



HAL
open science

The balance between quantitative and qualitative literary stylistics: How the method of "motifs" can help

Dominique Legallois, Thierry Charnois, Meri Larjavaara

► To cite this version:

Dominique Legallois, Thierry Charnois, Meri Larjavaara. The balance between quantitative and qualitative literary stylistics: How the method of "motifs" can help. Legallois, Charnois, Larjavaara. The Grammar of Genres and styles, De Gruyter Mouton, 2018, 9783110589689. 10.1515/9783110595864-008 . hal-03546266

HAL Id: hal-03546266

<https://hal.science/hal-03546266>

Submitted on 27 Jan 2022

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.

The balance between quantitative and qualitative literary stylistics: How the method of “motifs” can help

Dominique Legallois, University of Paris 3 Sorbonne Nouvelle

Thierry Charnois, University of Paris 13

Meri Larjavaara, Åbo Akademi University

At a time when computational methods of extraction of the morphosyntactic specificities of a given author or a literary genre are becoming more and more accessible thanks to new computational tools, linguists and specialists in stylistics and literary discourse can no longer remain unaware of the contribution of quantitative data. Admittedly, a quantitative approach is only one aspect of the stylistic analysis of texts, but this criterion has to be taken into account since statistically overrepresented features of language in texts can be considered as linguistic singularities, peculiar to an author, to a text or to a genre. Some analyses aiming to identify the linguistic specificity of a corpus give extremely interesting results, as evidenced by the small but growing number of monographs or papers demonstrating the approach: for example, in the French textometric tradition, Brunet 1985 on E. Zola, Magri (2009) on the comparison between travel writing and fiction, Kastberg Sjöblom (2006) on J.M.G. Le Clézio, or in other traditions, Burrows (1987) on J. Austen, Hoover (2007) on H. James, Ho (2011) on J. Fowles, Oakes (2014) for various applications, etc.

However, computational stylistics remains marginal for “academic” literary specialists in stylistics, despite the current development of digital humanities. The reasons for this marginal position may stem from the complexity of the methods used, the technical nature of analyses, or a difference of “culture” between text practitioners. There may however be other, deeper reasons than these: namely, the fact that statistical and automated analyses are invariably paradigmatic. They concern *discrete* descriptors, that is to say, units (lexical forms, lemmata, morpho-syntactic categories, punctuation marks) with no direct syntagmatic relations between one another, even if correlations can always be calculated. The contextual interpretation of these descriptors remains vague, therefore, since the main focus is commenting on tables of quantified units, rather than utterances. Highlighting the specificity of a descriptor is of course essential, but to grasp the role of the descriptor in its textual environment is a step that is still too rarely taken. This is the limitation of the discrete approach and may explain why its adoption by specialists in stylistics has been rather lukewarm so far.

The main aim of this paper is to show how the study of an important area within stylistics, namely the characterization of an author's style, can benefit from a new method in corpus linguistics, the discovery of sequential patterns or "motifs", i.e. contiguous strings of word forms/lemmas/POS tags. Motif analysis can be viewed as complementary to discrete approaches and constitutes a more powerful paradigm than other non-discrete analyses such as lexical bundles or clusters because motifs combine annotations at different levels of abstraction.

As will be seen from the analyses given below, this type of complex unit, while not claiming to be exhaustive since a style cannot be reduced to a set of lexico-grammatical patterns, nonetheless provides a more accurate vision of the stylistic characteristics of an author.

The study is based on a corpus of 60 novels by 12 19th-century French writers (see below). The aim is to identify the syntactic motifs favored by each author.

In the first part of the chapter, we examine the issue of the definition of style proposed in a recent paper by Hermann and colleagues (Hermann and al. 2015), in which the quantitative approach features prominently. Through a critical discussion of this study, we show that there are in fact two major trends in stylistics, and even two types of stylistics: a stylistics of identification which has its roots in the stylometry of the nineteenth century, and a stylistics of characterization. The first one (essentially quantitative) is a stylistics based on discriminatory features; the second (essentially qualitative) is interpretative. In order to ensure that quantitative and qualitative approaches are complementary, it is necessary to take not only discrete units, but also non-discrete units such as motifs into account in the analysis.

The second part deals with the presentation of the method of motifs. We formulate a definition of the term and we introduce three possible statistical methods of analysis: an endogenous method, an exogenous method, and a combination of the two.

In the third part, we apply the method of motifs to a literary corpus, and we extract some of the syntactic patterns favored by each author. For reasons of space, the motifs of only five authors (Balzac, Hugo, Gaboriau, Stendhal, and Flaubert) will be discussed.

The fourth part focuses on the qualitative interpretation of motifs. Taking the example of Hugo, we show that some groups of motifs contribute to realizing the same aesthetic project.

1. Stylistics, stylometry and style

1.1 Two stylistics: Identifying and characterizing

Stylometry and corpus stylistics are still unpopular among literary stylisticians. The quote below, although formulated 25 years ago, still resonates in the “non-computational” stylistic tradition:

When stylistic features of a text have been transformed into numerical form, they acquire a status that actually prevents them from being perceived as language-for-communication as such. That is to say, in the very act of transforming textual qualities into counts, their essential process-like character is irretrievably lost. [. . .] Thus no level of (mathematical) sophistication is able to overcome the problem that the processes of meaning constitution have been eliminated before the analysis is undertaken. (Van Peer, 1989: 302)

The problem of the reception of stylometry can be explained therefore by the loss of the qualitative dimension in data interpretation. To solve this problem, it is necessary to take both quantitative and qualitative orientations into account when considering the notion of style.

As it is impossible in this paper to give an overview of the history of stylistics, even less of style, we will draw on a very recent and interesting study by Herrmann et al. (2015) that relates the development of the notion of style in three different traditions, those of German, Dutch and French language and literary studies, since 1945. The paper identifies 6 basic conceptions of style across these traditions. Style can be seen:

1. as constituting a higher-order artistic value (assessed through aesthetic experience);
2. as a holistic gestalt of single texts: style cannot be reduced to descriptive categories or classes;
3. as an expression of the individuality, subjectivity and/or emotional attitude of an author or speaker;
4. as an artifact that presupposes (hypothetical or factual) selection/choice among a set of (more or less synonymous) alternatives;
5. as a deviation from some type of norm, involving (quantitative or cognitive) contrast;
6. as any property of a text that can be measured computationally. (Herrmann et al. 2015: 30)

After considering the various criticisms levelled at these conceptions, and how they have been addressed in the three different traditions, the authors propose a definition of style that is an attempt to provide a common ground for both mainstream and computational literary stylistics:

Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively. (Herrmann et al. 2015: 44)

While this definition may be relevant for a conception of style that encompasses any textual genre¹, we claim that it does not seem sufficiently precise for the analysis of literary style, for three reasons.

Our first criticism is that the vast majority of computational approaches are based on comparative methods (comparisons of genres, or authors, or texts). This comparative approach is usual for a stylistics whose aim is identification: identification of the author, the date of the text, or a particular type or school (Homes 1998). As is well known, the primary aim of stylometry is identification. The term stylometry was coined in 1897 by the Polish philosopher Lutoslawski (1863-1954)², who worked on the dating of Plato's dialogues (Lutoslawski 1897a, 1897b). The concept of stylometry had of course some forerunners before the term *stylometry* itself became established, but Lutoslawski was the first to define stylometry as the science of measuring stylistic affinities:

of two works of the same author and of the same size, that is nearer in time to a third, which shares with it the greater number of peculiarities, provided that their different importance is taken into account, and that the number of observed peculiarities is sufficient to determine the stylistic character of all three works (Lutosławski 1897a: 152)

Lutoslawski considered that the stylistic characteristics of an author constitute a kind of fingerprint that escapes the author's consciousness. As such, the features are objective and not subject to interpretation. Although it is similar to a fingerprint, the style of an author can obviously evolve.

One must however keep in mind a statistical principle due to the specificity of language: comparative methods *necessarily* produce results, that is to say, they systematically

¹ The authors defend a conception of style which is assumed to apply to any text (whether literary or not).
² Pawlowski and Pacewicz (2004).

highlight differences (Kilgarriff 2005; Loiseau 2010). Statistical methods guarantee the detection of identifying features, but not necessarily that of characterizing features.

The difference between these two types of features is fundamental. For example, in our corpus of French 19th century writers (see Section 2.1 below for the composition of the corpus), the forms *ça* (this/that) and *on* (we/one), and verbs in the imperfect (*imparfait*) are over-used in Zola compared to the eleven other writers. How is one to interpret this result? While these data may help to identify a novel by Zola, do they say anything about the style of this author? Simply knowing that *ça* and *on* are comparatively overused in the novel, is not enough then. Without a precise examination of these items in context or a concordance analysis, it is difficult to give these units a status: over-used features (keywords) do not necessarily have an interpretative value.

The distinction between identifying and characterizing features was theorized by the French linguist Rastier (2001) who (in reference to the work of Carlo Ginzburg) differentiated morellian features from spitzerian features. The former owe their name to the Italian doctor Morelli, who revolutionized the attribution of paintings in the late nineteenth century by identifying clues, including anatomic features such as ear lobes, which are unattended to by counterfeiters and experts. Morellian features are not necessarily interpretable, they simply make it possible to identify a text, an author, or a genre. Spitzerian features, from the name of the stylistician L. Spitzer, contribute to characterizing a text, an author, or a genre: they should be considered as creating a system and expressing an aesthetic project. That is the reason why – and this is our second criticism – the definition by Herrmann and al. lacks the notion of the convergence and interaction of features. For formal features to be considered as interpretive features, it is necessary to show that they create more or less closed systems of interrelated elements. Another criticism is the disjunction “quantitatively *or* qualitatively”. We argue to the contrary that a stylistic analysis concerning a text, a type, or an author, may be exclusively quantitative if its aim is identification; if the aim is characterization, however, quantitative and qualitative approaches are necessary and complementary: a qualitative approach cannot claim to analyze one or more features if they are not characteristic in terms of frequency (over-used or under-used in the corpus).

Traditional qualitative analysis cannot allow any shortcuts and must use at least a minimal amount of quantitative information. Moreover, a mere quantitative approach would remain purely descriptive, and does not constitute a literary stylistic analysis. To highlight correlations between features and dimensions (by factorial analysis) is not enough to discover aesthetic effects and communicative intentions.

We therefore replace the definition of style given above by the following one, which applies to literary stylistics that aims at characterization:

Style is a property of texts constituted by an ensemble or several ensembles of interrelated formal features which can be analyzed quantitatively AND qualitatively.

There again, this definition may not be sufficient; it provides nevertheless a consistent frame of analysis for our study of the 12 writers.

1.2 Discrete and non-discrete descriptors

We call “descriptors” all the units generally considered in stylometric analysis, regardless of their nature, i.e. word forms, lemmata, POS, but also average sentence or word length, average word length, type/token ratio (vocabulary richness), and vocabulary growth (homogeneity of text). These descriptors are discrete since they are atomic units. They are completely relevant as morellian features; they can also contribute to characterizing style, but they are difficult to interpret because these units are decontextualised abstractions, even if correlations between them are possible. More syntagmatic units or non-discrete units (continuous or discontinuous) are still little used in analyses of style. Salem (1987) recommended considering repeated segments, that is to say repeatedly occurring sequences of words. These segments are also referred to as ‘n-grams’ or lexical bundles or clusters. Other non-discrete units such as P-Frames (Fletcher, 2003) are more flexible than lexical bundles, since they provide systematic groupings of lexical bundles, which vary in only one position. Repeated segments and P-Frames have proved to be relevant and sufficient for the study of phraseology. Römer (2010) takes these units into account to establish the phraseology of a text. This kind of research is centered on particular genres or academic registers, but there are few studies that take repeated segments into account in a literary stylistic perspective (see however Mahlberg (2013) on Dickens). Table 1 shows the key lexical bundles in *Madame Bovary* by Flaubert with respect to the eleven other writers. Three association measures are used: Calculation of Specificities (Lafon 1984), log-likelihood⁴, and T.score.

key lexical bundles	sub freq	tot freq	specificities	loglikelihood	t.score
de_temps_à_autre,	31	157	84.53	165.87	5.41
les_uns_après_les	9	38	27.50	51.69	2.93

⁴ See Bertels and Speelman (2013) for a comparison between log-likelihood and Specificities.

ce n'est rien !	6	9	26.51	50.54	2.43
elle_se_mettait_à	5	5	25.79	51.64	2.22
se_mit_à_lui	8	34	24.53	45.82	2.76
au_clair_de_lune,	5	7	22.74	43.28	2.22
;_et,_à_travers	4	4	20.63	41.31	1.99
à_la_croix_rouge,	4	4	20.63	41.31	1.99
dans_la_côte_du	4	4	20.63	41.31	1.99
haut_de_la_côte	4	4	20.63	41.31	1.99
se_passa_la_main	4	4	20.63	41.31	1.99
temps_à_autre,_comme	4	4	20.63	41.31	1.99

Table 1: The first 12 key lexical bundles in *Madame Bovary* by Flaubert

However, the question of the granularity of the linguistic forms inevitably arises: lexical bundles can be considered too specific since they are multiword sequences. They do not permit any generalization. For example:

- (1) il eut un geste de (lit. He had a gesture of)⁵
- (2) elle fit un mouvement de (lit. She made a movement of)

are two different lexical bundles, but it would probably be illuminating to group them together as instances of a more abstract unit, in order to investigate the syntactic characteristics of a text. A part of speech n-gram could be useful in an identification task. For example:

PRONOUN + VERB + DET + NOUN + PREP

could be a 5-gram POS. The problem is that such n-grams are too generic, and may match sentences that are very different from (1) and (2)⁶, such as:

- (3) il mangea une pomme en (deux minutes)
lit. He ate an apple in (two minutes)

In view of these respective drawbacks, it is necessary to adopt a mixed and partly non-supervised method that we call the *motif method*. “Motif” is a term that is rather problematic since it is used by several scholars with different but closely-related meanings (Ganascia 2001, Quiniou and al. 2012, Köhler 2015, Longrée and Mellet 2013, Longrée and Mellet, this volume).

2. Motifs: definition and method

2.1 The corpus

The literary corpus consists of 60 novels by 12 19th century French authors. Five novels per writer were selected:

Balzac (*La Cousine Bette*, *Les Illusions perdues*, *Le Lys dans la vallée*, *Le Colonel Chabert*, *Le Père Goriot*), Dumas (*Les trois mousquetaires*, *Vingt ans après*, *Les Quarante-cinq*, *La Reine Margot*, *Le Prince des Voleurs*), Flaubert (*Madame Bovary*, *Les trois contes*, *L'éducation sentimentale*, *Bouvard et Pécuchet*, *Salammbô*), Gaboriau (*Le Crime d'Orcival*, *Le Dossier 113*, *Monsieur Lecoq*, *La Clique dorée*, *L’Affaire Lerouge*), Hugo (*Notre Dame de*

⁵ “Lit.” means that the English translation is a literal one

⁶ This method can however give satisfactory results; cf. Frontini et al., this volume.

Paris, Les Misérables, Les Travailleurs de la mer, L'Homme qui rit, Quatre-vingt-treize), Huysmans (*Les Sœurs Vatar, À rebours, En route, Là-bas, La Cathédrale*), Maupassant (*Bel Ami, Fort comme la mort, Pierre et Jean, Mont Oriol, Une vie*), Sand (*La petite Fadette, La Mare au diable, Indiana, Consuelo, François le Champi*), Stendhal (*Le Rouge et le Noir, Armance, La Chartreuse de Parme, Lucien Leuwen, Le Vert et le Rose*), Sue (*Le Juif Errant, Atar-Gull, Le Morne au Diable, Les Mystères de Paris, Mathilde - Mémoires d'une jeune femme*), Verne (*De la Terre à la Lune, 20000 lieues sous les mers, Deux ans de vacances, L'Île Mystérieuse, Michel Strogoff*), Zola (*Nana, L'Œuvre, l'Assommoir, la Terre, Germinal*).

The motif method is relatively simple. It consists of several steps: specific annotation, the extraction of segments, statistical calculation of segments, and finally identification of motifs from segments.

1st step

Identifying patterns that may be composed of different types of elements: word forms, lemmata and POS. Specifically, the results of a tagger (the French language tagger *Cordial*⁷) are modified with regular expressions as follows:

- we keep invariable forms such as prepositions, conjunctions, frequent adverbs, etc.;
- we keep the lemmata of several frequent verbs (aspectual and modal verbs, auxiliary verbs, etc.) and reduce personal pronouns to their canonical forms;
- we keep the POS labels (common noun, proper noun, adverb, verb, adjective, etc.) of the other words. Verb tense categories are kept. Some semantic categories for nouns and adverbs are also retained. For example, adverbs of manner are labelled ADVMAN, nouns expressing parts of the body are labelled NCCOR, abstract nouns NCABS⁸.

For example, this extract (from *Madame Bovary*)

Nous étions à l'étude, quand le proviseur entra, suivi d'un nouveau habillé en bourgeois et d'un garçon de classe qui portait un grand pupitre. Ceux qui dormaient se réveillèrent, et chacun se leva comme surpris dans son travail.

(we were at prep when the Headmaster came in, followed by a 'new boy' not wearing school uniform, and by a school servant carrying a large desk. Those who had been asleep woke up, and we all rose to our feet as though we had been interrupted at our work.)⁹

is annotated as follows:

nous être à le NC, quand le NC VPS, PASS de un NC PASS en NC et de un NC de NC qui VIMP un ADJ NC. celui qui VIMP se VPS, et chacun se VPS comme ADJ dans DETPOSS NC.

2nd step

Extracting segments of variable length for each novel. For example, from *Madame Bovary*

Segments	Rank	Frequency	Segments
le NC de le NC	1	918	le NC de le NC
NC de le NC ,	2	428	NC de le NC ,
le NC de le NC ,	3	353	le NC de le NC ,

⁷ <http://www.cordial.fr/>

⁸ The lexicon of abstract nouns is based on syntactic tests; for example: *faire preuve de N* (to show N), *manifeste du N* (lit. demonstrate N), *ressentir du N*, *éprouver du N* (to feel N).

⁹ Translation by M. Mauldon (Oxford world's classics).

à le NC de le	4	244	à le NC de le
	...		
avec de le NC ADJ	1127	9	avec de le NC ADJ
avec le NC de DETPOSS	1128	9	avec le NC de DETPOSS
ce être comme un NC	1129	9	ce être comme un NC
	...		
plus ADJ de le NC	2525	6	plus ADJ de le NC
plus ADJ et plus ADJ	2526	6	plus ADJ et plus ADJ
	...		
si ADJ , si ADJ ,	2540	6	si ADJ , si ADJ ,
	...		
VPS un NC , un NC ,	10050	3	VPS un NC , un NC ,

- Table 2: Segments from *Madame Bovary*

Only segments which appear at least twice in all the novels by each author are kept.

3rd step

Calculating the segments.

This step concerns statistical calculation. There are three possible methods:

1- For every author, segments are extracted on the basis of Mutual Information (MI); calculation is therefore said to be endogenous because it does not presuppose a comparison between texts. For example, for Balzac:

Segments	Frequency	Mutual Information
je ne savoir quoi de	18	23,175499
. INT ! oui ,	23	20,530413
je ne avoir jamais PASS	17	19,915905
je ne vouloir pas être	11	19,040236
! dire il en PRES	26	19,010582
. je ne avoir jamais	12	18,271523
NC pour le ADJORD fois	10	18,194517
NC . INT ! oui	10	18,152967
. je ne vouloir pas	18	17,848886
. dire il en PRES	85	17,552335
il se mettre à INF	13	17,337274
environ NUM NUM NC de	5	17,289822

- Table 3: The top 12 segments in Balzac extracted with MI.

Note that Mutual Information tends to highlight strong associations between elements without taking into account how frequent, or infrequent, the associations between the elements are.

2- For every author, segments are extracted with their raw frequency. Then, the frequency of the segments of every author is compared with that of the others; the calculation is exogenous since it is based on a comparison. Calculation of Specificities is therefore used.

Segments	Specificities	sub freq	tot freq
le NC de le NC	Inf ¹⁰	3896	55482
NC de le NC ,	Inf	1349	21739
le NC de DETPOSS NC	Inf	1145	11015
le NC de le NC ,	Inf	982	16062

¹⁰ "Inf" means that the association is very strong.

à le NC de le	Inf	891	13822
NC de le NC .	Inf	798	13030
à le NC de le NC	Inf	737	11664
le NC , le NC	Inf	697	10418
le NC de un NC	Inf	686	7730
NC de le NC de	Inf	662	8995
le NC de le NC .	Inf	604	9796
le NC de NP ,	Inf	566	6404

- Table 4: The top 12 segments in Balzac calculated with Specificities

3 – The previous two methods can be merged: segments calculated by MI are extracted for each author, and then the authors are compared with one another in order to define the key segments.

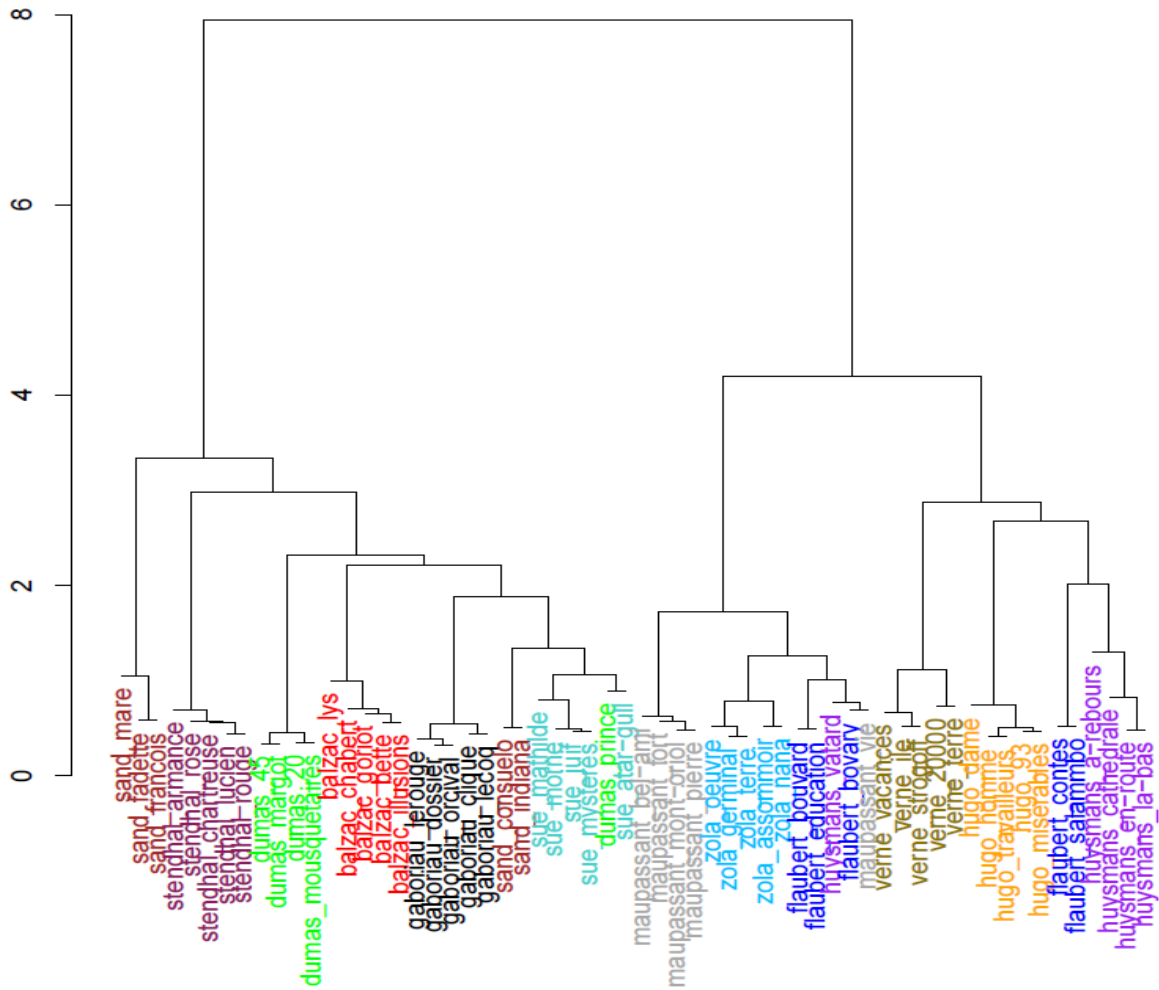
Segments	Specificities	sub freq	tot freq
je ne savoir quoi de	Inf	18	18
! dire il en PRES	Inf	26	42
, dire il en PRES	Inf	85	126
ne pouvoir être PASS que	Inf	12	12
dire il en PRES DETPOSS	Inf	18	26
dire il en PRES le	Inf	49	49
un NC ... INT !	Inf	13	13
, dire il en se	Inf	40	61
en lui PRES un NC	Inf	21	39
, dire il avec un	Inf	12	12
, VPS il en PRES	Inf	15	15
je me être PASS ,	Inf	16	16

- Table 5: The top 12 segments in Balzac (extracted with MI and calculated with specificities)

In sum, various statistical strategies can be adopted to present different views of the data. In the present analysis, all three strategies were used.

If the aim is identification, the non-discrete method of sequences proves to be much less competitive than the discrete method. A clustering analysis (with the classic delta distance)¹¹ based on discrete descriptors (and not on sequences) gives a consistent representation of the corpus:

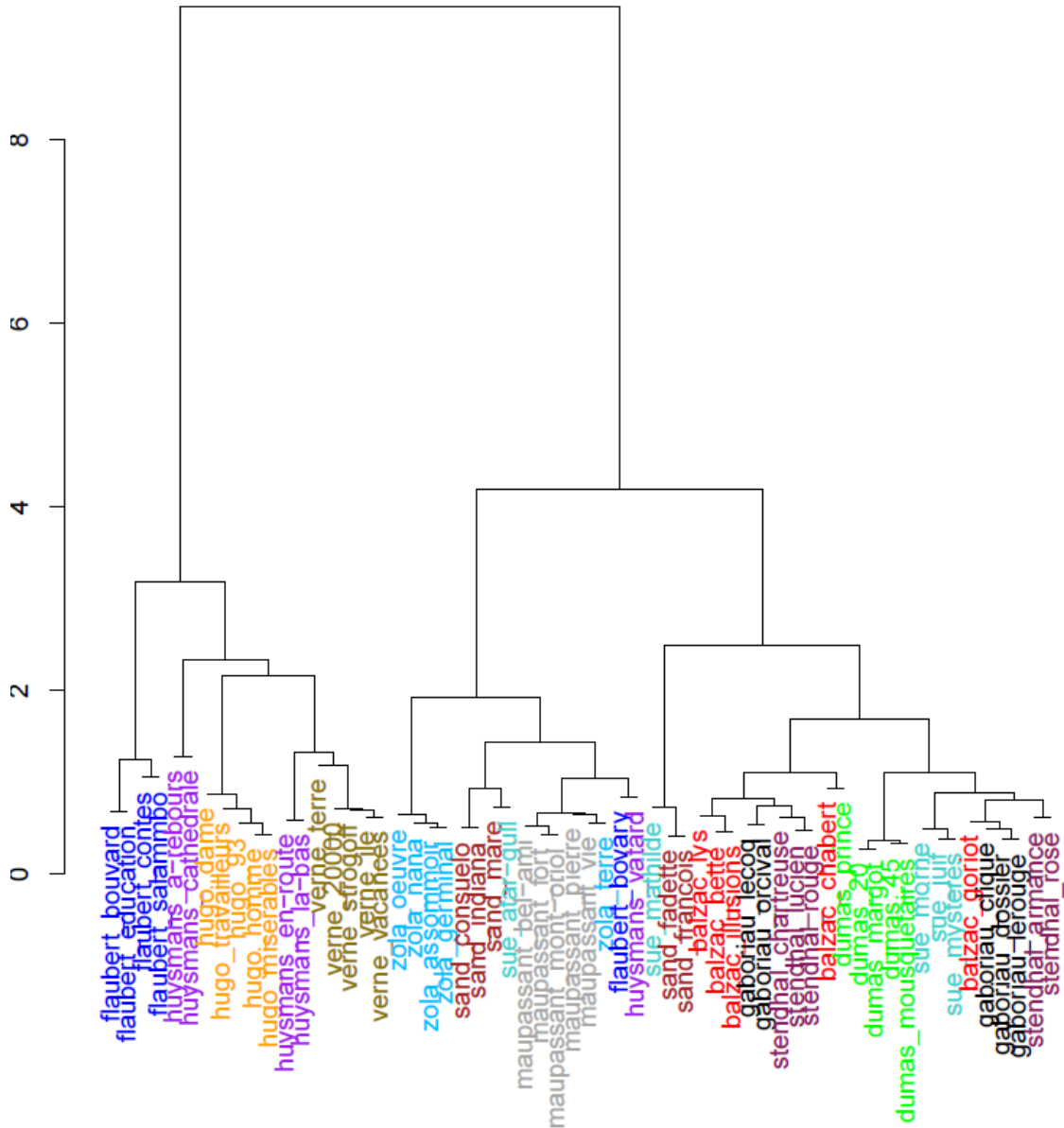
¹¹ The R package Stylo was used. See Eder et al. (2016).



- Fig. 1: Cluster analysis of the 60 novels based on discrete descriptors

The two main clusters conform to a time division: the former includes Stendhal, Sand, Sue, Balzac, Dumas, that is to say, writers of the first half of the 19th century. The latter includes all the writers of the second half. Only Gaboriau (1832-1873) belongs to the first cluster. This clustering can be considered to provide an acceptable consistency (relative to the chronology). On the other hand, a clustering based on sequences of five elements gives less consistent results¹²:

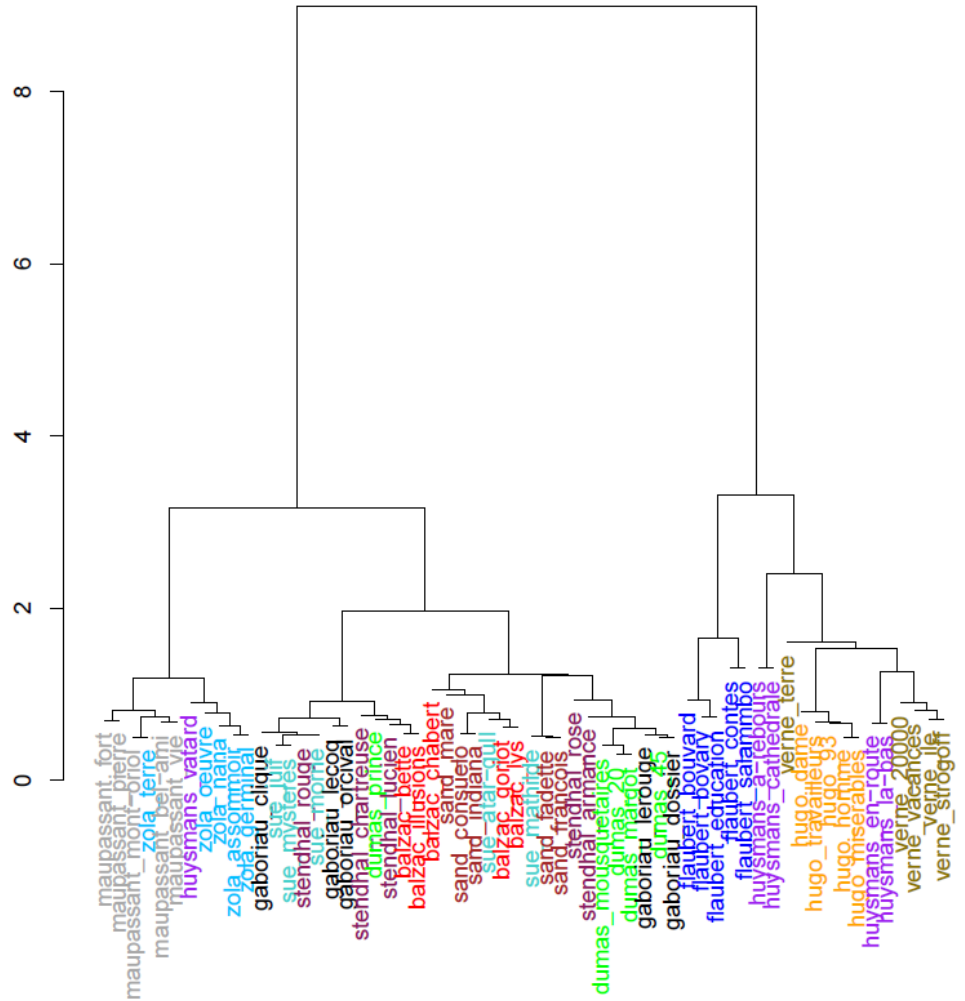
¹² The clustering is automatically made from sequences and not from motifs.



- Fig. 2: Cluster analysis of the 60 novels based on 5-grams

In the first main cluster, novels from the first half of the century (plus Gaboriau) are again grouped together, but in a sub-cluster one also finds novels by Maupassant and Zola (second half of the century). The second main cluster remains homogenous.

The clustering based on seven elements (7-grams):



- Fig. 3: Cluster analysis of the 60 novels based on 7-grams

is rather similar to the previous one, but some novels change places. For instance, *Madame Bovary* was isolated in the five-elements based clustering while in this clustering, it is now grouped with the other novels by Flaubert. This demonstrates that clustering is relatively sensitive to the size of sequences and that sequences are perhaps not the best unit to use when undertaking a stylistics of identification.

4th step

Whatever the method, it is necessary not to confuse segments and motifs. We consider as motifs only the sequences that are syntactically well formed and interpretable (because, for example, the same lexical paradigm is present in the pattern), or that have a recognizable functional or expressive role. For example, Zola uses the sequence **le NC ADJ de un NC qui VP** in order to characterize a character's typical behavior:

(4) Maheu seul marchait pesamment, butait à chaque tour contre le mur, de **l'air stupide d'une bête qui ne voit plus sa cage** (*Germinal*)

Maheu alone was walking heavily up and down the bare room, stumbling against the wall at every turn, with **the stupid air of an animal which can no longer see its cage**

(5) Bijard, (...) roulait toujours la tête, **du mouvement ralenti d'un animal qui a de l'embêtement**. (*L'Assommoir*)

lit. Bijard was still rolling his head **with the slow movement of an animal that is being goaded**

(6) Il resta quelques minutes encore, se pâma devant d'autres études, fit le tour de l'atelier avec **les coups d'œil aigus d'un parieur qui cherche la chance** (*L'œuvre*)

He remained for a few minutes longer, going into raptures before other sketches, while making the tour of the studio with **the keen glances of a speculator in search of luck**

Very interestingly, Mahlberg (2013) shows that the five-word cluster “with the air of a” appears 46 times in Dickens:

(7) ...and offered Mr. Pickwick a pinch of snuff *with the air of a* man who had made up his mind to a Christian forgiveness of injuries sustained (*Pickwick Papers*).

This cluster is functionally identical to Zola's motif but our method can identify a more lexically flexible and variable pattern than a simple lexical bundle.

It should be noted that the method generates a very large number of sequences (hundreds or even thousands for a single novel); distinguishing the motifs is therefore very time-consuming.

However, a sequence such as **le NC , et ,** , for instance, cannot be a motif, because it is not coherent and interpretable; for example, this sequence in Verne's novels:

(8) Nous marchions directement vers l'**Ouest, et**, le 11 janvier, nous doublâmes ce cap Wessel (*Vingt-mille lieues sous les mers*)

We were traveling due **west, and**, on January 11 we rounded Cape Wessel

(9) Alors Cyrus Smith éleva **la voix, et**, à l'extrême surprise de ses compagnons, il prononça ces paroles (*L'île mystérieuse*)

Then Smith raised **his voice, and**, to the extreme surprise of his companions, pronounced these words

In general, sequences extracted by MI are best able to be interpreted as motifs.

3. Results

For reasons of space, it is not possible to give exhaustive results, nor even to give a satisfactory synthesis of motifs for each of the twelve authors. We will therefore present an illustrative but very limited set of results for only five novelists: Balzac, Hugo, Stendhal, Flaubert, and Gaboriau, the author of popular novels. The motifs are in bold. The discourse function (DF) they express is in italics. All are statistically over-represented in the author's work. Their frequency in the author's texts (x) and their frequency in the entire corpus (y) are indicated as a ratio (x/y)

3.1 Balzac

DF: Reducing a set of references to an evaluative category

- , **enfin tous le NC** (+ extension) (lit. That is to say all the N/in short all of the N)
8/8

This motif is typical of Balzac since it was not identified in the other authors' novels. It is a formula in which the noun phrase subsumes, at the same time as it qualifies, all of the previously mentioned references:

(10) Rien ne démontrera mieux la singulière puissance que communiquent les vices, et à laquelle on doit les tours de force qu'accomplissent de temps en temps les ambitieux, les voluptueux, enfin tous les sujets du diable (*La Cousine Bette*).

Nothing can demonstrate more completely the strange capacity communicated by vice, to which we owe the strokes of skill which ambitious or voluptuous men can occasionally achieve or, in short, any of the Devil's pupils.

(11) - Si tu vas chez tes belles madames, je veux que tu effaces ce monstre de De Marsay, le petit Rastignac, les Ajuda-Pinto, les Maxime de Traille, les Vandenesse, enfin tous les élégants (*Les Illusions Perdues*).

If you are going to your fine ladies' houses, you shall eclipse that monster of a de Marsay and young Rastignac and any Ajuda-Pinto or Maxime de Trailles or Vandenesse, in short all of the dandies.

(12) - ...je m'aperçus que la femme, autrefois si imposante par ses sublimités, avait dans l'attitude, dans la voix, dans les manières, dans les regards et les idées, la naïve ignorance d'un enfant, les grâces ingénues, l'avidité de mouvement, l'insouciance profonde de ce qui n'est pas son désir ou lui, enfin toutes les faiblesses qui recommandent l'enfant à la protection (*Le Lys dans la Vallée*).

I now perceived that the woman, once so dignified in her bearing, showed in her attitude, her voice, her manners, in her looks and her ideas, the naive ignorance of a child, its artless graces, its eager movements, its careless indifference to everything that is not its own desire,-- in short all the weaknesses which commend a child to our protection.

The marker *enfin* expresses an enunciative stance, and therefore subjectivity, not so much at the level of the class in question as in the very process of categorization.

DF: Contrasting two categories

Two motifs fulfill this function:

- **il être un NC** (+ sub. relative) (there be a N) 32/371

The existential sentence with the verb *être* (*to be*) is more salient in Balzac than the existential sentence with the verb *avoir* (*to have*) when the discourse function is to contrast two categories (usually categories of people):

(13) Il est des femmes qui s'éprennent de la grandeur comme d'autres de la petitesse (*Les Illusions Perdues*)

Lit. There are women who are as much attracted by greatness as others by pettiness

(14) Il est des personnes que nous ensevelissons dans la terre, mais il en est de plus particulièrement chéries qui ont eu notre cœur pour linceul (*Le Lys dans la Vallée*)

Lit: There are people we bury in the ground, but there are some particularly cherished ones whose shroud is our heart.

The motif is of course also used in situations where there is no contrast but a strong assertion:

(15) Mon cher monsieur Crevel, répliqua la Lorraine, il est des noms qu'on ne prononce pas ici (*La Cousine Bette*)

My dear Monsieur Crevel, replied Lisbeth, there are certain names we never utter here.

- **plus ADJ que ne le être** (+ sujet) (more ADJ than is) 12/17

This segment is a comparative device expressing a relation of superiority:

(16) Le militaire, en temps de guerre, n'est-il pas également réservé à des spectacles encore plus cruels que ne le sont les nôtres ? (*La cousine Bette*).

Lit. Is not the soldier in time of war brought face to face with sights even more dreadful than those that we see?

(17) Je croyais à de pures amitiés, à des fraternités volontaires, plus certaines que ne le sont les fraternités imposées (*Le lys dans la vallée*).

I believed in pure friendship, in a voluntary brotherhood, more real than the brotherhood of blood.

DF: Characterizing by a remarkable property

- **un de ce NC ADJ** (one of this N ADJ) 55/257

The referent is said to belong to a class with a highly distinctive feature; this feature is supposedly known by the addressee or by the reader.

(18) Votre fille est une de ces beautés effrayantes pour les maris (*La cousine Bette*).

Your daughter is one of those beauties who rather alarm intending husbands

(19) La duchesse tourna sur Eugène un de ces regards impertinents qui enveloppent un homme des pieds à la tête (*Le Père Goriot*)

The Duchess gave Eugene one of those insolent glances that measure a man from head to foot.

Weinrich (1974) considers this stylem as specific to Balzac. It has been analyzed by Bordas (2001) who clearly showed its usage in the 19th century novel.

- **VPS le plus ADJ NC** (lit. $V_{\text{simple past}}$ the more ADJ N) 10/21

The motif marks the high degree of a property. In this structure, the adjective is anteposed instead of being in its usual postposed position:

(20) Je subis alors une conversation folle, pendant laquelle il me fit les plus ridicules confidences. (*Le Lys dans la Vallée*).

lit: I then had to put up with a mad dialogue, during which he told me the most ridiculous secrets.

(21) Elle éprouva la plus vive émotion de sa vie, elle sentit pour la première fois la joie inondant son cœur (*La Cousine Bette*).

She felt the deepest emotion of her life; for the first time she felt the full tide of joy rising in her heart.

3.2 Hugo

DF: Amplification for expressive purposes

- **PASS , PASS , PASS , PASS ,** (+ etc.) 23/33; **ADJ , ADJ , ADJ , ADJ ,** (+ etc.) 167/402; **VP ,VP ,VP , VP ,** (+ etc.) 10/13

This set of patterns is highly characteristic of Hugo's style: the juxtaposition can reach spectacular lengths.

(22) C'est en faisant la bouche en cœur du côté de ma vieille caboche que tu as taillé, coupé, tourné, viré, traîné, limé, scié, charpenté, inventé, écrabouillé, et fait plus de miracles à toi tout seul que tous les saints de le paradis (*Les Travailleurs de la Mer*)

It was in the midst of all this misery, alongside of my old craft, that you shaped, and cut, and turned, and twisted, and dragged about, and filed, and sawed, and carpentered, and schemed, and performed more miracles there by yourself than all the saints in paradise

(23) Gauvain avait devant lui l'impossible devenu réel, visible, palpable, inévitable, inexorable (*Quatre-vingt-treize*).

lit. Gauvain had before him the impossible become real, visible, palpable, inevitable, inexorable

(24) L'émeute est une sorte de trombe de l'atmosphère sociale qui se forme brusquement dans de certaines conditions de température, et qui, dans son tournoiement, monte, court, tonne, arrache, rase, écrase, démolit, déracine, entraînant avec elle les grandes natures et les chétives, l'homme fort et l'esprit faible, le tronc d'arbre et le brin de paille (*Les Misérables*).

Revolt is a sort of waterspout in the social atmosphere which forms suddenly in certain conditions of temperature, and which, as it eddies about, mounts, descends, thunders, tears, razes, crushes, demolishes, uproots, bearing with it great natures and small, the strong man and the feeble mind, the tree trunk and the stalk of straw.

The sentence grows by adding synonymous expressions, which creates an effect of variation in the meaning (Wulf 2014).

DF: Expressing the existence of a property in a human patient

- **il y avoir dans le NC** (+ GN) (there be in the N) 87/158

(25) Il y a dans le désespoir de la femme on ne sait quoi de faible qui est terrible (*Quatre-vingt-treize*).

There is in a woman's despair some indescribable weakness which is terrible to behold.

(26) Il y avait dans la femme le fond d'une brute et dans l'homme l'étoffe d'un gueux (*Les Misérables*)

In the wife there was a brutish streak and in the husband, the makings of a ruffian.

- **il y avoir de le NC dans (+GN)** 59/81

The impersonal construction here “massifies” count nouns:

(27) Il y avait de la crinière dans sa perruque (*Quatre-vingt-treize*)

Lit: There was horsehair in his wig

(28) Il y a du songe dans le tonnerre (*Les travailleurs de la mer*)

Lit: There is dream in thunder

- **le NC avoir un NC, le NC** (the N have a N, the N) 17/22

The informational structure places the focus on the identification of the “possessed” object:

(29) La révolution a un ennemi, le vieux monde, et elle est sans pitié pour lui, de même que le chirurgien a un ennemi, la gangrène, et est sans pitié pour elle (*Quatre-vingt-treize*).

Revolution has an enemy, the old world, and it feels no pity for it, just as the surgeon has an enemy, gangrene, and feels no pity for it.

(30) Le Passé a un synonyme, l'Ignoré (*L'Homme qui Rit*)

Lit: The past has a synonym, the Ignored

DF: Metaphorizing

Hugo is a writer of metaphor, a trope which he uses in all his novels, and not only in those that are sometimes referred to as philosophical (*Les Travailleurs de la mer* and *L'Homme qui Rit*). There are countless occurrences of this motif in the corpus:

- **, ce être le NC (+ expansion)** (, it be the N) 516/1386

This form is often used to express an opposition or at least a differentiation:

(31) La guerre étrangère, c'est une écorchure qu'on a au coude ; la guerre civile, c'est l'ulcère qui vous mange le foie (*Quatre-vingt-treize*).

A foreign war is a scratch on the elbow; a civil war is an ulcer which eats away your liver

(32) Dédalus, c'est le soubassement ; Orpheus, c'est la muraille ; Hermès, c'est l'édifice (*Notre Dame de Paris*).

Lit. Daedalus is the foundations; Orpheus is the wall; Hermes is the building

- **le NC être un NC (+ expansion)** (the N be a N) 254/332 ; **un NC être un NC (+expansion)** (a N be a N) 79/113

(33) La jeunesse est un plan incliné (*L'Homme qui Rit*)

Youth is an inclined plane

(34) Une prison est un habit de pierre (*Quatre-vingt-treize*)
Lit. A prison is a garment of stone

DF: Constructing contrasts

- **le NC VP, le NC VP** (the N $V_{\text{simple present}}$, the N $V_{\text{simple present}}$) 26/29 ; **le NC être ADJ, le NC être ADJ** (the N be ADJ, the N be ADJ) (35/41)

Contrast is created by simple juxtaposition, exploiting antonymic relations and lexical chiasmus:

(35) le savant conjecture, l'ignorant consent et tremble (*Les Travailleurs de la Mer*)
Lit. the scholar speculates, the ignoramus agrees and trembles

(36) le petit est grand, le grand est petit (*Les Misérables*)
lit. the little one is tall, the tall one is little

(37) le ciel est noir, l'océan est blanc (*L'Homme qui Rit*)
lit. The sky is black, the ocean is white

- **il y avoir le NC (+ et/comme) il y avoir le NC** (there be the N (and/like there be the NC) 14/19

Here the existential construction is used as a device to contrast two realities:

(38) Il y a les fumées paisibles et il y a les fumées scélérates (*Quatre-vingt-treize*)
Lit. There is peaceful smoke and there is villainous smoke

(39) Il y a la bravoure du prêtre comme il y a la bravoure du colonel de dragons (*Les Misérables*)
Lit. There is the bravery of the priest and there is the bravery of the colonel of dragoons

(40) Dans la nuit il y a l'absolu ; il y a le multiple dans les ténèbres (*L'homme qui Rit*).
Lit. At night there is the absolute; there is multiplicity in darkness

3.3 Gaboriau

Gaboriau is the French father of the detective novel.

DF: Constructing reasoning

The first three motifs correspond to hypothetical constructions; two motifs have the canonical form of *si (if)* hypothetical sentences. The third uses a different device: a thematic infinitive followed by *ce être se INF*.

- **si (+ P), ce être que il (+ P)** (if (+P), it be that he) 72/247

(41) S'il se tait, c'est qu'il n'a rien trouvé de plausible (*Le Crime d'Orcival*)

Lit. If he remains silent, it is because he found nothing plausible

(42) Or, pensait Lecoq, s'il accepte cette lutte, c'est qu'il entrevoit quelque chance d'en sortir vainqueur (*Monsieur Lecoq*)

Lit. However, Lecoq thought, if he accepts this struggle, it is because he foresees some hope of emerging victorious from it

- **si (+P), ce en être faire de NP/ DETPOSS NC** (if (+P), it be the end of NP/detposs N) 17/36

(43) Si Trémorel est jugé, c'en est fait de Laurence (*Le Crime d'Orcival*)

Lit. If Trémorel is judged, it is the end of Laurence

(44) Si vous faites du bruit, dit-il, si vous donnez l'éveil, c'en est fait de nos espérances (*Le Dossier 113*)

If you make any noise, he said, or raise an alarm, all our hopes are ruined.

- (Inf. + expansion), **ce être se INF** (It be to Inf) 10/41

(45) Rester, c'était s'exposer à une explication pénible, à des insultes, à une collision peut-être ... (*L'Affaire Lerouge*)

Lit. To stay was to run the risk of a painful explanation, insults, maybe a confrontation

(46) Laisser voir ses opinions, c'était se créer sans nécessité ni utilité une situation impossible (*Le Dossier 113*)

Lit. Revealing his own opinions would only needlessly create an impossible situation

DF: Introducing direct speech

The following two patterns are specific to the introduction of direct speech:

- **, et après un NC (+ expansion) :** (, and after a N :)9/16

(47) D'un coup d'œil, le vieux brocanteur avait embrassé ces détails, et après un sourire de remerciement à sa sœur : (*La Clique dorée*)

Lit. At a glance, the old bric-à-brac dealer had taken in these details, and after a smile of thanks to his sister:

(48) Il se redressa un peu interdit, et après un moment de méditation : (*Monsieur Lecoq*)

Lit. Surprised, he sat up, and after a moment of thought:

- **, et ce être de un NC ADJ que (+ P) :** (, and this be of a N ADJ that(+P) :) 8/8

(49) ..., et c'est d'une voix étranglée qu'il murmura : (*La Clique Dorée*)

Lit. ...and it was with a choked voice that he murmured

(50) ..., et c'est d'une voix brutale qu'il répondit à sa mère : (*Le Dossier 113*).
Lit. ..., and it was in a violent tone of voice that he answered his mother:

3.4 Flaubert

DF: Expressing the invasion of feelings

- **un NC ADJ le VIMP** (a N ADJ him V_{past}) (23/76)

(51) Une angoisse permanente l'étouffait (*L'Education Sentimentale*)
Lit. A constant anxiety stifled him

(52) Une tranquillité singulière l'occupait (*Salammbô*)
Lit. He was filled with a singular tranquility

(53) Des réflexions douloureuses l'assaillaient (*Bouvard et Pécuchet*)
Lit. Painful thoughts assailed him

DF: Making an experience known thanks to an approximation

- (transitive verb) **comme le NC de un NC** (like the N of a N) (58/242)

This hedging construction mitigates the representation of an experience:

(54) Les enfants qui chantaient des hymnes, les gerbes de lilas, les festons de verdure, lui avaient donné comme le sentiment d'une jeunesse impérissable (*Bouvard et Pécuchet*)
Lit. The children singing hymns, the wreaths of lilac, the festoons of greenery, gave him a sensation like that of imperishable youth

(55) Frédéric éprouva comme la sensation d'un coup de fouet (*L'Education Sentimentale*)
Lit. Frederick felt a sensation like that of a whiplash

(56) j'ai senti dans mon cœur comme le froid d'une épée (*Salammbô*)
lit. I felt in my heart the coldness of a sword

The motifs briefly presented in this section illustrate the relevance of the method: the motifs emerged from an automated and unsupervised analysis. Only a few of those listed above have been previously identified by stylisticians. Others require comments and developments.

4. Discussion

The motif is, we contend, a necessary and fundamental unit to link qualitative and quantitative stylistics. The examples introduced above have already shown that motifs are associated with expressive aims. It is however necessary to go further in order to highlight the convergence of some motifs towards the author's aesthetic project or world-view. We will take the example of Victor Hugo, in whose work we have identified a "family" of motifs that all in various ways express the same theme: the need to adjust categories and denominations to objects. A

characteristic feature of Hugo's aesthetics is the difficulty of describing referents or experiences using pre-established categories. Hence the need to adjust categorization, designation, or even perception.

A- The categorization of the object or experience is based on an approximate category. Several motifs express this type of categorization:

- **quelque chose comme le NC** (something like the N) 24/28 / **on ne savoir quoi de ADJ** (one does not know what ADJ) 24/24 / **on ne savoir quel NC** (one does not know what NC) 99/104

(57) C'était quelque chose comme le lever d'une montagne d'ombre entre la terre et le ciel (*Les Travailleurs de la Mer*).

Lit. It was something like the rising of a mountain of shadow between earth and heaven

(58) Elle se souvenait d'on ne sait quoi de lumineux et de chaud (*L'Homme qui Rit*)

Lit. She remembered something indescribably bright and hot

(59) Leur souffle sous le voile ressemble à on ne sait quelle tragique respiration de la mort (*Les Misérables*)

Lit. Their breath under the veil resembles some indescribably tragic breathing of death

- **Avec un sorte de NABS** (with a sort of N_{abstract}) 23/78

(60) Cosette considérait la poupée merveilleuse avec une sorte de terreur (*Les Misérables*)

Lit. Cosette looked at the marvelous doll with a sort of terror

(61) Il lui touchait les cheveux avec une sorte de précaution religieuse (*Les travailleurs de la mer*)

He touched her hair with a sort of religious awe

- **être un espèce de NC** (be a kind of N) 25/46 / **être un sorte de NC** (be a sort of N) 39/83

(62) Ce qui fait qu'une mère est sublime, c'est que c'est une espèce de bête (*Quatre-vingt-treize*)

Lit. what makes a mother sublime is that she is a kind of beast

(63) C'était une sorte de ruche monstrueuse qui y bourdonnait nuit et jour (*Notre Dame de Paris*)

It was a sort of monstrous hive, which buzzed night and day.

B- The categorization is corrected

- **être plus que un NC (ADJ), ce être un NC** (be more than a N (ADJ), it be a N) 7/10

(64) Tellmarch était plus qu'un homme isolé, c'était un homme évité (*Quatre-vingt-treize*)
lit. Tellmarch was more than an isolated man, he was avoided

(65) C'est plus qu'une morte, c'est une sainte (*Les Misérables*)
Lit. She is more than a dead person, she is a saint

C. The denomination is adjusted

- **ce qu'on pouvoir INF** (what one can + Inf) 17/52

(66) Notre-Dame de Paris n'est point du reste ce qu'on peut appeler un monument complet, défini, classé.

Lit. Notre-Dame of Paris is not what one can call a complete, defined, classified monument

- **ce NC que on VP** (this N that we V_{simple present}) 27/56

(67) C'est cette décadence qu'on appelle renaissance. (Notre Dame de Paris)

Lit. It is this decadence which is called revival

(68) Il y a un dieu pour ces ivrognes qu'on appelle les amoureux (*Les Misérables*)

Lit. There is a god for the drunkards who are called lovers

In these examples, the referent is first classified in a detrimental category (decadence, drunkard), and then it is positively recategorized.

D. The perception is adjusted

This indirect and approximate characterization of an object or situation is also present in a seemingly different motif (which is mostly used in *Quatre-vingt-treize*):

- **qui être le NC de (le NC/NP)** (that be the N of) 77/140

This relative clause is very particular: it re-categorizes with a definite noun phrase a previous indeterminate phrase that corresponds to the initial perception of a whole:

(69) Un rassemblement se pressait devant un perron de quelques marches qui était l'entrée de la mairie (*Quatre-vingt-treize*)

Lit. A crowd gathered in front of a flight of a few steps which was the entrance to the town hall

(70) À l'est apparaissait une blancheur qui était le lever du jour, à l'ouest blêmissait une autre blancheur qui était le coucher de la lune (*Quatre-vingt-treize*)

Lit: In the East a whiteness appeared which was the sunrise, and in the West another whiteness paled which was the setting of the moon

(71) Derrière cette tour se perdait dans la brume une grande verdure diffuse qui était la forêt de Fougères (*Quatre-vingt-treize*).

Lit. Behind this tower a broad expanse of greenery which was the forest of Fougères was lost in the mist

A flight of a few steps, a whiteness, a broad expanse of greenery are parts (or, better, signs) of a whole that cannot be directly identified.

Like other patterns, but in a very different way, this motif expresses the hugolian theme of the difficult relationship between the two modalities of perception and knowledge.

These examples show how a stylistics of motifs can contribute not only to identifying authors, but also to characterizing their writing.

5. Conclusion

To sum up, motifs are units chosen from segments of annotated units. They are both quantitative and qualitative in nature, in that they allow computational stylistics not to be confined to the selection of “morellian” features.

If the method is to make a real contribution to the knowledge of styles, it is advisable to list the cases where it could provide the specialist of literary texts with a genuine device for observation and analysis. We mention just briefly two possible directions for future research: How can motifs be made more informative? What types of analysis does the method allow?

It is obvious that the annotation could be enriched, notably by semantic properties. Speech and perception verbs are already considered in the method, but it would be possible to take other categories into account as well, such as psychological verbs (*émouvoir (to move)* *effrayer (to scare)*, etc.). Similarly, action nouns, deverbal or even event nouns could be specifically annotated, provided reliable lists were available. This improvement would no doubt make it possible to identify more semantic specificities than those illustrated above.

Moreover, the method still lacks indications about the topological location of motifs in the text: do some motives tend to appear at the beginning / end of the novel, chapter, paragraph, or sentence? This is a matter of “textual colligation” (Hoey 2005), which is still technically difficult to implement, but necessary to study because it can be assumed that a relationship exists between certain forms and the narrative and textual phases of novels.

Finally, correlations between motifs should also be calculated: is a pattern preferentially used with some other motif in the same context (chapter, paragraph)? This calculation would make it possible to identify not only single motifs, but sets of motifs, giving rise, perhaps, to significant constellations.

The application of the method is a crucial question, as it is essentially the applications that can contribute to its validation.

References

Bertels A. and Speelman D. ‘Keywords Method’ versus ‘Calcul des Spécificités’. A comparison of tools and methods” *International Journal of Corpus Linguistics* 18:4 (2013), 536–560.

Brunet E. 1985 *Le Vocabulaire de Zola*, suivi de l’Index complet et synoptique des [Rougon-Macquart](#), 3 tomes, Genève-Paris, Slatkine-Champion

Berenike J. Herrmann / Karina van Dalen-Oskam / Christof Schöch (2015) Revisiting Style, a Key Concept in Literary Studies, *Journal of literary theory*, 9/1, 25–5

Salem 1987 *Pratique des segments répétés. Essai de statistique textuelle*. Publication de l’INALF, collection « Saint-Cloud », Klincksieck, Paris

Fletcher, W., 2003. Exploring words and phrases from the British national corpus. <http://kwicfinder.com/BNC/>

Bordas Eric, (2001). « Un stylème dix-neuviémiste. Le déterminant discontinu un de ces... qui... », *L’Information Grammaticale*, Vol. 90, n° 1 p. 32-43.

Burrows John (1987) *Computation into Criticism: A Study of Jane Austen's Novels*, Clarendon Press.

Eder Macie M., Kestemont Mike and Rybicki Jan (2016). “Stylometry with R: A package for computational text analysis”. *R Journal*, 16(1), advance access available at: <https://journal.r-project.org/archive/>

Ganascia, Gabriel. 2001. Extraction automatique de motifs syntaxiques. In *Actes de Traitement Automatique du Langage Naturel 2001 (TALN 2001)*. Tours, 2-5 July.

Holmes, David I., “The Evolution of Stylometry in Humanities scholarship”. *Literary and Linguistic Computing* 13: 3 (1998), pp. 111-17.

Ho Yufang (2011) *Corpus Stylistics in Principles and Practice. A Stylistic Exploration of John Fowles' The Magus*. London/New York: Continuum.

Hoover, D. L. (2007). "Corpus Stylistics, Stylometry, and the Styles of Henry James". *Style* 41: 174-203.

Kastberg Sjöblom Margareta (2006). *L'écriture de J.M.G. Le Clézio. Des mots aux thèmes*. Paris, Honoré Champion.

Kilgarriff, Alan (2005). "Language is never, ever, ever, random". *Corpus Linguistics and Linguistic Theory*, 1, 2, 263-276.

Köhler Reinhard (2015) "Linguistic motifs", in *Sequences in Language and Text*, Edited by Mikros, George K. / Macutek, Ján, Mouton De Gruyter, p. 89–108.

Lafon, Pierre (1984). *Depouillements et Statistiques en Lexicométrie*. Geneva–Paris: Slatkine-Champion.

Loiseau Sylvain, (2010). "Les paradoxes de la fréquence". *Energieia*, 2, 20-55.

Longrée Dominique and Mellet Sylvie (2013). « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours ». *Langages* 189, 68-80.

Lutosławski Wincenty (1897a). *The Origin and Growth of Plato's Logic*. London, Longmans.

Lutosławski Wincenty (1897b). "On stylometry". *Classical Review* 11, 284–28.

Magri Véronique (2011). *Le Voyage à pas comptés. Pour une poétique du récit de voyage au XIXe siècle*. Paris, Honoré Champion.

Mahlberg Michaela (2013) *Corpus Stylistics and Dickens's Fiction*. New York: Routledge.

Oakes Michael (2014). *Literary Detective Work on the Computer*, Benjamins.

Pawłowski, Adam and Pacewicz, Artur (2004). “Wincenty Lutosławski (1863–1954). Philosophe, helléniste ou fondateur sous-estimé de la stylométrie?” *Historiographia Linguistica*, Vol. 31, 2-3, 423-447.

Quiniou Solen, Cellier Peggy, Charnois Thierry and Legallois Dominique (2012). “What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics?” *Lecture Notes in Computer Science*, Springer.

Rastier François (2001). « Vers une linguistique des styles ». *L'Information Grammaticale*, 89, 2001: 3-6.

Römer, Ute (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construct.*, 3, 95-119.

van Peer, Willie (1989). “Quantitative studies of literature: a critique and an outlook”. *Computers and the Humanities* 23 (4/5): 301–7.

Weinrich Harold., 1974, *Le Temps*, Paris, Le Seuil.

Wulf Judith (2014). *Étude sur la langue romanesque de Victor Hugo. Le Partage et la composition*. Paris, Classiques Garnier.

Categorial symbols:

ADJ: adjective

DETPOSS: possessive determiner

INF: infinitive

NC: common noun (N)

NCABS: abstract noun

NP: proper name

VPS: simple past verb

VP: present verb

PASS: past participle