuils absolus correspondants sont les suivants : 16 pour *Poésie*, 12 pour *Correspondance* et 51 pour *Roman*. Cela signifie que seuls les motifs apparaissant dans plus de 16 séquences sont conservés pour *Poésie*, par exemple. En ce qui concerne la contrainte *gap*, nous avons considéré différentes valeurs dans les expérimentations menées (cf. section 2.2) : [*1,1*], [*1,2*], [*1,3*] et [*1,5*]. On remarquera que fixer la contrainte *gap* à [*1,1*] revient à considérer des motifs de type *n*-gramme. En effet, les motifs extraits sous cette contrainte correspondent à des sous-séquences de mots consécutifs du corpus.

3.1.3. Paramètres pour la fouille de données séquentielles d'itemsets

Nous considérons ensuite des séquences d'itemsets dans lesquelles chaque itemset représente un mot avec sa forme, son lemme et sa catégorie morpho-syntaxique. Pour fouiller ces séquences d'itemsets, nous avons choisi CloSpan (Yan *et al.*, 2003) qui extrait des motifs clos. CloSpan ne permet de définir qu'une seule contrainte : le seuil *minsup*. Nous avons choisi sa valeur empiriquement et avons ainsi fixé *minsup* à 0,15%. Comme il n'est pas possible de définir de contrainte *gap* dans CloSpan, nous avons dû donner une valeur plus élevée que précédemment à *minsup* afin de limiter le nombre total de motifs générés et donc le temps de calcul. Néanmoins, certains motifs intéressants risquent de ne pas être extraits car leur support sera trop faible (*e.g.* le support absolu pour *Roman* est 1 000).

3.1.4. Sélection de motifs émergents

Pour sélectionner les motifs émergents des corpus, nous fixons le seuil ρ à 1,001 afin qu'il soit juste au-dessus de 1. Ce seuil est utilisé à la fois pour les motifs d'items mais également pour les motifs d'itemsets.

3.2. Analyse quantitative des motifs

Dans cette sous-section, nous présentons les résultats quantitatifs sur les motifs d'items et sur les motifs d'itemsets. Le nombre de motifs extraits étant important, cette analyse quantitative nous permet de sélectionner les motifs qui sont ensuite analysés, d'un point de vue linguistique, pour une tâche de stylistique (cf. section 3.3).

Le tableau 3 donne le nombre de motifs extraits à partir des trois corpus, en considérant deux types de motifs : des motifs d'items (avec différentes valeurs pour la contrainte *gap*) et des motifs d'itemsets. Le ratio de motifs émergents est également donné pour chaque type de motifs.

Corpus	Motifs d'items avec contrainte gap				Motifs d'itemsets
	[1,1]	[1,2]	[1,3]	[1,5]	
Poésie	18 816 (30,7 %)	37 933 (27,0 %)	55 762 (24,3 %)	86 901 (22,6 %)	2 245 326 (11,4 %)
Correspondance	16 936 (50,2 %)	36 849 (50,7 %)	56 755 (50,4 %)	96 549 (50,0 %)	10 128 288 (57,4 %)
Roman	78 210 (6,1 %)	175 645 (5,3 %)	282 967 (4,9 %)	512 647 (4,6 %)	11 681 913 (71,2 %)
Total	113 962 (16,7 %)	250 427 (15,3 %)	395 484 (14,2 %)	696 097 (13,2 %)	24 055 527 (59,8 %)

Tableau 3: Nombre de motifs des corpus et ratio des motifs émergents (entre parenthèses)

Ainsi, parmi les 18 816 motifs extraits de *Poésie* en fixant la contrainte *gap* à [1,1], 30,7 % de ces motifs sont des motifs émergents (cela représente 5 776 motifs). Nous constatons tout d'abord qu'en considérant uniquement les motifs émergents, le nombre de motifs à analyser diminue énormément. De plus, cela permet de se focaliser sur les motifs les plus spécifiques pour un genre de texte donné. Ainsi, dans la suite des analyses, nous nous intéressons uniquement aux motifs émergents. Nous constatons également qu'en augmentant la contrainte *gap*, le pourcentage de motifs émergents d'items a tendance à diminuer : cela signifie que les motifs additionnels qui sont extraits sont plutôt des motifs non-spécifiques des genres de texte étudiés. Pour l'analyse stylistique qui suit en section 3.3, nous fixons donc la contrainte *gap* à [1,3], ce qui donne un bon compromis entre le nombre total de motifs d'items extraits et leur pertinence. Nous pouvons enfin remarquer qu'il y a beaucoup plus de motifs d'itemsets extraits que de motifs d'items (en moyenne, plus de 50 fois plus).

Nous étudions ensuite la distribution des motifs émergents selon leur longueur. La courbe représentée par la figure 2 donne le nombre relatif de motifs en fonction de leur longueur, pour les motifs d'items (la longueur est donnée en nombre d'items) et pour les motifs d'itemsets (la longueur est alors donnée en nombre d'itemsets). Nous constatons alors que la majorité des motifs d'items comportent entre 2 et 5 items alors que la plupart des motifs d'itemsets comportent entre 4 et 11 itemsets. Les motifs d'itemsets représentent ainsi des patrons linguistiques plus longs. De plus, il y a beaucoup de motifs d'items de longueur 2 mais ils sont moins informatifs - d'un point de vue linguistique - que les motifs plus longs : nous ne considérons que les motifs de longueur au moins 3 lors de l'analyse stylistique.

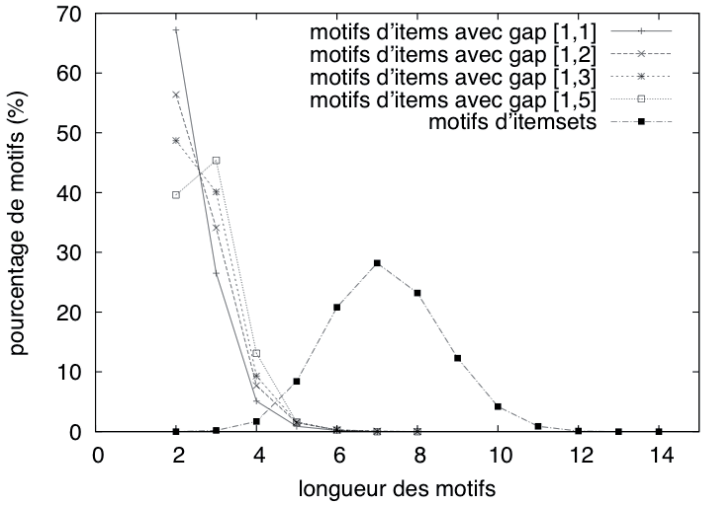

Figure 2 : Distribution des motifs émergents selon leur longueur

Nous étudions enfin la distribution des motifs émergents selon leur taux de croissance. La courbe représentée par la figure 3 donne le nombre relatif cumulé de motifs émergents en fonction de leur taux de croissance, pour l'ensemble des trois corpus.

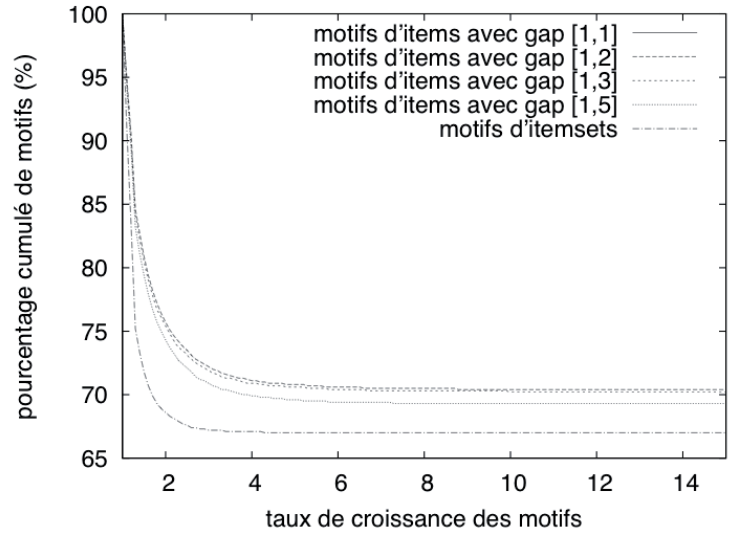

Figure 3 : Distribution des motifs émergents selon leur taux de croissance

Cela signifie que, par exemple, 67,1 % des motifs émergents d'itemsets ont un taux de croissance supérieur à 4 ou encore que 67,1 % des motifs émergents d'itemsets sont conservés en fixant un seuil $\rho=4$ sur les taux de croissance. Nous pouvons constater que la plupart des motifs émergents ont un taux de croissance infini puisque le nombre cumulé de motifs émergents reste constant pour les taux de croissance supérieur à 10. Cela signifie que la plupart des motifs émergents n'apparaissent que dans un seul genre de texte (et pas une seule fois dans les autres genres considérés). Lors de l'analyse stylistique, nous ne considérons ainsi que les motifs émergents – d'items et d'itemsets – ayant un taux de croissance infini.

Pour finir, seuls les motifs d'itemsets contenant à la fois des catégories morpho-syntaxiques et des formes de mots ou des lemmes sont considérés lors de l'analyse stylistique. Les motifs contenant uniquement soit des catégories morpho-syntaxiques, soit des formes de mots ou des lemmes sont ainsi supprimés car les premiers sont très généraux et les seconds sont trops spécifiques. En fait, la grande majorité des motifs d'itemsets contiennent à la fois des catégories morpho-syntaxiques et des mots puisque ces motifs représentent 93,5 % de l'ensemble des motifs d'itemsets. Cela rejoint également les conclusions de (Biber, 2009), à savoir que les motifs extraits contiennent généralement à la fois des éléments variables et des éléments fixes (les motifs uniquement composés de catégories morpho-syntaxiques ne contiennent ainsi que des éléments variables alors que les motifs uniquement composés de mots ne contiennent que des éléments fixes).

3.3. Analyse stylistique des motifs émergents

Dans cette sous-section, nous présentons une analyse stylistique s'appuyant sur les motifs émergents extraits précédemment. Nous nous focalisons plus particulièrement sur *Poésie*. Nous nous intéressons tout d'abord aux motifs émergents d'items. En les étudiant, nous avons ainsi découvert des motifs intéressants, caractéristiques de la poésie. Le tableau 4 donne des exemples de motifs identifiés ; dans ces motifs, le symbole * est utilisé pour représenter une contrainte *gap* d'un ou plusieurs mots[6].

Motifs d'items	Exemples
des * plus * que	il a **des** morsures **plus** venimeuses **que** celles de ta bouche **des** cailloux anguleux **plus** brillants **que** des marbres
on * et * on	une rose qu'**on** respire **et** qu'**on** jette sur des tombeaux divins qu'**on** brise **et** qu'**on** insulte ?
le/la/l' * qui * et * qui	**la** nuit **qui** m'oppresse **et qui** trouble mes yeux **le** grelot **qui** résonne **et** le troupeau **qui** bêle
le * du * qui * dans	**le** vent **du** soir **qui** meurt **dans** le feuillage **le** bruit **du** vieux **qui** bêche **dans** la nuit
est * un * qui	**est**-ce **un** goéland **qui** bat de l'aile ? ta grâce **est** comme **un** luth **qui** vibre au fond du bois

Tableau 4: Exemples de motifs d'items caractéristiques de Poésie

De plus, chaque motif est illustré par des exemples de séquences de *Poésie* correspondant audit motif. Les motifs présentés permettent d'observer des structures grammaticales relativement indépendantes du lexique. En effet, les éléments fixes de ces motifs sont des mots grammaticaux alors que les éléments variables (c'est-à-dire ceux qui remplissent les *gaps*) sont généralement des mots lexicaux (des noms, des verbes ou encore des adjectifs, par exemple). Nous illustrons

6 Le symbole * correspond au symbole *?* utilisé dans (Renouf et Sinclair, 1991). Notons que le symbole * est également utilisé dans (Biber, 2009) mais il représente un lexème variable alors que, dans notre approche, ce symbole représente un *gap* d'un ou plusieurs mots.

également l'intérêt des contraintes de *gap* dont la valeur est donnée sous la forme d'un intervalle. En effet, le motif *des * plus * que* permet d'identifier, entre autres, deux séquences (cf. tableau 4) qui remplissent le premier *gap* avec un nombre de mots différent : dans la première, le mot *morsures* remplit ce *gap* alors qu'il correspond à *cailloux anguleux* dans la deuxième séquence. Cela montre ainsi la capacité de généralisation des motifs d'items avec contrainte de *gap* (*e.g.*, par rapport à des motifs de type *n*-grammes).

La table 5 illustre la correspondance entre les motifs d'items présentés dans la table 4 et les motifs d'itemsets correspondants.

Motifs d'items	Motifs d'itemsets
des * plus * que	des N plus ADJ que
on * et * on	N qu'on V et qu'on V
le/la/l' * qui * et * qui	le N qui V et (qui) V, le N qui V et le N qui V
le * du * qui * dans	le N du N qui V dans le N
est * un * qui	est-ce un N qui V, est comme un N qui V

Tableau 5 : Correspondance entre motifs d'items et motifs d'itemsets

Nous constatons que plusieurs motifs d'itemsets peuvent correspondre à un même motif d'items. De plus, les motifs d'itemsets extraits permettent d'obtenir les catégories morpho-syntaxiques des éléments variables. Ainsi, dans le cadre de l'étude stylistique d'un genre de texte, le travail des linguistes consiste à sélectionner les motifs pertinents parmi les motifs émergents d'itemsets qui sont extraits automatiquement : cela leur permet d'obtenir directement des patrons grammaticaux caractéristiques du genre de texte considéré.

Les patrons grammaticaux considérés ici correspondent en fait à des *cadres collocationnels* au sens de (Renouf et Sinclair, 1991), c'est-à-dire à des collocations sur des unités grammaticales et non sur des unités lexicales. Cependant, contrairement à eux, nous ne choisissons pas *a priori* les patrons étudiés par la suite mais nous les découvrons automatiquement à partir des corpus utilisés. Nous pouvons également rapprocher nos travaux de ceux de (Biber, 2009) - qui travaille également sur les cadres collocationnels - avec certaines différences. En effet, notre approche permet d'extraire directement des motifs d'items avec des *gaps* ainsi que des motifs d'itemsets (correspondant directement à des patrons grammaticaux) alors que Biber extrait d'abord des séquences fréquentes, à partir de corpus, séquences qu'il analyse ensuite une à une afin d'identifier les éléments variables et les éléments fixes pour pouvoir enfin construire des patrons type qu'il étudie par la suite. La notion de cadres collocationnels a ainsi fait l'objet d'un certain nombre de travaux en linguistique anglaise de corpus mais n'a cependant pas eu d'écho en linguistique française[7]. Or, l'analyse des cadres collocationnels peut être pleine d'enseignement si elle est associée à une véritable théorie de l'usage qui considère que les formes grammaticales émergent de l'usage linguistique (c'est-à-dire des approches *corpus-driven*) et ne sont nullement le produit de règles intégrées (c'est-à-dire des approches *corpus-based*). C'est

7 Nous pouvons toutefois mentionner le travail de (Longrée *et al.*, 2008) sur l'utilisation de motifs mais l'approche employée est plutôt *corpus-based*.

pourquoi il est intéressant de disposer d'approches permettant d'extraire automatiquement de tels cadres collocationnels, ce qui est le cas de notre approche.

4. Conclusion

Dans cet article, nous avons présenté une première étude sur l'utilisation de techniques de fouille de données en proposant une méthodologie reposant sur l'extraction de motifs séquentiels et appliquée dans le cadre de la stylistique. Pour cela, nous avons considéré deux types de motifs séquentiels : des motifs d'items (les items correspondent aux formes des mots) et des motifs d'itemsets (s'appuyant sur les formes des mots, les lemmes et les catégories morpho-syntaxiques). De plus, nous nous sommes focalisés sur un type particulier de motifs séquentiels : les motifs émergents. Une analyse quantitative des motifs extraits de trois corpus (représentant différents genres de texte : *Poésie*, *Correspondance* et *Roman*) a tout d'abord montré que les motifs constituent un paradigme plus puissant que les *n*-grammes pour construire des patrons linguistiques. Une analyse linguistique a confirmé cela puisque des patrons grammaticaux caractéristiques de *Poésie* ont pu être directement identifiés à partir des motifs émergents extraits pour ce genre. Nous avons également comparé notre méthodologie à celle proposée dans (Biber, 2009) en montrant qu'elle permettait d'obtenir directement des motifs caractéristiques des genres de texte.

Cependant, de nombreux motifs sont extraits à l'aide de notre approche et les linguistes doivent identifier, parmi ces motifs, ceux qui sont réellement intéressants d'un point de vue analytique. Une des perspectives de nos travaux concerne la conception d'outils permettant de filtrer, d'ordonner et d'explorer les motifs extraits afin de faciliter leur analyse par les linguistes. De plus, afin de limiter le nombre de motifs extraits, il serait intéressant de définir de nouvelles contraintes utilisées directement lors du processus de fouille : par exemple, imposer la présence d'un certain type d'item dans les motifs extraits (*e.g.* un verbe).

Remerciements

Ce travail bénéficie du soutien de la région Basse-Normandie et de l'ANR (projet Hybride ANR-11-BS02-002).

Références

Agrawal R. et Srikant R. (1995). Mining sequential patterns. In *Proc. of ICDE'95*.

Biber D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, 14(3).

Dong G et Li J. (1999). Efficient mining of emerging patterns : Discovering trends and differences. In *Proc. of SIGKDD'99*.

Dong G et Pei J. (2007). *Sequence Data Mining*. Springer.

Hunston S. et Francis J. (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia.

Legallois D. et François J. (2006). Autour des grammaires de constructions et de patterns. *Cahiers du CRISCO*, 21.

Longrée D., Luong X. et Mellet S. (2008). Les motifs : un outil pour la caractériseration topologique des textes. In *Actes de JADT'08*.

Nanni M. et Rigotti C. (2007). Extracting trees of quantitative serial episodes. In *Proc. of KDID'07*.

Renouf A. et Sinclair J. (1991). Collocational Frameworks in English. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Longman.

Yan X., Han J. et Afshar R. (2003). Clospan: Mining closed sequential patterns in large databases. In *Proc. of SDM'03*.