

A Peculiarity-based Exploration of Syntactical Patterns: a Computational Study of Stylistics

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

▶ **To cite this version:**

Mohamed-Amine Boukhaled, Francesca Frontini, Jean-Gabriel Ganascia. A Peculiarity-based Exploration of Syntactical Patterns: a Computational Study of Stylistics. Workshop on Interactions between Data Mining and Natural Language Processing DMNLP'15 ECML/PKDD 2015 Workshop, Sep 2015, Porto, Portugal. pp.31-40. hal-01198413

HAL Id: hal-01198413

https://hal.sorbonne-universite.fr/hal-01198413

Submitted on 12 Sep 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Peculiarity-based Exploration of Syntactical Patterns: a Computational Study of Stylistics

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

Abstract. In this contribution, we present a computational stylistic study and comparison of classic French literary 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 peculiar overrepresentation behavior controlled by the author's purpose with respect to a linguistic norm. The analyzed results show the effectiveness in extracting interesting syntactic patterns from novels, and seem particularly promising for the analysis of such particular texts.

Keywords: Computational Stylistics, Interestingness Measure, Sequential Pattern Mining, Syntactic Style

1 Introduction

Computational stylistics is a subdomain of computational linguistics located at the intersection of several research areas such as natural language processing, literary analysis and data mining. The goal of computational stylistics is to extract style patterns characterizing a particular type of texts using computational and automatic methods (Craig 2004). When investigating the writing style of a particular author, the task will automatically explore linguistic forms of his style, which is not only distinguishing features, but also the deliberate overuse of certain structures by the author compared to a linguistic norm (Mahlberg 2012). However, the notion of style in the context of computational stylistics appears to be wide enough, and is manifested on several linguistic levels: lexicon, syntax, semantics and pragmatics. Each level has its own markers of styles and its own linguistic units that characterize it.

Many works have been done in the literature to analyze the stylistic traits on these different linguistic levels (Biber 2006, Biber & Conrad 2009, Ramsay 2011, Frontini et al. 2014; see Siemens & Schreibman, 2013 for a discussion and overview). In this contribution, syntactic style will be targeted.
In their study Quiniou et al. (2012) have shown the interest of using sequential data mining methods for the stylistic analysis of large texts. They have shown that relevant and understandable patterns that are characteristic of a specific type of text can be extracted using sequential data mining techniques such as sequential pattern mining.
However, the process of extracting textual patterns is known by its property of producing a large amount of patterns, even from a relatively small sample of text. Thus, a measure of interest is to be applied to identify the most important and relevant patterns for the characterization of the text's style in question.
In this paper, we present a computational stylistic study of classic texts of French literature based on a data-driven approach where the discovery of interesting linguistic forms is done without any prior knowledge. Specifically, the proposed method is based on the assessment of the peculiar overrepresentation of syntactic patterns extracted using sequential data mining technique from texts with respect to a norm corpus. This method is intended to quantitatively support a textual analysis by focusing on the verification of the degree of importance of each syntactic pattern (syntagmatic segments with potential gaps), and by extracting the syntactic patterns that characterize the syntactical style of a work by a particular author.

2 Approach for extracting relevant syntactic patterns

Our method consists of two steps. First, a sequential pattern mining algorithm is applied to the texts in order to extract recurrent syntactic patterns. Second, a peculiarity-based interestingness measure that evaluates of the overrepresentation (in terms of frequency of occurrence with respect to a norm corpus) is applied to the set of extracted syntactic patterns. Thus, each syntactic pattern will be assigned an interestingness value indicating its importance and its relevance for the characterization of text's syntactic style. In what follows, we present in section 2.1 the corpus used in our experience, and its dividing protocol into two parts: text to analyze and text used as norm. Then, section 2.2 introduces some elements necessary to understand the process of extracting sequential syntactic patterns. Finally, the formulation and the statistical details of the proposed interestingness measure are presented in Section 2.3.

2.1 Analyzed Corpus

In our study, we used four novels, belonging to the same genre and the same literary time span, written by four famous classic French authors: Balzac's *"Eugenie Grandet"*, Flaubert's "*Madame Bovary*", Hugo's "*Notre Dame de* Paris" and Zola's "*Le ventre de Paris"*. This choice is motivated by our particular interest in studying the style of the classical French literature of the 19th century. At the time of the analysis of the syntactic patterns, each text written by one of the four authors is contrasted with texts written by the three other authors. That is to say that these three texts will be considered as norm corpus from which we will evaluate the hypothesis of the overrepresentation of syntactic patterns in the fourth remaining text, as explained later in this section.

2.2 Extraction of syntactic patterns

In our study we consider a syntagmatic approach. The text is first segmented into a set of sentences, each sentence is then represented by a sequence of syntactic labels (POS-tag)[1] corresponding to the words of the sentence using Treetagger (Schmid 1994). This produces at the end a set of syntactic sequences for each text. For exemple, the sentence "*Le silence profond régnait nuit et jour dans la maison.*" Will be represented by the sequence:

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

Then, sequential patterns of a certain length with their supports (a number indicating how many sentences contain the pattern) are extracted from this syntactic sequential database using a sequential pattern extraction algorithm (Viger et al. 2014). Syntactic pattern consists of a sequential syntagmatic segment (with possible gaps) present in the syntactic sequences. It can be considered as a kind of generalization of the notion of n-gram widely used in the field of automatic language processing. Examples of syntactic patterns present in the sequence of the example above:

- $< DET >< NOM >< ADJ >$
- $< NOM >< ADJ >< VER >< NOM >$
- $< KON >< NOM > <*^{2}> < DET >< NOM >$

To avoid the effect of statistical fluctuations on the analysis of patterns with low supports, we considered a support's threshold of 1%. That is to say that we focus only on patterns that are present in at least 1% of the sentences of the analyzed text. However, as sequential pattern mining is known to produce a large quantity of patterns even from relatively small samples of texts,

[1] Frech treetagger tagset:
http://www.cis.unimuenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html

[2] <*> denotes a gap that can be filled with any POS tag

an interestingness measure should be applied on these patterns in order to identify the most important ones. This interestingness measure is explained in the next section.

2.3 Evaluation of the relevance of syntactic patterns

Our hypothesis to evaluate the relevance of a syntactic pattern is based on the fact that the most relevant ones should significantly reflect the stylistic choice of the author and should thus be characterized by a significant peculiar quantitative behavior, this peculiar behavior translate into a support's over-representation in his texts.
However, to capture this overrepresentation one cannot refer only to the absolute frequency of occurrence (support) Indeed, more frequent use of a syntactic pattern by an author (which translates into a relatively high support) does not necessarily indicate a stylistic choice since it can be very well a property imposed by the grammar of the language or by syntactic features that are characteristic of text's genre.
Thus, to assess the over-representation of a pattern, we use an empirical approach based on the comparison of the support of a syntactic pattern in a text to that found in a norm corpus. A ratio α between these two quantities is calculated as follow:

$$\alpha = \frac{\text{frequency of a pattern in the norm corpus}}{\text{frequency pattern in the text}}$$

In our experiments we found empirically that the distribution of the ratio α exhibits a Gaussian behavior. Indeed, the values of the α ratio are normally distributed around a central value (see Fig. 1). This is due to the fact that the frequency of occurrence of a syntactic pattern in a text is highly correlated with the frequency of occurrence in the norm corpus with a few exceptional special cases or outliers (see Fig. 2). These outliers represent the patterns of special interest for our study because they represent a certain linguistic deviation that is specific to the author's style compared to what one would expect to see in the norm corpus.

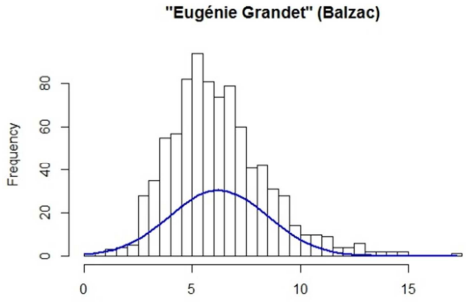


Fig. 1. Gaussian behaviour of the ratio α in Balzac's "*Eugénie Grandet*" novel

The configuration described above allows us to use an outlier detection method based on Gaussian distribution and Z-score to identify such special patterns (Chandola et al. 2009). The over-representation of a pattern in this case will result in a greater negative aberrant behavior compared to other patterns. The most over-represented patterns will be those associated with lowest values of standard z-score Z. The z-score values are calculated as follows:

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

Where α_i and Z_i are respectively the ratio α and the z-score corresponding to the i-th syntactic pattern. $\hat{\alpha}$ and S are respectively the mean and standard deviation of the ratio α .

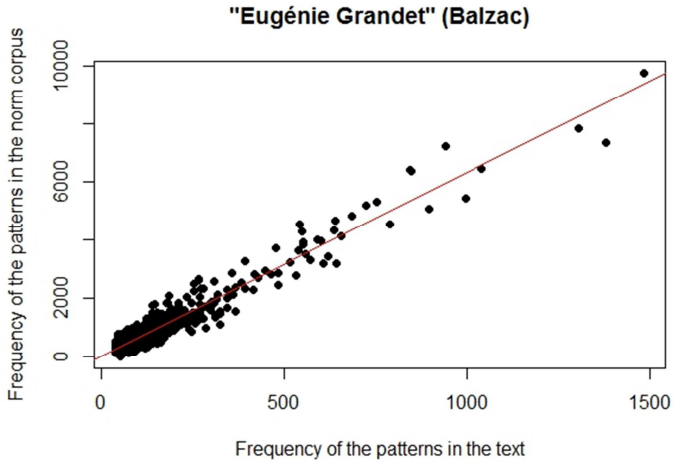


Fig. 2. Frequencies of syntactic patterns in a text with respect to their frequencies in the norm corpus for the studied novel. Each point in the graph represents a syntactic pattern. The plotted lines represent the linear regression lines capturing the expected behaviour of the α ratio

3 Results and Discussion

In this section, we present some examples of relevant syntactic patterns extracted from our corpus. Using the proposed method, the extracted patterns seem to have a strong relevance to characterize the style of the authors of our corpus but also to the novels' content and the literary genre in which it operates. In the Flaubert's *Madame Bovary*, several extracted patterns well represent the rhythmic rather than functional role of punctuation that is peculiar to the style of Flaubert (Mangiapane 2012). For example pattern (1) captures instances of a comma preceding the conjunction, followed by a parenthetical clause.

Pattern (1) <PUN> < KON>< PUN> <PRP>, with support= 113, sample instances of the pattern in the text:

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

In *le Ventre de Paris* of Zola, and in the same direction, the syntactic patterns extracted as relevant clearly represent the use of nested clauses to describe situations or attitudes in the novel such as in the pattern (2), or to describe public places and objects in displays in long lists as in the pattern (3):

Pattern (2) : <PUN> <PRP> <PRP> <NOM>, support= 104, sample instances of the pattern in the text (bold text):
« 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 .»

Pattern (3): <NOM> <PUN> <PRP> <NOM> <ADJ>, support= 68, sample instances of the pattern in the text (bold text):

- angles , à fenêtres étroites
- très-jolies , des légendes miraculeuses
- écrevisses , des nappes mouvantes

In *Eugénie Grandet* of Balzac, other different communicative functions are performed by the syntactic patterns and their textual instances, for example:

Pattern (4): <PUN> <VER> <NAM> <PRP>, support= 49, which is used as post-introducer of direct speech. This rather formulaic way of specifying (in a parenthetical form) the utterer of a reported speech is common to all, but seems to be strongly preferred by Balzac, while the other authors have

shown a more varied style in introducing dialogues. Sample instances of the pattern in the novel:

- , dit Grandet en
- , reprit Charles en
- , dit Cruchot en

Pattern (5): <NUM> <NUM> <NOM>, support= 54, is a pattern used to refer to money, which is typical for the novel scenario where money plays a very important role. Sample instances of the pattern in the novel:

- vingt mille francs
- deux mille louis
- sept mille livres

Pattern (6) : <ADV> <VER> <PRO> <ADV>, support= 59, is used to express negative questions :

- n' avait -il pas
- ne disait -on pas
- ne serait -il pas

Pattern (7) : <PUN> <NOM> <PUN> <VER>, support= 44, represent the punctuation extensively used to mimic spoken intonation and even to reproduce performance phenomena such as stutter. :

- , messieurs , cria
- , madame , répondit
- , mademoiselle , disait

The few analyzed examples indicate that the presented technique is effective in extracting interesting syntactic patterns from a single text, and this seems particularly promising for the analyses of such classic literary texts.
On the other hand, this technique, as well as other similar ones, prompts the question of what is really captured by significant patterns. Some structures may be significant because they are typical of an author's style, its fingerprint - as we may say borrowing a metaphor often used in attribution studies, or they may be dictated by functional needs, due to the particular topic of the novel, or to the conventions of the chosen genre. This is particularly true for syntactic analysis, where the functional constraints on the authorial freedom are more evident. Much further works have to be carried out concerning this issue.

4 Conclusion

In this paper, we have presented an objective interestingness measure to extract 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 peculiar overrepresentation behavior controlled by the author's purpose. To evaluate the effectiveness of the proposed method, we conducted an experiment on a classic French Corpus. The analyzed results show the effectiveness in extracting interesting syntactic patterns from this type of text.
Based on the current study, we have identified several future research directions such as exploring other statistical measures to assess the interestingness of a given syntactic pattern, and expanding the analysis to include morpho-syntactic patterns (form and lemma words). Finally, we intend to experiment with other languages and text sizes using standard corpora employed in the field of computational stylistics at large.

Acknowledgment

This work was supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

References

Biber, D., 2006. *University language: A corpus-based study of spoken and written registers*, John Benjamins Publishing.

Biber, D. & Conrad, S., 2009. *Register, genre, and style*, Cambridge University Press.

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.

Frontini, F., Boukhaled, M.A. & Ganascia, J., Linguistic Pattern Extraction and Analysis for Classic French Plays.

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).

Quiniou, S. et al., 2012. What about sequential data mining techniques to identify linguistic patterns for stylistics? In *Computational Linguistics and Intelligent Text Processing*. Springer, pp. 166–177.

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

Schmid, H., 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*. pp. 44–49.

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

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