Analyse des relations et des dynamiques de corpus de textes littéraires par extraction de motifs graduels
Analysis of the relations and dynamic of literary texts corpus by gradual itemsets extraction

Amal Oudni[1] Mohamed Amine Boukhaled [1] Gauvain Bourgne[1]
[1] Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris
{Amal.Oudni, Mohamed.Boukhaled, Gauvain.Bourgne}@lip6.fr

Résumé :
La stylistique computationnelle permet de déterminer l'ensemble des traits caractéristiques formels d'une œuvre littéraire. Cet article propose une approche de stylistique computationnelle qui se focalise sur les relations et les dynamiques entre les textes d'un corpus. Elle permet de mettre en évidence des corrélations de co-variation globales et locales entre les fréquences d'utilisation des catégories syntaxiques que nous illustrons sur une collection de textes écrits par Balzac. Pour ceci, nous proposons un outil d'analyse littéraire basé sur les méthodes de résumés linguistiques sous forme de motifs graduels et de motifs graduels caractérisés par des intervalles d'intérêt.

Mots-clés :
Résumés linguistiques, motifs graduels, caractérisation, stylistique computationnelle, catégories syntaxiques, analyse littéraire.

Abstract:
Computational stylistic allows to determine the formal characteristics of a literary work. This paper proposes a computational stylistic approach that focuses on the relationships and dynamics between the texts of a corpus. It allows to highlight the correlations of global and local co-variation between the frequencies of linguistic categories which we illustrate on a collection of texts written by Balzac. In this context, we propose a literary analysis tool based on linguistic summaries in the form of gradual itemsets and gradual itemsets characterized by interest intervals.

Keywords:
Linguistic summaries, graduals itemsets, characterization, computational stylistics, syntactic labels, literary analysis

1 Introduction

La stylistique computationnelle est un sous-domaine de la linguistique informatique qui s'intéresse à l'extraction et l'analyse des motifs de style caractérisant un type particulier de textes à l'aide de méthodes statistiques et automatiques [5]. Initialement, les techniques de stylistique computationnelle ont été utilisées pour étudier les questions relatives au style d'écriture [16]. Les premiers travaux se sont principalement intéressés aux traits lexicaux et grammaticaux. Récemment, des traits plus complexes ont été pris en compte [9]. Dans ce contexte, des méthodes de fouille de données ont été utilisées et leur intérêt a été mis en évidence pour l'analyse stylistique des grands textes [4].

Dans cet article, nous nous intéressons à l'analyse stylistique dynamique des textes. Plus précisément, nous considérons de nouveaux traits qui concernent les catégories syntaxiques caractéristiques d'un corpus de textes pour extraire des relations de co-variation entre ces dernières et mettre en évidence les dynamiques entre les textes de ce corpus. Nous proposons une approche basée sur des résumés linguistiques d'une forme particulière, appelée motifs graduels : on peut les illustrer par un exemple du type « plus la fréquence de citation de Molière par un auteur augmente, moins celle de Racine augmente » ou un exemple du type « plus la fréquence d'utilisation des noms communs chez le style d'un auteur augmente, plus elle augmente pour les déterminants » qui exprime l'existence d'une relation entre les différents marqueurs de style chez un auteur. Nous avons choisi, pour la mise en œuvre de cette approche, des textes classiques de la littérature française écrits par Balzac.

Les résumés linguistiques ont été d'abord in-

troduits dans une variante floue par [17], puis développés et présentés dans une forme computationnelle [13], en utilisant le calcul de Zadeh des propositions linguistiques quantifiées [18]. Ils sont généralement définis comme des textes constitués de quelques phrases en langage naturel [17, 13]. Ils peuvent extraire différents types d'informations, comme les co-occurrences dans le cas de règles d'association [1], les événements séquentiels dans le cas de motifs séquentiels [2] ou les tendances graduels entre les attributs décrivant les données dans le cas des motifs graduels [12, 7]. Dans cet article, nous nous concentrons sur ce dernier et ses enrichissements : nous considérons les résumés linguistiques de la forme « plus/moins A, plus/moins B » où A et B sont des attributs. Ils résument les données à travers leurs tendances internes, exprimées comme des corrélations entre valeurs d'attributs. De plus, nous considérons une variante enrichie, appelée motifs graduels caractérisés : ce sont des motifs graduels auquels est ajoutée une clause linguistiquement introduite par l'expression « *surtout si* » [15]. Ils peuvent être illustrés par l'exemple « plus l'âge augmente, plus le salaire augmente surtout si l'âge est entre $[30, 40]$ ». Dans cet article, nous proposons d'étendre cette méthode d'extraction de motifs graduels pour le traitement de données temporelles, en discutant le rôle spécifique de l'attribut temporel.

L'article est organisé comme suit : la section 2 rappelle les définitions formelles des motifs graduels et leur caractérisation, en décrivant leur principe, leur interprétation et les critères de qualité proposés pour leur évaluation. Elle présente de plus l'extension aux données temporelles. La section 3 présente les données textuelles sur lesquelles l'approche proposée est appliquée et la section 4 détaille et discute les résultats obtenus.

2 Approche proposée

Parmi la variété des approches possibles de résumés linguistiques, nous proposons d'utiliser la méthode basée sur l'extraction de motifs graduels pour deux raisons : ils sont adaptés à une caractérisation globale des relations entre différents textes d'un corpus et sont capables de résumer et de réduire les données représentées par des attributs numériques à des quantités d'informations interprétables, facilitant ainsi la compréhension de leur contenu par l'expert.

Cette section présente les deux formes de résumés linguistiques que nous extrayons à partir de données numériques, en précisant pour chacun d'eux son principe, son approche d'extraction et son critère de qualité dans la section 2.1. Nous décrivons les motifs graduels caractérisés dans la section 2.2. La section 2.3 présente l'application proposée aux données textuelles et enfin la section 2.4 souligne l'originalité de l'approche proposée.

2.1 Motifs graduels

Notations et définition. Pour définir formellement les motifs graduels, on a besoin d'abord de définir les items graduels. On note $\mathcal{D}$ un ensemble de données constitué de n objets décrits par m attributs numériques.

Un *item graduel* A^* est un couple constitué d'un attribut A et d'une variation, notée $* \in \{\geq, \leq\}$: $A^{\geq}$ et $A^{\leq}$ représentent le fait que les valeurs d'attribut augmentent (dans le cas de $\geq$) ou diminuent (dans le cas de $\leq$). Un *motif graduel* est un ensemble d'items graduels, interprété comme leur conjonction. A un motif $M = \{(A_j, *_j), j = 1..k\}$, on associe sa longueur, k, définie comme le nombre d'attributs qu'il implique, et le pré-ordre induit $\preceq_I$ défini sur $\mathcal{D}^2$ tel que $o \preceq_I o'$ ssi $\forall\ j \in [1, k]\ A_j(o) *_j A_j(o')$ où $A_j(o)$ représente la valeur de l'attribut A_j pour l'objet o.

Il est important de noter la différence entre ces motifs graduels et les règles d'association floues [8, 11] : ces dernières sont une généralisation floue des règles d'association classiques. L'interprétation de ces règles s'applique à chaque objet individuellement : elle

considère qu'une présence (floue) implique au sens flou une présence (floue), et que chaque objet a une contribution individuelle avec son propre degré d'appartenance. Au contraire, la gradualité des motifs graduels exprime une tendance globale à travers l'ensemble de données : elle s'applique à un sous-ensemble d'objets de manière transversale.

Plusieurs interprétations de ces motifs graduels ont été proposées, sous la forme de régression [12], de corrélation d'ordres induits [3, 14] ou d'identification de sous-ensembles d'objets compatibles [6, 7]. Chacune de ces interprétations conduit à la définition d'un support et à des méthodes d'identification automatique de motifs graduels.

Dans cet article, nous considérons l'interprétation comme contrainte de covariation par identification de sous-ensembles d'objets compatibles [6, 7] : elle consiste à identifier un sous-ensemble, appelé *chemin*, d'objets D de $\mathcal{D}$ qui peuvent être ordonnés de façon à ce que tous les couples de D vérifient le pré-ordre induit. Ainsi pour un motif M, $D = \{o_1, ..., o_p\} \subseteq \mathcal{D}$ est un chemin si et seulement s'il existe une permutation π telle que $\forall l \in [1, p-1]$, $o_{\pi_l} \preceq_M o_{\pi_{l+1}}$. Les motifs graduels dépendent ainsi de l'ordre induit par les valeurs d'attribut et non des valeurs elles-mêmes.

Un tel chemin est dit *complet* si aucun objet ne peut lui être ajouté sans violer la contrainte d'ordre imposée par M. On note $\mathcal{L}(M)$ l'ensemble des chemins complets associés à M. On appelle *chemin maximal* un chemin complet de longueur maximale.

Critère de qualité. Le support graduel de M, $SG_{\mathcal{D}}(M)$, est défini comme la longueur de son chemin complet maximal normalisée par le nombre total d'objets :

$$SG_{\mathcal{D}}(M) = \frac{1}{|\mathcal{D}|} \max_{D \in \mathcal{L}(M)} |D| \qquad (1)$$

2.2 Motifs graduels caractérisés

Cette section rappelle le principe des motifs graduels caractérisés [15] ainsi que leur formalisation, en illustrant sur un exemple.

Principe. L'objectif de la caractérisation d'un motif graduel M est d'identifier un ensemble d'attributs $J \subset M$ et une région R conduisant à l'expression linguistique « surtout si $J \in R$ » où la validité du motif M doit augmenter (voir [15] pour plus de détails). La région R induit une restriction $\mathcal{D}'$ de l'ensemble de données $\mathcal{D}$, en considérant uniquement les données satisfaisant la contrainte de valeur exprimée par la région R. Le principe consiste à maximiser à la fois le support du motif considéré M sur $\mathcal{D}'$ et le nombre d'objets dans $\mathcal{D}'$ [15].

La figure 1 représente un ensemble de données décrit par deux attributs pour lesquels le motif graduel $M = A^{\geq}B^{\geq}$ est supporté par le chemin représenté par •. Son support graduel est $14/23 = 60\%$. Or, on peut observer que la co-variation entre A et B est particulièrement valable dans la partie centrale du graphique, tandis que les données qui ne sont pas en accord avec le motif se trouvent surtout dans les parties où A prend des valeurs basses ou élevées. Plus précisément, si les données sont limitées aux objets pour lesquels A prend des valeurs dans l'intervalle $[32; 53]$, graphiquement délimité par les lignes verticales sur la figure 1, alors le support du motif augmente à $9/10 = 90\%$. Ceci motive l'extraction du motif graduel caractérisé $A^{\geq}B^{\geq}$ *; surtout si* $A \in [32; 53]$.

La caractérisation des motifs graduels est interprétée comme un accroissement de la validité, quand les données sont restreintes aux objets satisfaisant la clause de caractérisation. Cependant, pour qu'elle soit informative, une telle caractérisation ne doit pas limiter les données drastiquement : il est facile d'atteindre 100% de support, par exemple en restreignant les données à un unique couple de points satisfaisant l'ordre induit par le motif graduel

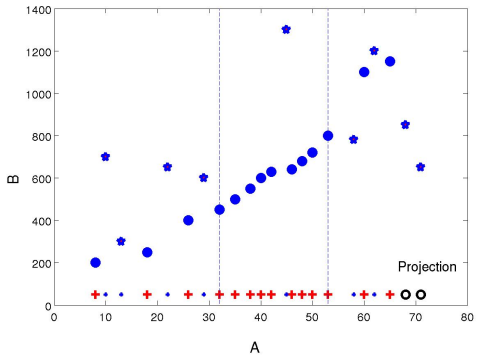

Figure 1 – Exemple de caractérisation d'un motif graduel, conduisant à « plus A, plus B, surtout si $A \in [32; 53]$ »

considéré. Pourtant, la caractérisation résultante serait trop spécifique et non pertinente. Selon le même principe, dans l'exemple précédent, restreindre les données à l'intervalle plus petit $[32; 42]$ augmente le support à 100%, mais conduit à une caractérisation trop spécifique.

Le principe des motifs graduels caractérisés est donc de trouver un compromis entre un support élevé et un nombre élevé d'objets lors de la restriction des données à un sous-ensemble de données définie par les intervalles considérés.

Formalisation. Le principe illustré ci-dessus peut être formalisé comme suit : pour un motif graduel M, la caractérisation est notée comme « M, *surtout si* $J \in R$ ». L'ensemble d'intervalles R définit une région qui induit une restriction $\mathcal{D}'$ de l'ensemble de données $\mathcal{D}$, en considérant uniquement les données satisfaisant la contrainte de valeur exprimée par R.

Le principe exposé dans la section précédente consiste alors à maximiser à la fois le support du motif considéré M sur les données restreintes $\mathcal{D}'$, et le nombre d'objets satisfaisant les contraintes d'ordre sur $\mathcal{D}'$, c'est-à-dire

$$\max_R |\mathcal{D}'| \tag{2}$$

$$\max_R SG_{\mathcal{D}'}(M) \tag{3}$$

où SG représente le support graduel rappelé dans l'équation (1).

Un compromis doit être trouvé entre ces deux objectifs qui peuvent être contradictoires : en effet, une augmentation de la taille du sous-ensemble $\mathcal{D}'$ peut conduire à la diminution de la proportion d'objets compatibles avec l'ordre induit par le motif considéré.

Approche de caractérisation. Il a été proposé dans [15] de décomposer la tâche d'identification des attributs caractéristiques et de leurs intervalles d'intérêt associés, en considérant successivement chaque attribut composant le motif graduel considéré M ainsi que chaque chemin supportant M : le calcul du support graduel restreint $SG_{\mathcal{D}'}(I)$ et l'intervalle d'intérêt sont basés sur la restriction des chemins associés à M. Une fois qu'une restriction est identifiée pour chaque chemin, un post-traitement est effectué sur les différents chemins : les restrictions sont combinées pour sélectionner les limites optimales qui correspondent aux limites de la plus grande restriction identifiée.

Pour cela, une méthode basée sur l'utilisation d'outil de morphologie mathématique a été proposée [15]. Étant donné un motif graduel M, un chemin maximal D et un attribut A pour lequel un intervalle d'intérêt est recherché, l'information de chemin est codée par un processus de transcription en une séquence de symboles $\{+, -\}$, où $+$ représente un objet vérifiant M et $-$ celui qui ne le vérifie pas.

Ainsi, une succession de symboles de $+$ et de $-$ est obtenue. On cherche ensuite la région où M est vrai, ce qui est équivalent à la recherche de la plus longue séquence de $+$. Dans l'approche proposée, cette séquence de $+$ est prolongée, en incorporant certains symboles $-$, de manière à augmenter la taille de l'ensemble de données restreint représenté par la plus longue séquence de $+$ sans détériorer la proportion de $+$ dans la séquence considérée. Ceci est réalisé par un filtre morphologique, appliquée à la succession de symboles obtenue après la transcription des

données, comme justifiée et discuté dans [15].

La partie inférieure de la figure 1 indique la transcription obtenue pour l'exemple illustratif. Plus de détails sur le processus de transcription et l'étape de filtrage morphologique sont donnés dans [15].

La qualité des motifs graduels caractérisées est évaluée par le support graduel caractérisé, $SG_{\mathcal{D}'}$, défini dans l'équation 3.

2.3 Extension proposée pour les données temporelles

Cette section décrit l'extension de l'approche d'extraction de motifs graduels caractérisés pour les données temporelles. Nous proposons une représentation paramétrique où le temps est le paramètre de tous les attributs de la base de données. Chaque point ne représente pas un vecteur de p attributs numériques $x = (A_j)_{j=1...p}$, mais un vecteur de valeurs d'attribut à chaque date, noté comme $x = (t, A_1(t), \ldots, A_p(t))$ pour chaque temps $t \in [0, T]$.

Nous proposons également de conserver le temps comme un attribut comme les autres, ce qui conduit à une nouvelle expression linguistique des motifs graduels adaptée. Deux types d'items graduels peuvent être identifiés : le premier considère le temps comme un attribut, conduisant à des items graduels de la forme $M = \{(t, \leq)(A_{*_i}, *_i),\ i = 1...k\}$. Pour ce type, l'expression classique « plus le temps augmente » n'est pas pertinente. Par conséquent, nous proposons de remplacer le résumé linguistique par la forme « A_i a tendance à la hausse » (si $*_i = \geq$) et « A_i a tendance à la baisse » (si $*_i = \leq$). L'exemple « l'utilisation des interjections dans les textes de Balzac a tendance à la hausse, surtout entre $[1832, 1847]$ » illustre un tel cas.

Le second type ne considère pas t comme un item, mais il considère une donnée pour chaque date indépendamment : il applique les motifs graduels classiques aux vecteurs $(A_1(t), \ldots, A_p(t))\ \forall\ t \in [0, T]$. Ce cas peut être illustré par l'exemple « plus l'utilisation des noms communs élevée, plus l'utilisation des déterminants est élevée ».

2.4 Originalité

La principale différence entre les approches existantes pour l'analyse stylistique et l'approche proposée dans cet article provient de l'exploitation des méthodes basées sur les motifs graduels. En effet, celles-ci se focalisent sur l'interaction entre les formes linguistiques d'une part, et sur leurs évolutions et variations au fils du temps d'une autre part. Ces approches permettent donc une analyse dynamique : au lieu de se baser uniquement sur la quantification et l'analyse statique des formes linguistiques pour la génération d'une sorte de propriétés stylistiques génériques, les motifs graduels permettent d'introduire un autre niveau d'analyse.

3 Application aux textes de la Comédie Humaine de Balzac

3.1 Données considérées

Le corpus utilisé pour cette analyse est constitué de 81 textes tirés de la Comédie Humaine de Balzac. Cette dernière est une collection de textes de natures différentes (romans, nouvelles, contes et essais) dont l'écriture s'étale sur plusieurs années. Le fait que ces textes soient écrits sur une plage de temps considérable (de 1827 à 1848) permet de tracer le changement de style de Balzac au fil du temps et d'avoir une étude dynamique qui s'intéresse à la variation et l'évolution de la forme linguistique des écrits de cet auteur.

Le choix de ce corpus est à la fois motivé par le fait que tous ces textes sont écrits par le même auteur, ce qui permet d'avoir une étude moins biaisée et bien focalisée sur le style d'écriture propre à lui, et par notre intérêt particulier pour la littérature française classique du $19^{\text{ème}}$ siècle.

3.2 Représentation numérique des données

Il faut noter que les approches de motifs graduels s'appliquent à des données numériques ou floues. Une représentation numérique des données textuelles est donc nécessaire. Pour cela, ce traitement a été réalisé en trois étapes : tous les textes ont été d'abord segmentés en un ensemble de phrases, puis le corpus a été analysé syntaxiquement en utilisant TreeTagger [10]. Chaque texte est ensuite représenté par un vecteur R_k de fréquences d'apparition des catégories syntaxiques dans ce texte, telles que $R_k = r_1, r_2, \ldots, r_k$ où r_i avec $1 \leq i \leq 14$ représente la fréquence d'apparition d'une des 14 catégories syntaxiques utilisées dans cette analyse et qui sont listées dans la section ci-dessous. Pour avoir une étude moins biaisée vers les textes ayant une plus grande taille, une normalisation par rapport à cette dernière du vecteur de fréquences représentant chaque texte composant ce corpus a été effectuée.

3.3 Attributs considérés

Chaque texte est décrit par un attribut temporel qui représente la date, un attribut qui représente la taille des textes et $p = 14$ attributs numériques basiques qui correspondent aux fréquences des catégories syntaxiques suivantes : abréviation, adjectif, adverbe, déterminant, interjection, conjonction, nom propre, nom commun, valeur numérique, pronom, préposition, ponctuation, ponctuation de fin de phrase et verbe.

3.4 Exemple illustratif

Pour chercher les corrélations entre valeurs d'attributs, nous transformons les données représentées sous forme d'évolutions temporelles, comme illustré sur la figure 2, en des données qui peuvent être représentées sous forme de nuage de points, où chaque point représente le couple composé de la fréquence des noms communs et celle des déterminants utilisés pour chaque date donnée, comme

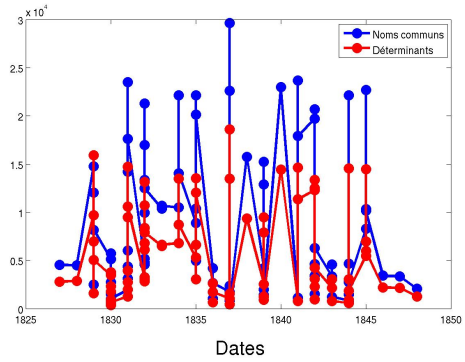

Figure 2 – Évolution temporelle des noms communs et déterminants

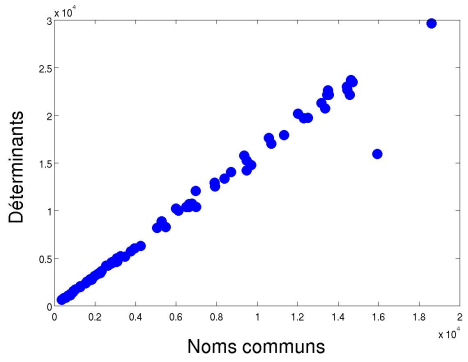

Figure 3 – Nuage de points des noms communs et des déterminants

illustré sur la figure 3. Pour cette paire d'attributs, une tendance globalement positive peut être observée et qui conduit au motif graduel « plus l'utilisation des noms communs augmente, plus l'utilisation des déterminants augmente ».

4 Expérimentations

Cette section décrit le protocole expérimental et présente les résultats obtenus à partir du corpus de textes présenté dans la section 3.

4.1 Protocole expérimental

Nous identifions des motifs dont $SG_{\mathcal{D}}$, tel que défini dans l'équation (1), est supérieur au seuil $s = 18\%$, des motifs graduels caractérisés

dont le support de caractérisation $SG_{\mathcal{D}'}$ est supérieur au seuil $s_c = 60\%$ et tels que l'intervalle caractéristique extrait incorpore au moins 5 textes.

4.2 Résultats obtenus

Nous obtenons sur l'extrait considéré de la Comédie Humaine deux types de motifs graduels. On observe d'abord des motifs graduels n'incluant pas d'attribut temporel, mettant en relations des attributs non temporels, tels que

— plus Balzac utilise des conjonctions, plus il utilise des verbes : $SG_{\mathcal{D}} = 100\%$

— plus Balzac utilise des noms, plus il utilise des déterminants : $SG_{\mathcal{D}} = 30\%$

Ces motifs semblent naturels dans la langue française. En effet, une conjonction s'accompagne toujours d'une ou deux propositions qui contiennent a priori un verbe, et un déterminant introduit en général un nom. Nous obtenons aussi des motifs comme

— plus Balzac utilise des noms propres, moins il utilise des pronoms : $SG_{\mathcal{D}} = 26.5\%$

— plus le texte est long, plus Balzac utilise des ponctuations non terminales : $SG_{\mathcal{D}} = 22.5\%$

Ces motifs semblent moins directement découler des règles du français. Ils peuvent être considérés comme des marqueurs stylistiques, s'appliquant à l'ensemble de l'œuvre en mettant en évidence des corrélations d'emploi traduisant des habitudes d'écritures. Ces marqueurs peuvent selon les cas traduire la régularité de la langue française en général ou celle du style de Balzac.

On observe également des motifs graduels dépendant de l'attribut temporel. Ceux-ci peuvent s'appliquer sur l'ensemble du texte, indiquant alors une évolution progressive de l'écriture, ou être caractérisés par une période, mettant en évidence par exemple une phase d'évolution ou de transition. Sur la Comédie Humaine, nous obtenons ainsi comme exemple du premier type, le motif

— la taille des textes montre une tendance à la hausse : $SG_{\mathcal{D}} = 25\%$

Comme exemple de motif graduel caractérisé, on peut citer

— l'utilisation des déterminants montre une tendance à la hausse, surtout entre janvier 1827 et décembre 1829 : $SG_{\mathcal{D}} = 18.75\%, SG_{\mathcal{D}'} = 86\%$

Ce motif indique une évolution sur le début de son œuvre. Ce type de motifs est intéressant en particulier pour étudier la dynamique d'un corpus et repérer des tendances ou des périodes de changement.

Ces résultats préliminaires donnent un nouvel outil d'analyse littéraire d'un corpus, original puisqu'il se centre sur les relations et les dynamiques entre les textes d'un corpus. Il permet de mettre en évidence des régularités d'usage, des tendances globales ou locales en mettant en évidence des périodes spécifiques.

5 Conclusion

Dans cet article, nous avons proposé une étude de stylistique computationnelle dynamique en utilisant une collection de textes écrits par Balzac sur plusieurs années. L'approche proposée est basée sur des méthodes de résumés linguistiques sous forme de motifs graduels classiques et caractérisés. Elle permet de mettre en évidence des corrélations de co-variations, globales ou locales, des fréquences d'utilisations de catégories syntaxiques.

Nous avons proposé d'étendre l'approche de caractérisation de motifs graduels aux données temporelles afin d'identifier des périodes spécifiques en mettant en évidence les phases d'évolution ou de transition.

Une première perspective de ce travail consiste à mettre en relations les résultats obtenus avec un corpus linguistique plus large et avec des avis d'experts afin de déterminer quels motifs sont spécifiques à Balzac, ou tout au moins plus saillants dans son œuvre.

Les perspectives de ce travail incluent également l'étude approfondie du choix des attributs numériques : les attributs syntaxiques simples comme proposé dans cet article ou des attributs plus complexes comme la fréquence de patrons syntaxiques. Une autre perspective consiste à utiliser les critères lexicaux. En effet, il serait intéressant d'utiliser les nombres d'occurrences de certains mots ou les notions de proximité sémantique ou encore les méta-données comme par exemple le temps de rédaction s'il est connu ou les mesures de réceptions du texte tels que le nombre d'exemplaires publiés ou vendus, de réédition, etc. Cela soulève néanmoins des problèmes concernant la récolte de toutes ces informations.

Références

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. of the Int. Conf. on VLDB*, pages 487–499, 1994.

[2] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. of the Int. Conf. on Data Engineering*, pages 3–14, 1995.

[3] F. Berzal, J. C. Cubero, D. Sanchez, M. A. V. Miranda, and J. M. Serrano. An alternative approach to discover gradual dependencies. *Inter. Journ. of Uncertainty, Fuzziness and Knowledge-Based Systems*, (5) :559–570, 2007.

[4] N. Béchet, P. Cellier, T. Charnois, and B. Crémilleux. Discovering linguistic patterns using sequence mining. In *Computational Linguistics and Intelligent Text Processing*, pages 154–165, 2012.

[5] H. Craig. Stylistic analysis and authorship studies. In *A companion to digital humanities*, pages 233–334, 2004.

[6] L. Di Jorio, A. Laurent, and M. Teisseire. Fast extraction of gradual association rules : a heuristic based method. In *Proc. of the ICSTST*, pages 205–210, 2008.

[7] L. Di Jorio, A. Laurent, and M. Teisseire. Mining frequent gradual itemsets from large databases. In *Advances in IDA*, pages 297–308, 2009.

[8] D. Dubois and H. Prade. Gradual inference rules in approximate reasoning. *Information Sciences*, 61(1–2) :103–122, 1992.

[9] V. W. Feng and G. Hirst. Patterns of local discourse coherence as a feature for authorship attribution. *Literary and Linguistic Computing*, 29(2) :191–198, 2014.

[10] S. Helmut. Probabilistic part-of-speech tagging using decision trees. In *Proc. of the EMNLP*, pages 154–165, 1994.

[11] E. Hüllermeier. Implication-based fuzzy association rules. In *Proc. of the Int. Conf. on PKDD*, pages 241–252, 2001.

[12] E. Hüllermeier. Association rules for expressing gradual dependencies. In *Proc. of the Int. Conf. on PKDD*, pages 200–211, 2002.

[13] J. Kacprzyk and R.R. Yager. Linguistic summaries of data using fuzzy logic. *Journ. of General Systems*, 30 :133–154, 2001.

[14] A. Laurent, M.-J. Lesot, and M. Rifqi. GRAANK : Exploiting rank correlations for extracting gradual itemsets. In *Proc. of the Int. Conf. on FQAS*, pages 382–393, 2009.

[15] A. Oudni, M.-J. Lesot, and M. Rifqi. Characterisation of gradual itemsets through ≪ especially if ≫ clauses based on mathematical morphology tools. In *Proc. of EUSFLAT*, pages 826–833, 2013.

[16] R. Siemens and S. Schreibman. *A companion to digital literary studies*. 2013.

[17] R.R. Yager. A new approach to the summarization of data. *Information Sciences*, 28 :69–86, 2001.

[18] L. A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications*, 9 :149–184, 1983.