Une mesure d'intérêt à base de surreprésentation pour l'extraction des motifs syntaxiques stylistiques

Mohamed-Amine Boukhaled, Francesca Frontini, Jean-Gabriel Ganascia

LIP6 (Laboratoire d'Informatique de Paris 6), Université Pierre et Marie Curie and CNRS (UMR7606),
ACASA Team, 4, place Jussieu,
75252-PARIS Cedex 05 (France)
{mohamed.boukhaled, francesca.frontini, jean-gabriel.ganascia}@lip6.fr

Résumé. Dans cette contribution, nous présentons une étude sur la stylistique computationnelle des textes de la littérature classiques française fondée sur une approche conduite par données, où la découverte des motifs linguistiques intéressants se fait sans aucune connaissance préalable. Nous proposons une mesure objective capable de capturer et d'extraire des motifs syntaxiques stylistiques significatifs à partir d'un œuvre d'un auteur donné. Notre hypothèse de travail est fondée sur le fait que les motifs syntaxiques les plus pertinents devraient refléter de manière significative le choix stylistique de l'auteur, et donc ils doivent présenter une sorte de comportement de surreprésentation contrôlé par les objectifs de l'auteur. Les résultats analysés montrent l'efficacité dans l'extraction de motifs syntaxiques intéressants dans le texte littéraire français classique, et semblent particulièrement prometteurs pour les analyses de ce type particulier de texte.

Abstract.

An Overrepresentation-based Interestingness Measure for Syntactic Stylistic Pattern Extraction

In this contribution, we present a computational stylistic study of the French classic literature texts based on a data-driven approach where discovering interesting linguistic patterns is done without any prior knowledge. We propose an objective measure capable of capturing and extracting meaningful stylistic syntactic patterns from a given author's work. Our hypothesis is based on the fact that the most relevant syntactic patterns should significantly reflect the author's stylistic choice and thus they should exhibit some kind of overrepresentation behavior controlled by the author's purpose. The analysed results show the effectiveness in extracting interesting syntactic patterns from classic French literary text, and seem particularly promising for the analyses of such particular text.

Mots-clés : Stylistique computationnelle, fouille de texte, motifs syntaxiques, mesure d'intérêt

Keywords: Computational stylistic, text mining, syntactic patterns, interestingness measure

1 Introduction

La stylistique computationnelle est un sous-domaine de la linguistique informatique qui se situe à l'intersection de plusieurs domaines de recherche comme le traitement automatique du texte, l'analyse littéraire et la fouille de données statistique. L'objectif de la stylistique computationnelle est d'extraire des motifs de style caractérisant un type particulier de textes à l'aide des méthodes statistiques et automatiques. En prenant le cas de l'étude du style d'écriture d'un auteur particulier, la tâche sera d'explorer automatiquement les formes linguistiques de son style qui ne sont pas seulement caractéristiques mais aussi volontairement surutilisées par cet auteur par rapport à une norme linguistique. Cependant, la notion de style dans le contexte de la stylistique computationnelle se révèle être assez large vu qu'elle se manifeste sur plusieurs niveaux linguistiques : lexical, syntaxique, sémantique et pragmatique. Chaque niveau possède ses propres marqueurs de styles et ses propres unités linguistiques qui le caractérisent.

Grace à la notion du style, la stylistique computationnelle interfère avec de nombreuses autres tâches connexes telles que l'attribution d'auteur (Stamatatos 2009), la classification stylistique de texte (Kessler et al. 1997), génération de texte basée sur style (Hovy 1990), l'évaluation automatique de la lisibilité et de la complexité du texte (Pitler & Nenkova 2008).

Les techniques de la stylique computationnelle ont été utilisées pendant plusieurs années pour étudier les questions relatives au style dans le contexte de l'analyse littéraire, voir (Siemens & Schreibman 2013) pour un aperçu et une discussion. D'un point de vue méthodologique, deux types différents d'approches ont émergé:

• L'approche conduite par classification, qui peut être simplifiée comme suit: une classification connue a priori se trouve dans la littérature (comme les comédies vs tragédies de Shakespeare); certaines caractéristiques linguistiques sont identifiées sur la base de leur pertinence et de leur capacité à reproduire cette classification. Ces caractéristiques linguistiques sont utilisées par la suite pour voir si la distinction a priori se maintient ou non quand on se base sur des techniques de classification automatique comme le clustering par exemple (Craig 2004).
• L'approche herméneutique, dans laquelle les textes littéraires sont analysés afin d'en extraire automatiquement , sans aucune connaissance préalable, les caractéristiques linguistiques intéressantes qui peuvent ensuite être utilisées par les experts du domaine pour produire une analyse critique mieux informée (Mahlberg 2012, Ramsay 2011).

Dans cette contribution, nous présentons une étude stylistique computationnelle des textes classiques de la littérature française basée sur une approche herméneutique conduite par données où la découverte des formes linguistiques intéressantes se fait sans aucune connaissance préalable. Plus précisément, la méthode proposée est fondée sur l'évaluation de la surreprésentation des motifs syntaxiques dans un texte par rapport à un corpus de norme. Cette méthode est destinée à soutenir une analyse textuelle en focalisant sur :
1) La vérification du degré d'importance de chaque motif syntaxique (segments syntagmatiques avec d'éventuel trous).
2) L'extraction automatique d'une liste de motifs syntaxiques caractérisant le style syntaxique d'une œuvre d'un auteur donné.

2 Approche pour l'extraction des motifs syntaxiques pertinents

Notre méthode est composée de deux étapes. D'abord, un algorithme d'extraction de motif séquentiel est appliqué sur les textes pour en extraire des motifs syntaxiques récurrents. Deuxièmement, une mesure d'intérêt basée sur l'évaluation de la surreprésentation (en termes de fréquence d'apparition) par rapport à un corpus de norme est appliquée à l'ensemble des motifs syntaxiques extraits. Ainsi, à chaque motif syntaxique sera affecté un poids en fonction de sa surreprésentation indiquant son importance et sa pertinence dans la caractérisation du style syntaxique du texte en question. Dans ce qui suit, nous présentons le corpus à traiter et le protocole de son découpage en deux éléments : texte à analyser et texte de norme dans la sous-section 2.1. Ensuite, la sous-section 2.2 introduit quelques éléments nécessaires pour la compréhension du processus d'extraction de motifs syntaxiques séquentiels. Enfin, la formulation et les détails statistiques de la mesure d'intérêt proposée sont présentés à la section 2.3.

2.1 Corpus analysé

Dans notre étude, nous avons utilisé quatre romans écrits par quatre célèbres auteurs français: *Eugénie Grandet* de Balzac, *Madame Bovary* de Flaubert, *Notre Dame de Paris* de Hugo et le *Ventre de Paris* de Zola. Ce choix est motivé par notre intérêt particulier pour la littérature française classique du 19ème siècle. Le fait que tous ces textes soit du même genre littéraire et écrits par des auteurs appartenant à la même époque permet de réduire l'effet qu'a la variation du genre et de l'époque sur le style d'écriture. Ce qui permet à son tour d'avoir une étude moins biaisée et bien focalisée sur le style d'écriture propres à ces auteurs. Au moment de l'analyse des motifs syntaxiques chaque texte écrit par un de ces quatre auteurs est mis en contraste avec les textes écrits par les trois autres auteurs. C'est à dire que ces trois texte seront considérés comme corpus de norme à partir duquel on va évaluer l'hypothèse de la surreprésentation des motifs syntaxiques dans le quatrième texte restant, comme expliqué dans la suite de cette section.

2.2 Extraction des motifs syntaxiques

Dans notre étude nous considérons une approche syntagmatique. Le texte est d'abord segmenté en un ensemble de phrases, puis chaque phrase est représentée par une séquence d'étiquettes syntaxiques (POS-tag) [1]correspondantes aux mots de la phrase. Ce qui permet de produire à la fin un ensemble de séquences syntaxique pour chaque texte. Par exemple, la phrase «*Le silence profond régnait nuit et jour dans la maison .* » sera représenté par la séquence :

$$< DET\ , NOM\ , ADJ\ , VER\ , NOM\ , KON\ , NOM\ , PRP\ , DET\ , NOM\ , SENT >$$

Puis, des motifs séquentiels d'une longueur déterminée avec leurs fréquences d'apparition (comptées par nombre de phrases et décrit plus souvent par le terme « support ») sont extraits de cette base de données séquentielle syntaxique en utilisant un algorithme d'extraction de motifs séquentiels (Viger et al. 2014). Un motif syntaxique consiste en un segment syntagmatique séquentiel (avec d'éventuels trous) présent dans la séquence syntaxique. Il peut être considéré

[1] Liste complète des étiquettes syntaxiques sur : http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html

comme une sorte de généralisation de la notion des n-gram très utilisée dans le domaine du traitement automatique de la langue. Voici quelques exemples de motifs syntaxiques présents dans la séquence de l'exemple cité ci-dessus [2]:

- $< DET >< NOM >< ADJ >$
- $< NOM >< ADJ >< VER >< NOM >$
- $< KON >< NOM >< * >< DET >< NOM >$

Pour éviter l'effet de la fluctuation statistique sur l'analyse des motifs avec basses fréquences, nous avons considéré une contrainte de seuil minimum de fréquence de 1%. C'est-à-dire que nous nous concentrions uniquement sur des motifs qui sont présents dans au moins 1% des phrases du texte analysé.
Cependant, comme le processus d'extraction de motifs séquentiels est connu par sa propriété de produire une grande quantité de motifs, et cela même dans des échantillons de textes relativement petits, une mesure d'intérêt doit être appliquée afin d'identifier les motifs les plus importants et pertinents pour la caractérisation du style syntaxique du texte en question. Cette mesure d'intérêt est expliquée dans la sous-section suivante.

2.3 Evaluation de la pertinence des motifs syntaxiques

Notre hypothèse est basée sur le fait que les motifs syntaxiques les plus pertinents devraient refléter de manière significative le choix stylistique de l'auteur et doivent ainsi se caractériser par une considérable surreprésentation dans ses textes. Cependant, pour capturer cette surreprésentation on ne peut pas se référer seulement à la fréquence brute, ou même relative, des motifs syntaxiques. En effet, une utilisation plus fréquente d'un motif syntaxique par un auteur (ce qui se traduit par une fréquence relative très élevée) n'indique pas nécessairement un choix ou un trait stylistique puisque ça peut être très bien une propriété imposée par la grammaire de la langue ou par les caractéristiques syntaxiques du genre du texte.
Ainsi, pour évaluer la surreprésentation des motifs dans un texte, on utilisera une approche empirique basée sur la comparaison de la fréquence d'apparition d'un motif syntaxique dans un texte par rapport à sa fréquence d'apparition dans un corpus de norme. Un ratio α entre ces deux quantités est calculé :

$$\alpha = \frac{Fréquence\ du\ motif\ dans\ le\ corpus\ de\ norme}{Fréquence\ du\ motif\ dans\ le\ texte}$$

Dans notre expérimentation nous avions constaté empiriquement que la distribution du ratio α suit un comportement Gaussien. En effet, les valeurs du ratio α sont réparties autour d'une valeur centrale (voir Fig. 1). Cela est dû au fait que la fréquence d'apparition d'un motif syntaxique dans un texte est fortement corrélée à sa fréquence d'apparition dans un corpus de norme avec quelques cas particuliers qui présentent une certaine aberrance (voir Fig. 2). Ce sont ces cas aberrants qui représentent un intérêt particulier pour notre étude parce qu'ils représentent une certaine déviation linguistique (propres au style de l'auteur) par rapport à ce qu'on s'attend de voir dans un corpus de norme.

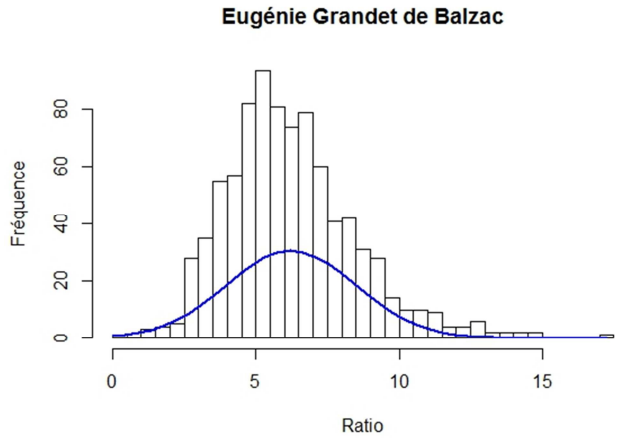

Fig. 1. Illustration du comportement gaussien du ratio α dans le roman *Eugénie Grandet* de Balzac

[2] Le symbole <*> correspond à un trous qui peux être remplacé par n'importe quelle étiquette syntaxique

Cela nous permet d'utiliser la méthode de détection des cas aberrants basée sur la distribution Gaussienne (Chandola et al. 2009). En effet, la surreprésentation d'un motif dans ce cas se traduira par un comportement aberrant négatif plus grand par rapport aux autres motifs. Les motifs les plus surreprésentés dans un texte seront ceux associés aux valeurs de z-score standard Z les moins élevées. Les valeurs de z-score sont calculées comme suit :

$$Z_i = \frac{\alpha_i - \hat{\alpha}}{s}$$

Où Z_i et α_i sont respectivement le ratio α et le z-score Z associé au i-éme motif syntaxique, α_i et s sont respectivement la moyenne et l'écart-type du ratio α

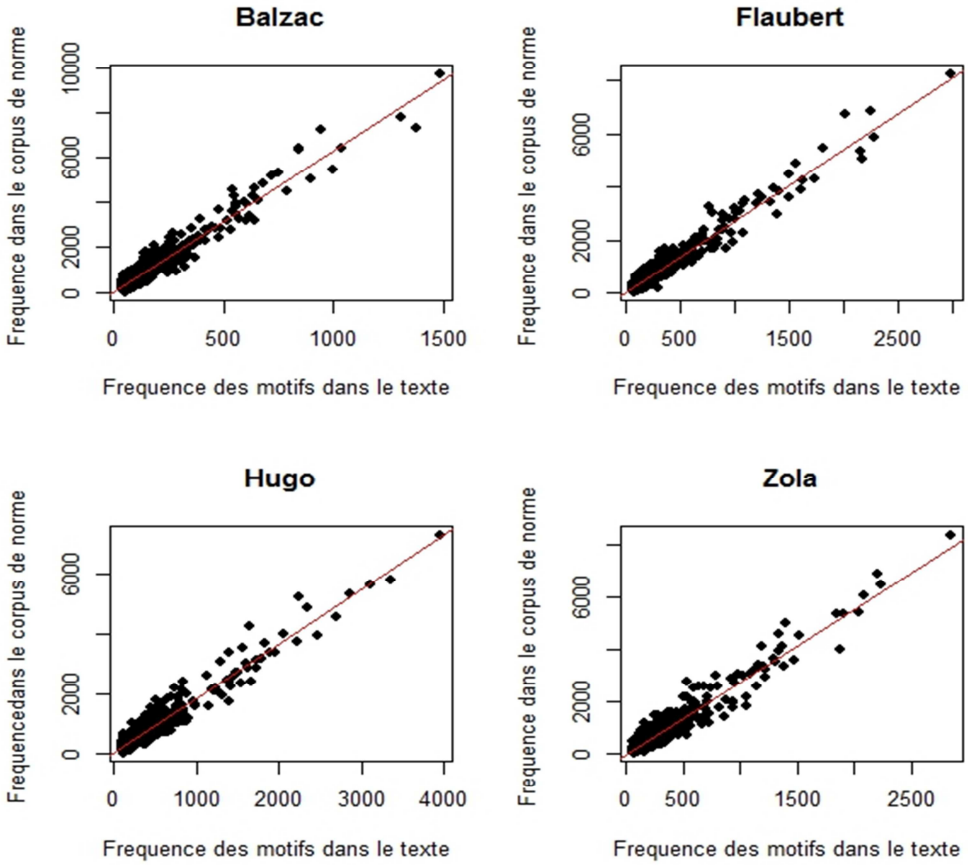

Fig. 2. Fréquence d'apparition d'un motif syntaxique dans un texte par rapport sa fréquence d'apparition dans un corpus de norme pour les romans étudiés, chaque point dans un graphe représente un motif syntaxique. Les lignes tracées représentent les courbes de régression linéaire capturant le comportement attendue du ratio α

3 Résultats et Discussion

Dans cette section, nous présenterons quelques exemples de motifs syntaxiques extraits et classés comme motifs caractéristiques du style des textes inclus dans notre corpus. En utilisant la méthode proposée, les motifs extraits semblent avoir une forte pertinence pour caractériser le style de l'auteur du texte en question, mais aussi pour le scénario du roman et le genre littéraire dans lequel il s'inscrit. Dans le roman *Madame Bovary* de Flaubert, les motifs extraits représentent bien la fonction rythmique plutôt que fonctionnelle que donne Flaubert à la virgule et au point-virgule (Mangiapane 2012). Par exemple dans le cas du motif (1) nous voyons la virgule précédant la conjonction, suivie par une clause imbriquée.

Motif (1) <PUN> < KON>< PUN> <PRP>, avec support= 113, exemples d'instances dans le texte :

- `, et , à`
- `, mais , avant`
- `; et , à`

Dans le *Ventre de Paris* de Zola, et dans le même sens, les motifs classés comme pertinent qualifient clairement l'utilisation des clauses imbriquée pour décrire des situations ou des attitudes comme dans le motif (2) ou bien pour les descriptions des lieux publics et des objets en affiche comme dans le motif (3) :

Motif (2) : <PUN> <PRP> <PRP> <NOM>, support= 104, exemple (en gras) d'instances dans le texte :
« Florent se heurtait à mille obstacles **, à des porteurs** qui se chargeaient **, à des marchandes** qui discutaient de leurs voix rudes ; il glissait sur le lit épais d' épluchures et de trognons qui couvrait la chaussée , il étouffait dans l' odeur puissante des feuilles écrasées .»

Motif (3): <NOM> <PUN> <PRP> <NOM> <ADJ>, support= 68, exemples d'instances dans le texte :
- angles , à fenêtres étroites
- très-jolies , des légendes miraculeuses
- écrevisses , des nappes mouvantes

Dans *Eugénie Grandet* de Balzac, nous pouvions constater d'autres différentes fonctions communicatives accomplies par les motifs syntaxiques et leurs instances textuelles, par exemples :

Le motif (4): <PUN> <VER> <NAM> <PRP>, avec support= 49, est utilisé pour faire une post introduction d'un discoure directe sous forme d'une clause imbriquée. Exemples d'instances dans le texte :
- , dit Grandet en
- , reprit Charles en
- , dit Cruchot en

Le motif (5): <NUM> <NUM> <NOM>, avec support= 54, est un motif utilisé pour parler des sommes d'argents, ce qui est typique pour le scénario du roman où l'argent joue un rôle très important . Exemples d'instances dans le texte :
- vingt mille francs
- deux mille louis
- sept mille livres

Le motif (6) : <ADV> <VER> <PRO> <ADV>, avec support= 59, est utilisé pour exprimer des questions négatives :
- n' avait -il pas
- ne disait -on pas
- ne serait -il pas

Le motif (7) : <PUN> <NOM> <PUN> <VER>, avec support= 44, représente la ponctuation largement utilisée pour imiter l'intonation orale et même de reproduire les phénomènes de performance tels que bégayer :
- , messieurs , cria
- , madame , répondi
- , mademoiselle , disait

Enfin pour le dernier texte, *Notre Dame de Paris* de Hugo en l'occurrence, nous avons remarqué que les motifs syntaxiques extraits comme étant caractéristique sont beaucoup plus pertinents pour décrire le contenu de ce roman et la manière dont l'auteur a utilisé pour introduire le lecteur dans l'histoire et le familiariser avec l'endroit ou se déroulera la plupart des évènements.
Par exemple, dans le motif (8) le nom propre est souvent un endroit, surtout au début du roman où les pièces descriptives sont plus fréquentes dans le but de guider le lecteur dans la topographie de Paris médiéval.

Motif (8) : <NOM> <PRP> <NAM> <PUN>, support= 340, exemples d'instances dans le texte :
- hôtel de Bourbon ,
- murailles de Paris ,
- dauphin de Vienne ;

Par ailleurs, le motif (9) est souvent utilisé pour présenter les personnages en indiquant d'abord leur nom et leur titre. Il convient de noter que le roman *Notre Dame de Paris* présente une pléthore de personnages secondaires.

Motif (9): <NAM> <PRP> <NAM> <PUN>, support = 118, exemples d'instances dans le texte :
- Marguerite de Flandre,
- Jehan de Troyes,

Les quelques exemples analysés indiquent d'une part que la technique présentée est efficace pour extraire des motifs syntaxiques intéressants dans des textes littéraires, et cela semble particulièrement prometteur pour les analyses de ce type de texte. D'autre part, la technique proposée, ainsi que d'autres techniques semblables, nous invite à poser plus de questions sur l'interprétation linguistiques et la significativité de ce qui est réellement capturé par ces motifs syntaxiques. Certaines structures syntaxiques peuvent être importantes car elles sont typiques du style de l'auteur (son empreinte stylistique), mais elles peuvent être aussi très bien dictées par les besoins fonctionnels, en raison de la question particulière de l'œuvre, ou les conventions du genre choisi. Ceci est particulièrement vrai pour l'analyse syntaxique, où les contraintes fonctionnelles imposées, qui limitent la liberté d'auteur, sont plus évidents.

4 Conclusion

Dans cette contribution, nous avons présenté une étude sur la stylistique computationnelle des textes de la littérature classiques française basée sur une approche conduite par données, où la découverte des motifs linguistiques intéressants se fait sans aucune connaissance préalable. Nous avons proposé une mesure objective capable de capturer et d'extraire des motifs syntaxiques stylistiques significatifs à partir d'une œuvre d'un auteur donné. Pour évaluer l'efficacité de la méthode proposée, nous avons mené une expérience sur quatre romans classique français très célèbres. Les résultats analysés montrent l'efficacité dans l'extraction de motifs syntaxiques très intéressants d'un point de vue stylistique à partir de ce type particulier de texte.
Sur la base de la présente étude, nous avons déduit plusieurs perspectives et futures directions de recherches. Premièrement, nous allons explorer l'utilité d'utilisation d'autres mesures statistiques pour évaluer l''intérêt stylistique d'un motif syntaxique donné. Deuxièmement, cette étude sera élargie pour inclure les motifs morphosyntaxiques (forme et lemme des mots). Troisièmement, nous avons l'intention d'expérimenter avec d'autres différentes langues en utilisant d'autres corpus largement employés dans le domaine de la stylistique computationnelle en général.

Remerciement

Ce travail a bénéficié d'une aide d'État gérée par l'Agence Nationale de la Recherche dans le cadre des Investissements d'Avenir portant la référence ANR-11-IDEX-0004-02

Références

Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), p.15.

Craig, H., 2004. Stylistic analysis and authorship studies. A companion to digital humanities, 3, pp.233–334.

Hovy, E.H., 1990. Pragmatics and natural language generation. Artificial Intelligence, 43(2), pp.153–197.

Kessler, B., Numberg, G. & Schütze, H., 1997. Automatic detection of text genre. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. pp. 32–38.

Mahlberg, M., 2012. Corpus stylistics and Dickens's fiction, Routledge.

Mangiapane, S., 2012. Ponctuation et mise en page dans Madame Bovary: les interventions de Flaubert sur le manuscrit du copiste. Flaubert. Revue critique et génétique, (8).

Pitler, E. & Nenkova, A., 2008. Revisiting readability: A unified framework for predicting text quality. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 186–195.

Ramsay, S., 2011. Reading machines: Toward an algorithmic criticism, University of Illinois Press.

Siemens, R. & Schreibman, S., 2013. A companion to digital literary studies, John Wiley & Sons.

Stamatatos, E., 2009. A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), pp.538–556.

Viger, P.F. et al., 2014. SPMF: A Java Open-Source Pattern Mining Library. Journal of Machine Learning Research, 15, pp.3389–3393.