# Towards a topological grammar of genres and styles: a way to combine paradigmatic quantitative analysis with a syntagmatic approach

Dominique Longrée, Sylvie Mellet

# Towards a topological grammar of genres and styles: a way to combine paradigmatic quantitative analysis with a syntagmatic approach

Dominique Longrée (1) et Sylvie Mellet (2)

(1) LASLA, Université de Liège et SeSLa, Université Saint-Louis Bruxelles, Belgium
(2) Université Côte d'Azur, CNRS, BCL, France

In order to contribute to the discussion herein about what makes a "Grammar of Genres and Styles", we would like to submit a methodological study based on textual analysis whose aim is to identify formal criteria for distinguishing between different discursive genres or authors' styles and characterizing them according to their linguistic properties and textual dynamics. In our previous work, we have used methods relying not only on a paradigmatic, quantitative analysis but also on syntagmatic approaches: sequences (Longrée and Luong 2003, 2005), text segmentations (Longrée, Luong, and Mellet 2004, 2006; Longrée and Mellet 2007), neighbourhoods (Mellet and Barthélemy, 2007; Luong, Julliard, Mellet and Longrée, 2007; Barthélemy, Longrée, Luong, and Mellet 2009) and bursts (Longrée, Luong, and Mellet 2008; Longrée and Mellet 2016). This work has led to a theoretical proposal to consider the text as a topological space and to introduce a new analytical unit that we call the "motif" (Longrée, Luong and Mellet 2008; Mellet and Longrée 2009, 2012; Longrée and Mellet 2013, 2014). With this methodological background in mind, we would like to assess here the benefits and limitations of both approaches – paradigmatic and syntagmatic – in the characterization of textual genres and author's styles.

## 1. The corpus and the methodology

As our previous work did, this study involves the analysis of a corpus of Latin classical texts, made up of literary works of various genres and authors. It follows a long philological tradition. For decades, classical philologists have tried to characterize styles and genres according to their lexical, lexico-grammatical, morphological, or even syntactic particularities, and they have therefore often used methods involving exhaustive counting, e.g. counting Sallustius' narrative infinitives, Caesar's historical presents (Mellet 1980), clausulas in the Latin prose (Aumont 1996), or Tacitus' postponed subordinate clauses (Seitz 1958; Kohl 1959). A book of J.P. Chausserie-Laprée, *L'expresssion narrative chez les historiens latins, Histoire d'un style,* published in 1969, is particularly representative of this kind of work: he studied a large sample of literary texts, from Caesar to Tacitus, in order to describe the evolution of historical narrative prose; he highlighted the interest of counting recurrent linguistic and stylistic phenomena; in particular, he counted occurrences of certain syntactic phrases according to their sentence positions in order to characterize different types of sentence structures.

Analysing Latin texts offers great advantages. First of all, it is a well-known, closed corpus. Secondly, since the 1960s, this corpus has been digitalized and tagged,[1] which allows for automatic counting and statistical processing.[2] This makes it possible to enhance philological studies with the modern methods of Textual Data Analysis, although it is still necessary to carry out a reflection on the theoretical concepts required for such an analysis.

We will first show that the paradigmatic approach as used by Biber (1988, 2006)[3] is useful for identifying pertinent generic classifications, but that these results are often rough and poorly informative (e.g. they build a tri-partition history / discourse / poetry). We will also show that the results of the classifications are better and more refined by taking into account a syntagmatic dimension, and that this second approach offers more accurate text characterization.[4]

Second, we will examine the available conceptual tools for the text syntagmatic approach, such as the grammatical n-grams; we will also introduce the notion of motif and we will address its relevance for our purposes.

This methodological exploration will allow us to select new, effective tools that we will use to characterize the respective styles of Caesar and Tacitus within the framework of a topological (Mellet and Barthélemy 2007) modelling of the texts. In this way, we will try to offer a new option to go beyond the paradigmatic / syntagmatic opposition of text analysis.

## 2. The potential and the limitations of the paradigmatic approach

The paradigmatic approach is herein illustrated by two Correspondence Analyses (CA). By "paradigmatic approach," we mean an analysis applied to grammatical features, building up closed classes. The first of the two CAs presented below represents the distances between the main works of the LASLA classical Latin Corpus[5] according to the parts of

[1] Digitalized, lemmatized and tagged corpus of the Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) of the University of Liege (http://www.cipl.ulg.ac.be/Lasla/tlatins.html).

[2] The software programs Hyperbase-Latin and Hyperbase Web Edition ("Bases, corpus, langage", Université Côte d'Azur) offer tools for such quantification and analysis (http://ancilla.unice.fr/ et http://hyperbase.unice.fr/).

[3] However, we must note that, in a paper published in 2009, Biber took into account the existence of multi-word patterns, including a sketch of the syntagmatic dimension.

[4] These two approaches overlap with the opposition between "language in the Mass" and "language in the Line". See Pawlowski 1999.

[5] Plaute = Plautus (*Amphitruo, Asinaria, Aulularia, Bacchides, Captiui, Casina, Curculio, Epidicus*); Caton = Cato (*De Agricultura*); Catulle = Catullus; Lucrèce = Lucretius; Gaules = Caesar, *Bellum Gallicum*; civile = Caesar, *Bellum civile*; 1_Discours, 2_Verrines, 3_Discours, 4_Discours, 5_Discours, 6_Discours, Philipp. = Cicero, *Orationes*, (all speeches divided into 7 chronologic groups); Traités_C = Cicero, *De Amicitia, De Officiis, De Senectute*; Salluste = Sallustius, *Catilina* and *Jugurtha*; GéorgEglog =Virgilius, *Georgicae* and *Eclogae*; Enéide = Virgilius, *Aeneida*; Horace = Horatius, *Carmina, Carmen Saeculare, Sermones, Epistulae;* Tibulle = Tibullus, Properce = Propertius, 1_Ovide, 2_Ovide = Ovidius (all works excepted *Metamorphosis*, *Tristes* and *Ponticae* divided into 2 chronologic groups), Quinte-Cur = Curtius; Consolatio = Seneca, *Consolationes* , Colère = Seneca, *De ira*; Bienfaits = Seneca, *De Beneficiis*; 1_Lucilius, 2_Lucilius = *Epistulae ad Lucilium* (divided into 2 chronologic groups); Traités_S = Seneca, all other treatises; Tragédies = Seneca, all

speech used in the different texts; the second CA shows the same works analysed according to the distribution of verb tenses and moods.


**Figure 1 – Correspondence Analysis – 36 textual partitions and 21 parts of speech (displaying only the distribution of textual partitions)**

This first CA (Figure 1) is applied to 21 parts of speech and 36 textual partitions. The POS are the following: substantive, verb, adjective, numeral, personal pronoun, possessive pronoun, reflexive pronoun, possessive reflexive pronoun, demonstrative pronoun, relative pronoun, interrogative pronoun, indefinite pronoun, adverb, relative adverb, interrogative adverb, negative adverb, interrogative-negative adverb, preposition, coordinating conjunction, subordinating conjunction, interjection.[6] The 36 textual partitions correspond to a very large sample of all Latin textual genres: theatre, treatise, poetry, history, speech and novel.[7] The first CA dimension (36% of information) opposes the speeches and the treatises (often written as a conversational debate) on the right side of the graph to all the other texts. Inside the theatre genre, it also opposes Plautus' comedies on the right ("Plaute") to Seneca's tragedies on the left ("Tragédies"). The second dimension (29% of information) opposes strongly in the

tragedies; Juvénal = Iuvenalis; Pétrone = Petronius; Mineures = Tacitus, *Agricola, Germania, de Oratoribus*; Histoires = Tacitus, *Historiae*; Annales = Tacitus, *Annales*.

[6] LASLA categories.

[7] See note 5.

left part of the graph history (above) to poetry (below). These two dimensions can account for 65% of the inertia. As expected, the atypical works are located near the crossing of the axes; indeed, the method cannot associate with other works the only novel of the corpus (Petronius' *Satyricon*) or the only philosophical poem (Lucretius' *De natura rerum*). In this way, this CA confirms the well-known generic classifications, and its informative power is weak.

The second CA is based on more specific grammatical criteria, i.e. the distribution of verb tenses and moods.

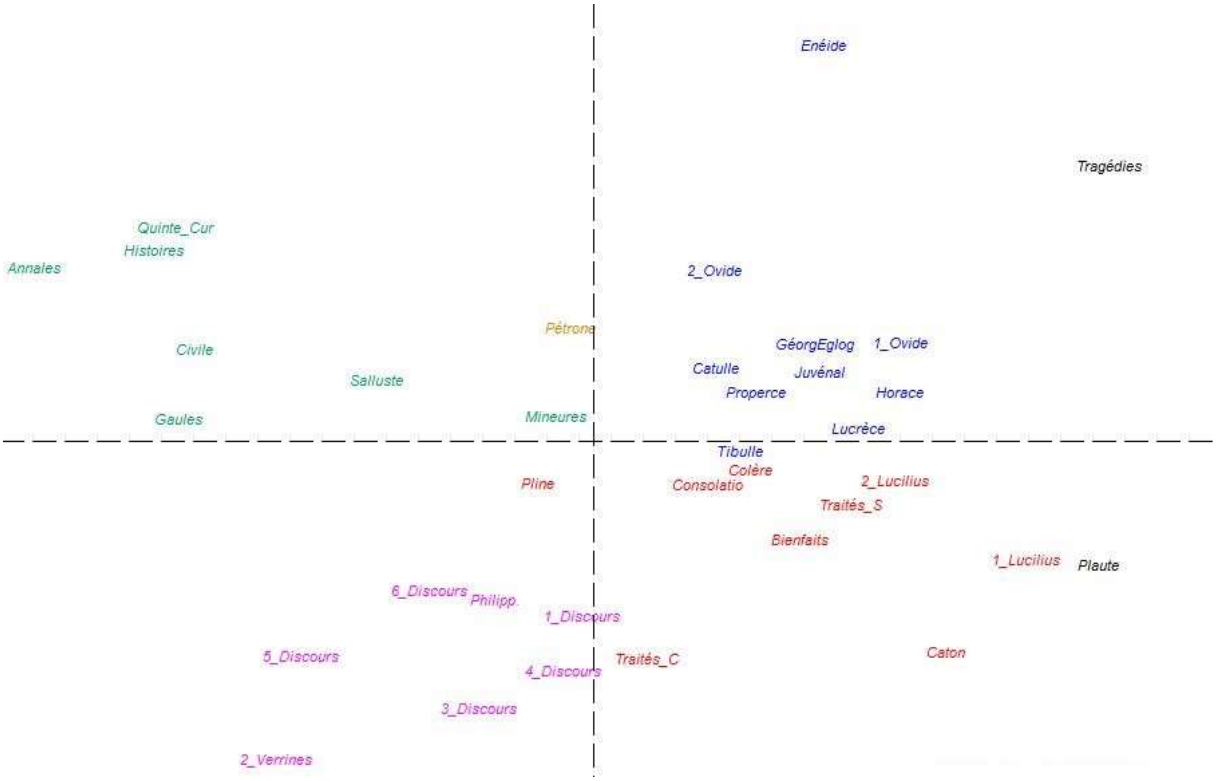


**Figure 2 – Correspondence Analysis –**
**36 textual partitions and 18 verb mood-tense associations**

This CA (Figure 2) more clearly shows groupings by text types, since each of the four most documented genres is located in a different quadrant of the graph: history to the upper left, speech to the lower left, treatise to the lower right, and poetry to the upper right. In addition, in a more refined way, we can also detect groupings by discourse modes: regarding the theatre genre, Plautus's Comedies are close to the treatises (they have in common the use of the two indicative futures), whereas Seneca's Tragedies are located in the same quadrant (even if off-centre) as the other versified texts. Near the crossing of the axes, we still find atypical works (Petronius's *Satyricon*, Tacitus's minor works), although Lucretius's *De natura rerum* has joined the other versified texts.

As we can see, the paradigmatic approach is effective for automatically classifying the texts, but it provides little new information to the linguist or the philologist. We thus made the assumption that introducing a syntagmatic approach would allow a more precise and original

classification of the texts to be obtained and, in addition, would permit the characterization of the genre and style of each text.

## 2. The contribution of the syntagmatic approach

The syntagmatic approach requires the definition of new analysis objects. Various types of such objects are available(Gledhill 2007): repeated segments, clusters, phraseological phrases, syntactic constructions, anaphoric or isotopic networks, etc. Those objects can be detected by a range of methods that can be grouped into two types: supervised and unsupervised (Ganascia 2001). Supervised detection applies to objects whose interest and meaning for textual analysis are already known, according to philological tradition. Such detection allows for a noiseless inventory and a thorough statistical investigation. While the second type of method, unsupervised detection, can reveal unanticipated structures, they are not always significant (Legallois, Quiniou, Cellier and Charnois 2012; Quiniou, Cellier, Charnois & Legallois 2012).

### 2.1. One isolated POS vs. one POS imbedded in a syntactic structure

The first test for validating the benefit of a syntagmatic approach lies within the ambit of the supervised method. We will assess the impact on genre characterization of taking into account the integration of parts of speech (POS) and grammatical categories into syntactic structures. With this aim in view, we will study the distribution of two features across the different texts of the corpus: first, the distribution of indicative perfects functioning as the predicate of a relative clause; and second, the distribution of the ablative forms used in the participial structure named *ablativus absolutus* 'absolute ablative'. We will compare the characterization power of these grammatical categories when considered in isolation (Figures 3 and 5) with the power of the same categories when included in the syntactic structures defined above (Figures 4 and 6).

Figure 3 shows the significant overuse[8] of Latin perfect indicative occurrences in texts belonging to various genres: history, speech, poetry, theatre, novel. This distribution does not allow a clear generic characterization of the texts. By contrast, the distribution of the perfect indicatives in the 3rd person singular functioning as predicates of relative clauses can clearly discriminate speeches and, to a lesser extent, treatises.

[8] Above the dotted line, the values have less than 5% chance of occurring by chance.

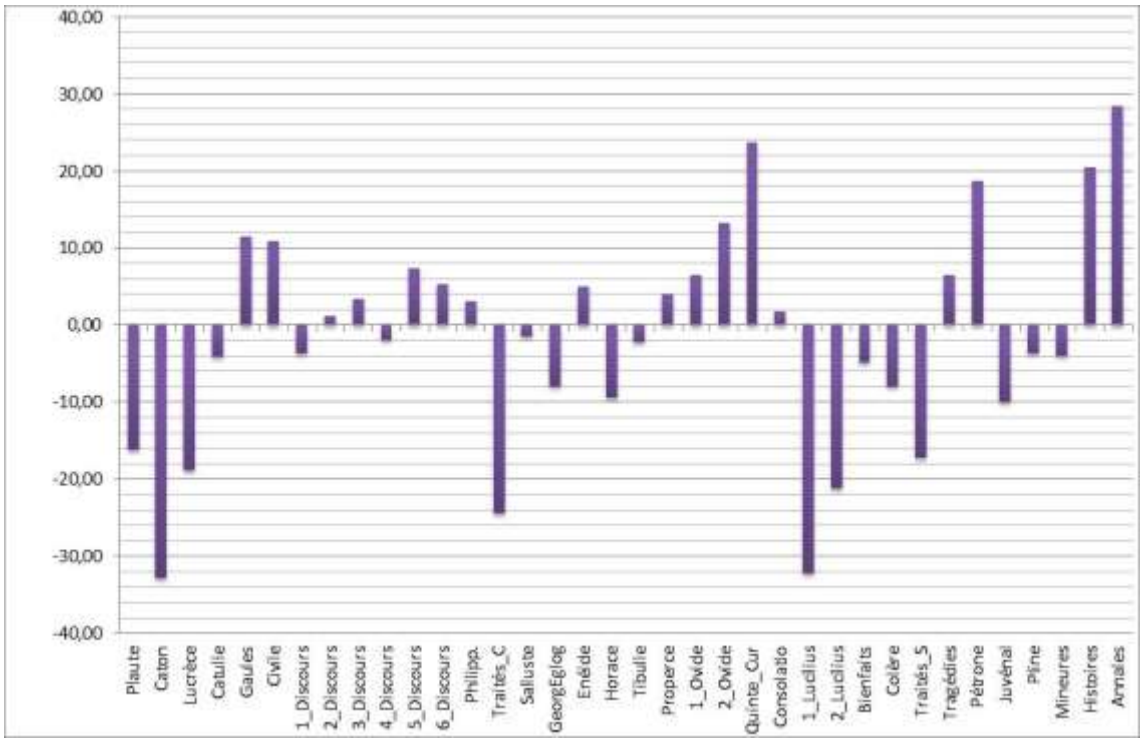


**Figure 3 – Distribution of Latin perfect verb forms**

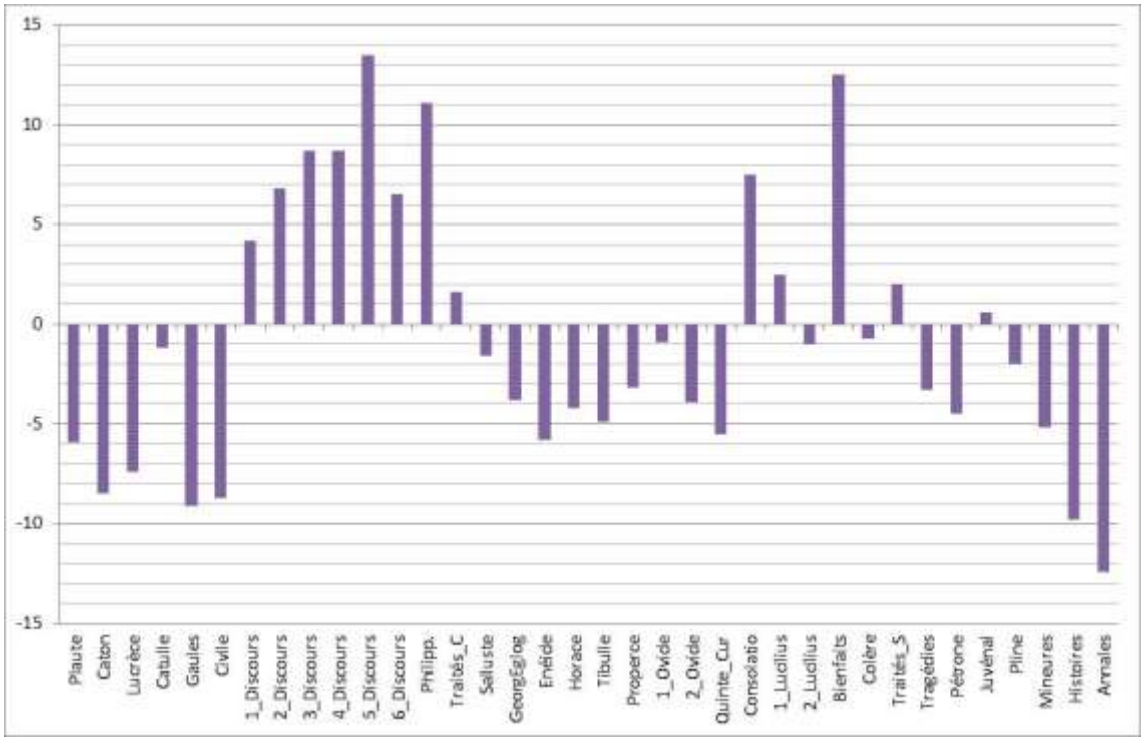


**Figure 4 – Distribution of Latin perfect indicative verb forms in the 3d person singular functioning as predicates of relative clauses**

The seven first vertical bars with a positive value represent the speeches of Cicero, and the other positive bars the treatises of Ciceron and Seneca. See examples (1) and (2),

including strings of Latin perfect indicative verb forms in the 3d person singular functioning as predicates of relative clauses, mined, in (1), from a speech of Cicero and, in (2), from a treatise of Seneca:

(1) *Cuius ut omittam innumerabilia scelera urbani consulatus,* ***in quo*** *pecuniam publicam maximam* ***dissipauit,*** *exsules sine lege* ***restituit****, uectigalia* ***diuendidit****, prouincias de populi Romani imperio* ***sustulit****, regna* ***addixit*** *pecunia, leges ciuitati per uim* ***imposuit,*** *armis aut* ***obsedit*** *aut* ***exclusit*** *senatum, ut haec, inquam, omittam…*
'For, to say nothing of his countless acts of wickedness during his consulate in the city, **during which** he **has squandered** a vast amount of public money, **restored** exiles without any law, **sold** our revenues to various people, **removed** provinces from the empire of the Roman people, **given** kingdoms for bribes, **imposed** laws on the city by violence, **besieged** the senate or **excluded** from it by force of arms, to say nothing, I say, of all this…'
(Cicero, *Philippica oratio*, 7, 15)

(2) *Nonnunquam enim magis nos obligat* ***qui dedit*** *parua magnifice,* ***qui*** *regum* ***aequauit*** *opes animo,* ***qui*** *exiguum* ***tribuit*** *sed libenter,* ***qui*** *paupertatis suae* ***oblitus est,*** *dum meam respicit,* ***qui*** *non uoluntatem tantum iuuandi* ***habuit*** *sed cupiditatem,* ***qui*** *accipere se* ***putauit*** *beneficium, cum daret,* ***qui dedit*** *tamquam non recepturus,* ***recepit*** *tamquam non dedisset,* ***qui*** *occasionem qua prodesset et* ***occupauit*** *et* ***quaesiit.***
'For sometimes indeed we feel under greater obligations **to one who has given** small gifts out of a great heart, **who matched** the wealth of kings by his spirit, **who bestowed** his little, but **gave** it gladly, **who** beholding my poverty **forgot** his own, **who had**, not merely the willingness, but a desire to help, **who though** he received a benefit when giving it, **who gave** it with no thought of having it returned, who, when it was returned, **had** no thought of having given it, **who** not only **sought**, but **seized**, the opportunity of being useful.'
**(**Seneca, *De Beneficiis*, 1, 7)

In the same way, the distribution of the ablative case is not clearly significant, as this case can function as a marker of numerous different syntactic functions, as for instance in (3) as a marker of the complement of the intransitive verb *frui* 'to enjoy':

(3) ***Insolito spectaculo*** *fruebantur…*
'They enjoyed **the strange spectacle…'**
(Tacitus, *Historiae*, 4, 62)

In Figure 5, we only observe underuses in all of Seneca's works, in the tragedies as well as in the treatises.

On the contrary, when the ablative is used in the particular participial construction called *ablativus absolutus* 'absolute ablative', as in (4),

(4) *Galli,* ***re gognita*** *per exploratores obsidionem relinquunt…*
'The Gauls, **the matter having been discovered** through their scouts, abandon the blockade'.

(Caesar, *Bellum gallicum*, 5, 49)

its distribution (Figure 6) strongly isolates and characterizes the historical works (from left to right, Caesar's *Gallic War* and *Civilian War,* Sallustius and Quintus Curtius' works, Tacitus' *Histories* and *Annals*).

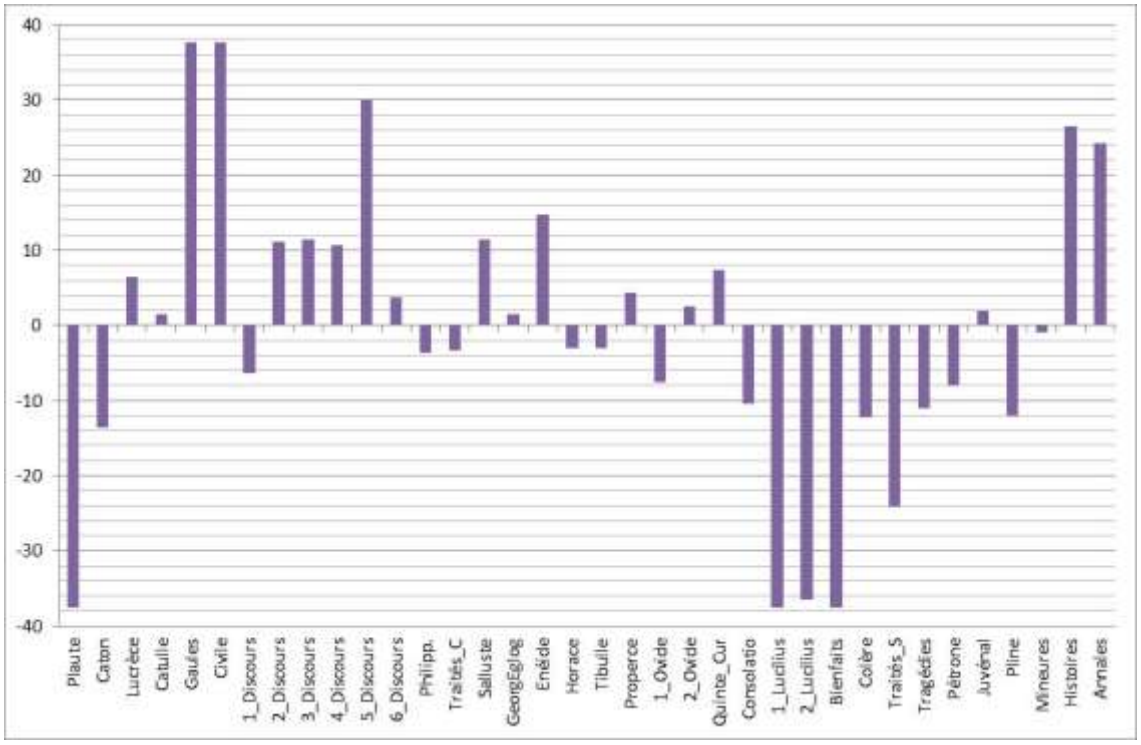


**Figure 5 – Distribution of ablative case occurrences**

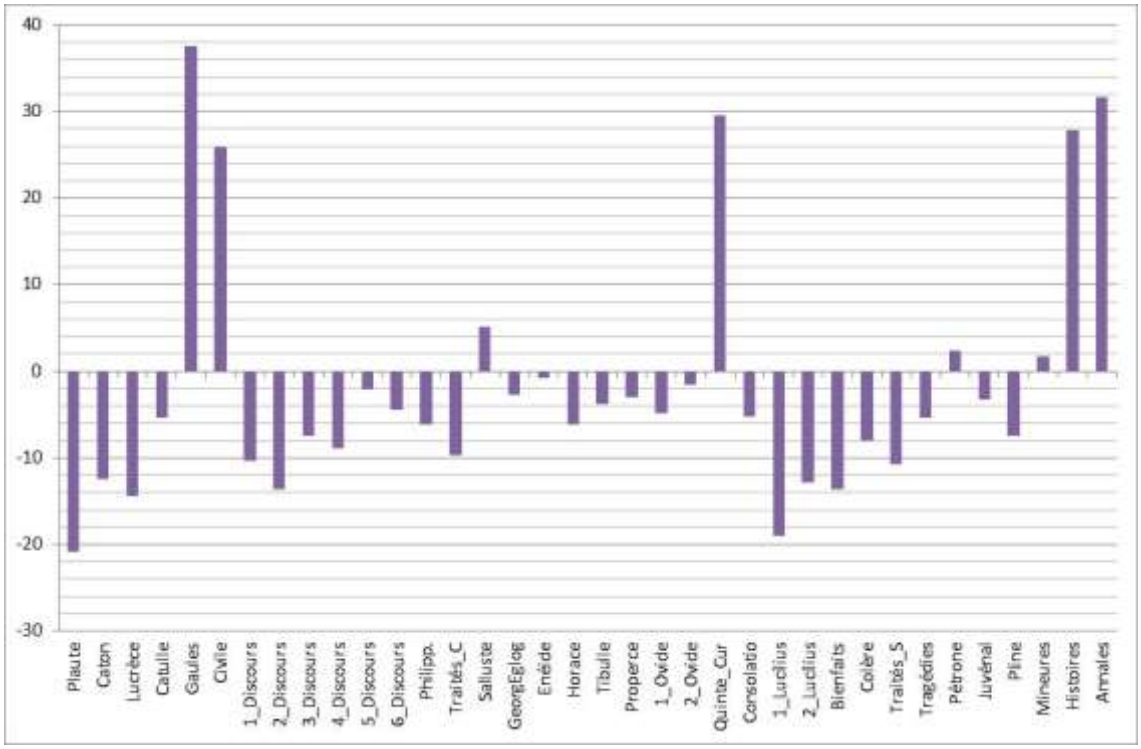

**Figure 6 – Distribution of the predicates (in the ablative case) of the participial construction called *ablativus absolutus* 'absolute ablative'**

As a conclusion of those two tests, we can note that the taking into account of the syntactic dimension produces far better results, although at this stage the results are not really unexpected and do not bring new information about the generic grouping of the corpus partition. This is not really surprising: even by leaning in this way on a syntactic dimension, this method does not take into account the real sequential structure of the text's linearity. Moreover, the choice of the studied syntactic structures relies upon the already acquired knowledge of the Latinist, and their detection is always supervised.

Therefore, studying the distribution of one grammatical category in one given structure in a supervised way is not enough. It is thus important to also analyse strings of several automatically detected grammatical categories. With this aim in view, the texts will of course be reduced beforehand to a string of morphosyntactical tags. Leaning on the POS-part of these tags, we will automatically and in an unsupervised way mine repeated strings of three POS-tags (POS-3-grams) and submit the results to a Correspondence Analysis. Then, we will focus our research on one particular POS-3-grams.

## 2.2. The POS-n-grams

The first analysis with the POS-3-grams results in a CA showing a clear bipartition between prose and poetry.

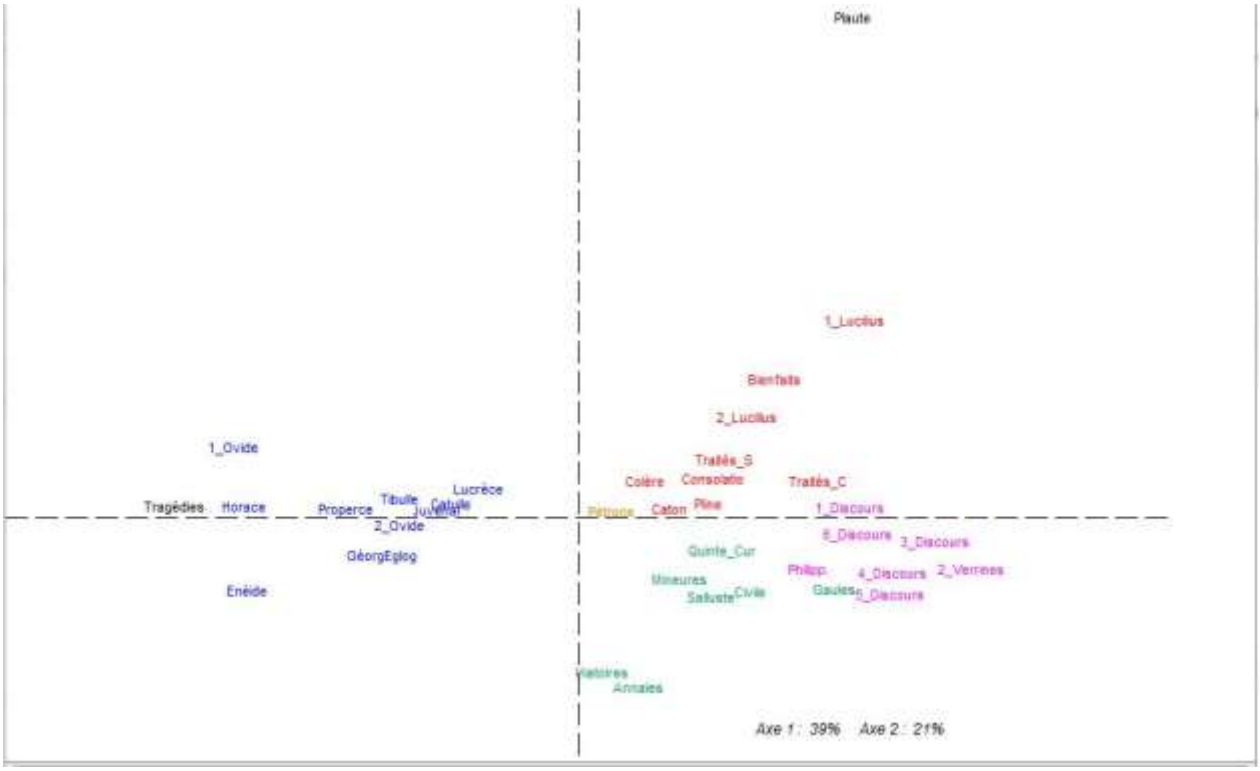


**Figure 7 – Correspondence Analysis – Distribution of POS-3-grams in 36 textual partitions (displaying only the texts distribution)**

This CA seems less informative than Figure 1 based on the distribution of isolated POS-tags, but in fact this bi-partitioned CA highlights other similarities or distances which can easily be interpreted by the philologist. For example, Lucretius' work integrates with the group of other poetic works (on the left), and Cicero's different works are more closely grouped together but without masking the distinction between treatises ("Traités_C") and speeches ("Discours", "Verrines", "Philipp").

Here, we can wonder which POS-3-grams are specific to poetry and help distinguish it from prose. When we display the POS-3-grams on the above CA (Figure 8), we observe on the far left that the POS-3-grams "cca" /adjective – adjective – substantive/ contributes strongly to this CA and shows a great proximity to the poetic works.

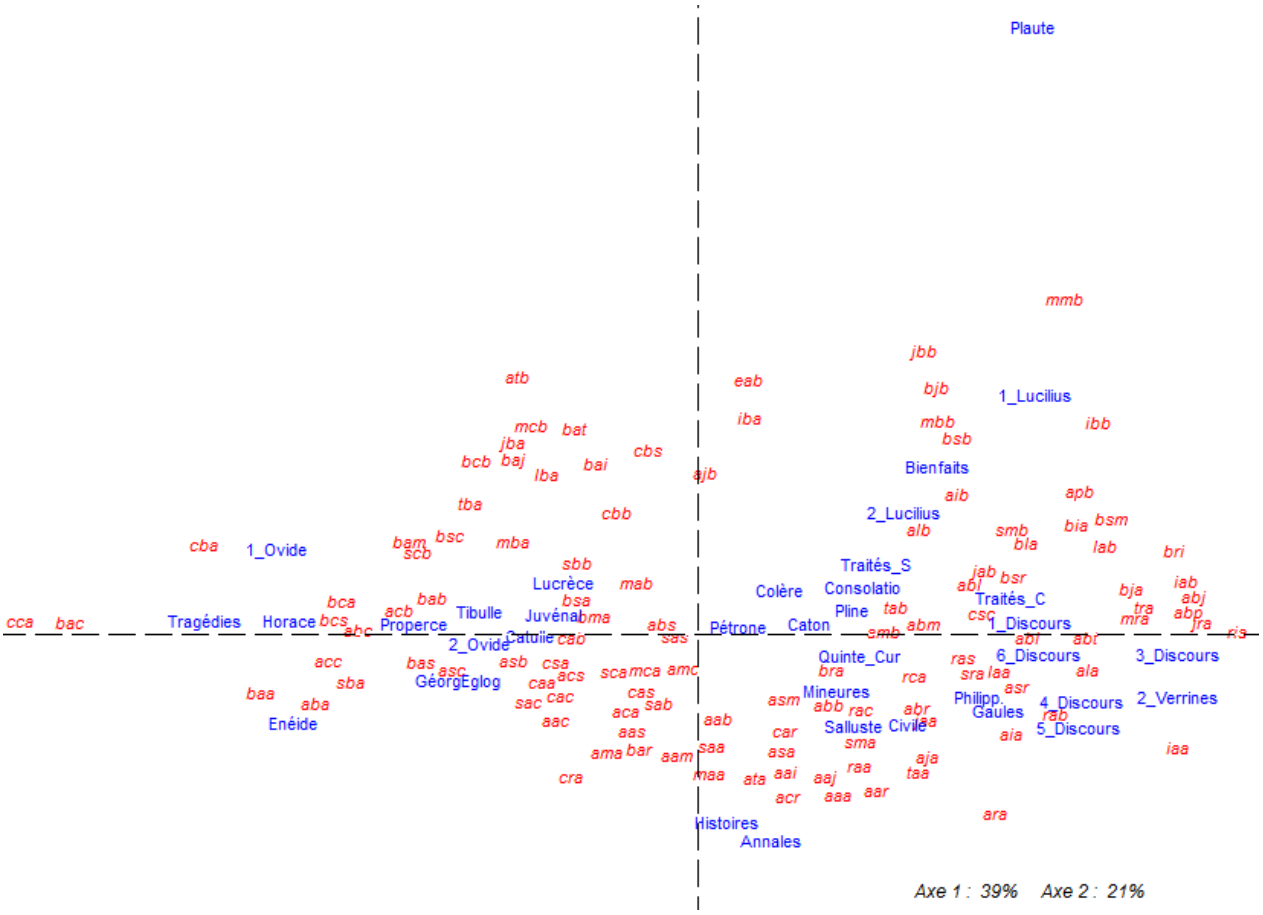


**Figure 8 – Correspondence Analysis –**
**Distribution of POS-3-grams in 36 textual partitions**
**(displaying both distributions)**

It was possible to confirm this by a graph distribution (Figure 9).

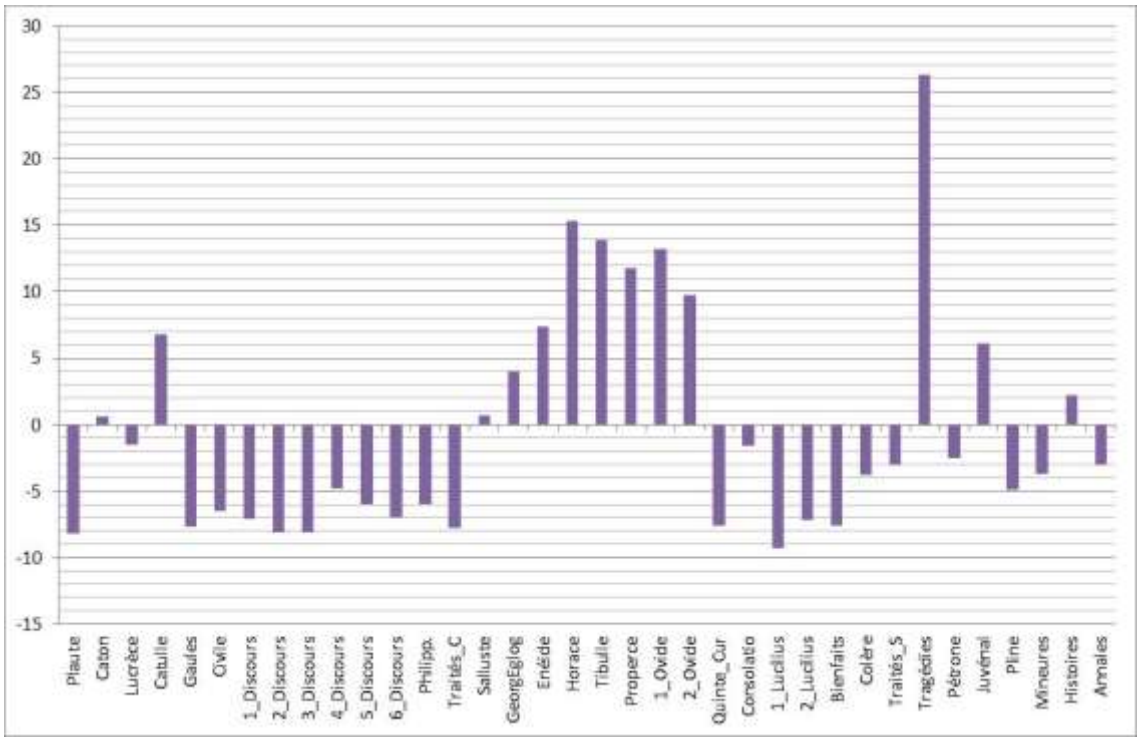


**Figure 9 – Distribution of the POS-3-grams /adjective – adjective – substantive/**

In this graph, there is only one non-poetic text, Tacitus's *Histories*, that presents a slightly significant overuse of the POS-3-Gram. This is not completely surprising: philologists have long emphasized the "poetical colour" of Tacitus's writing. With a longer POS-n-gram, the difference between the Poets and Tacitus is strengthened: for instance, the distribution of the POS-5-gram /adjective – adjective – substantive – verb – substantive/ (Figure 10) presents positive, significantly reduced variation for all poetic texts, while both of Tacitus' works show negative variations.

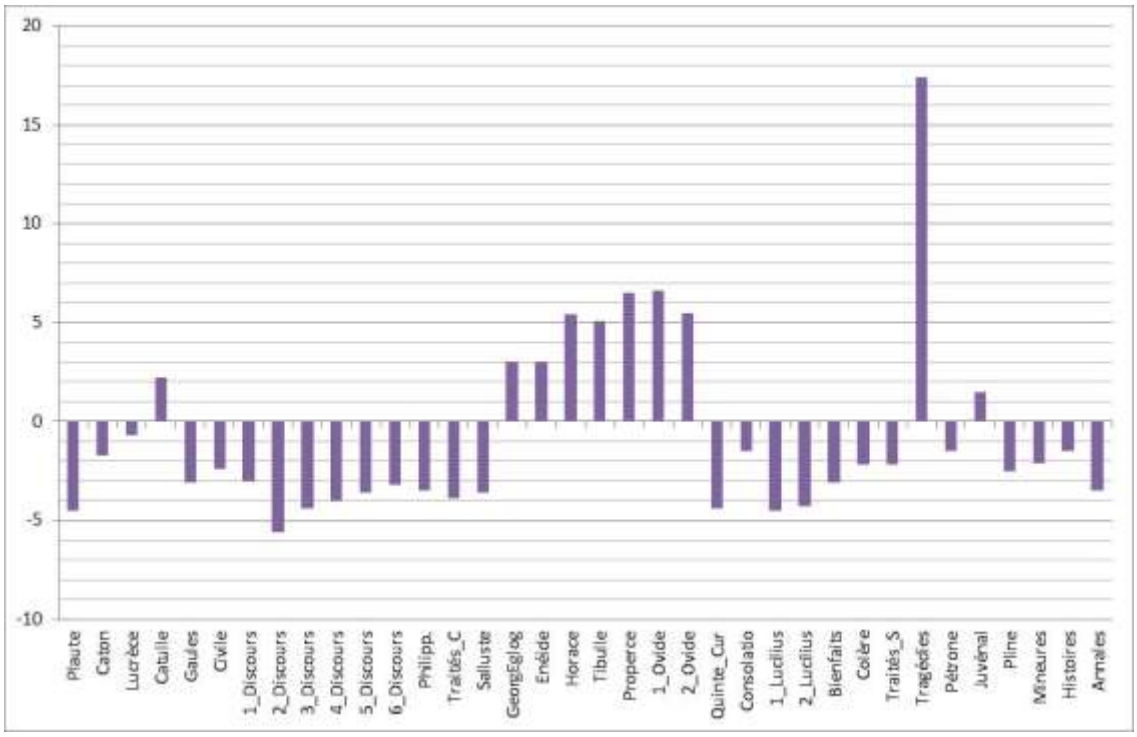


**Figure 10 – Distribution of the POS-5-grams /adjective – adjective – substantive – verb – substantive/**

POS-n-grams can also characterize authors' styles (Longrée, Mellet & Poudat 2010). The results are far better when they take into account not only the POS, but all of the information provided by the morphosyntactical tags. For instance, the string /adverb – adjective, 1st class, Nominative singular – adjective, 1st class, Nominative singular/ with or without coordination[9] between the two adjectives is a feature that is automatically and statistically detected[10] as characteristic of Sallustius's style, whereas this string is totally missing from Caesar's works. Therefore, this string helps distinguish between the writing of these two contemporary historians.

At the same time, the meaning of the POS-n-grams can be difficult to interpret. Indeed, POS strings do not necessarily correspond to a syntactic structure; there are not always grammatical links between the different elements of the string: for instance , the POS-4-grams /substantive 2d decl. accusative singular – coordination – preposition – substantive 2d decl. ablative singular/ is specific to historians, it does not correspond to a particular syntactic structure; the detection is the result of a pure statistical phenomenon due to the high frequencies of each of the constituents of the repeated segment. With our tagging and mining methods,[11] we extract different types of strings: some correspond to syntactic patterns, some

[9] This variation suggests that the repeated segments method is not sufficient to detect all the specific patterns of a text, style, or genre.

[10] By the method of the repeated segments applied to the texts reduced to a string of morphosyntactical tags.

[11] TXM, Hyperbase, Sdmc (Sequential Data Mining under Constraints).

to phraseological patterns, some are a pure succession of POS without lexical nor grammatical coherence. And for the time being, the Treebanks do not seem to offer a reliable solution.

In addition, the n-gram (or repeated string) is a frozen structure, which authorizes no variation, no addition, and no suppression. It is therefore not a suitable theoretical framework for collecting all of the various-shaped tokens in the texts of a syntagmatic pattern. Yet, the specific structures characterizing a genre or an author's style generally authorize some variations. For example, the two initial subordinate *dum*-clauses *Dum haec per provincias a Vespasiano ducibusque partium geruntur* [While these events are taking place in the provinces at the instigation of Vespasian and the party leaders] and *Dum ea geruntur* [While these events are taking place], both with the same predicate *geruntur* (indicative present of *gero*) and the same kind of subject (anaphoric pronoun neuter and plural) referring to previously narrated events are two tokens of the same pattern, but they are situated at the extreme ends of a continuum which goes from the simplest and shortest structure to the longest and the most complex. The generic pattern is characteristic and exclusive of historical prose: from the point of view of the grammar of genres, it is the same distinctive feature; therefore, to not be able to recognize and count all of its various-shaped occurrences would be a major drawback for the syntagmatic approach of genre characterization.

Finally, the n-gram corresponds to an exclusively sequential and localized approach to the text. This approach is not capable of globally comprehending the dynamics of the entire text, on which a grammar of genres and styles is based[12].

## 3. The notion of "motif"

We will therefore call now for the concept of "motif" in order to handle the different tokens of a given structure and to model them in one unified pattern. The identification of a unified pattern as a "motif" is legitimated by the fact that this "motif" always has the same textual function, regardless of its surface variations. By way of its repetition, the motif is indeed strongly related to the text dynamics and is one of its main meaning components.

What is a "motif"? Formally, the "motif" is defined as an ordered subset of the textual ensemble, formed by the recurring combination of n elements provided with its linear structure. Thus, if the text is formed by a certain number of occurrences of elements A, B, C, D, and E, a "motif" can be the recurring micro-structure ACD or AAA, etc., without here prejudging the nature (lexical, grammatical, metrical, …) of the elements A, B, C, D and E in question: the 'motif' is only the framework – or the collocational pattern – accommodating a range of parameters to be defined and which are capable of characterizing the diverse texts of a corpus, or even the different parts of a text.

[12] See the mixed conclusion in which resulted Magri and Purnelle (2012).

The concept of "motif" is one of the foundation stones of a topological approach to texts. This approach aims to account for the global dynamics of texts, both in their linearity and their reticularity (Viprey 1997, 2002a, 2002b; Legallois 2006).

The "motif" properties of recurrence and stability behind the surface variations make it a pivotal element in textual structuration. The motif is involved in particular in the temporal dynamics of the narration, in the relations between sentences, and between the different textual sequences such as descriptions, narrations, argumentations, and so on.

As general pattern, the "motif" is able to characterize a genre[13]; but its different realizations or tokens, – which we will call from now on "motif variations" – may be specific to different authors in a given genre (Longrée and Mellet 2014). We will exemplify this assertion by characterizing, in a contrastive way, the style of two Latin historians living at either end of the classical literary period, Caesar (1st century B.C.) and Tacitus (end of 1st century A.D.). With this aim in view, we selected three books of the *Gallic War,*[14] one book of the *Civilian War*, four books of Tacitus's *Annales* as well as the *Life of Agricola*, and a biography from the same author: the selection criterion was the size of the texts, in order to make a comparison possible. In this corpus, we will mine sets of characteristic motif tokens – verb tense sequences, sentence structures – and then we will observe their distribution across the texts and their meaningful collocations.

The sentence structures we will study are amongst those Chausserie-Laprée (1969) has detected as the most characteristic of narrative sentences: the variation in their use has been analysed as a marker of the diachronic evolution of narrative expression. He distinguished two main types of narrative sentences: the so-called "typical narrative sentences," and sentences with an "appended element".

All of the "typical narrative sentences" begin with a set of syntactical structures whose function is to describe the circumstances of the main action signified in the main clause. This set forms a narrative framework for the following elements and is mainly made up of participial constructions (*ablativus absolutus*) and circumstantial subordinate clauses (*cum*-clauses in the subjunctive), as in example (5):

(5) *Postridie eius diei, **refractis portis**, **cum** iam **defenderet** nemo, atque **intromissis militibus nostris**, sectionem eius oppidi uniuersam Caesar uendidit.*
'The day after, **the gates having been broken open**, **while** no one more **defended** them, and **our soldiers having been sent in**, Caesar sold the whole spoil of that town.' (Caesar, *Bellum gallicum*, 2, 33)

We have selected one "framing motif" made up of at least one occurrence of one of these two circumstantial elements: in our corpus, we have detected seven different sufficiently frequent realizations of this motif; the variations rely on the expansion of the motif to two, three, or four circumstantial elements and the intrusion of another element into the sequence.

[13] See also Stubbs and Barth (2013).

[14] *Gal*. 4 and 5 written by Caesar himself and *Gal*. 8 written by one of his legates, Hirtius.

All of the sentences with an "appended element" end with an unexpected circumstantial element which brings a complementary afterthought that provides more information about the action described in the preceding main clause, as in (6):

(6) *Pisonem Verania uxor ac frater Scribonianus, Titum Vinium Crispina filia composuere, / **quaesitis redemptisque capitibus** quae uenalia interfectores seruauerant.*
'For Piso, the last rites were performed by his wife Verania and his brother Scribonianus, for Vinius, by his daughter Crispina, **their heads** which the murderers had reserved for sale **having been searched out and purchased**.'
(Tacitus, *Historiae*, 1, 47)

We have detected three different sufficiently frequent realizations of this "appendage motif". With the seven realizations of the "framing motif", we have a set of 10 sequences that are potentially able to characterize the texts of the corpus. The data we collected have been treated by way of a Tree analysis, which allows the visualization of the grouping of the texts according to the chosen parameter, in our case, the distribution of the 10 sequences.

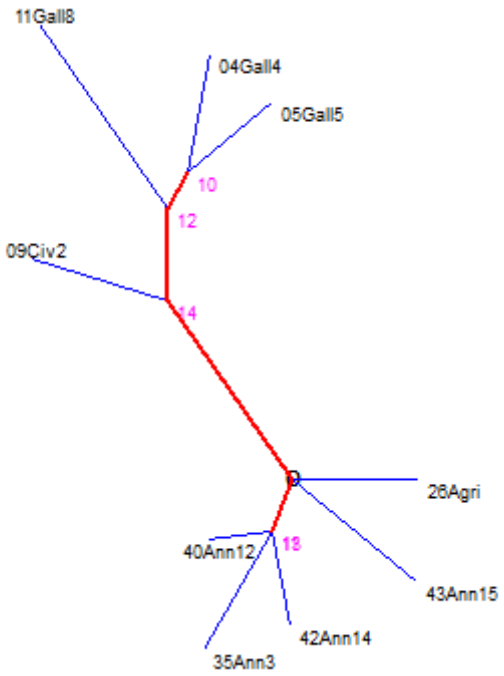


**Figure 11 – Classification of Caesar's and Tacitus texts according to the distribution of seven instances of a "framing motif" and three instances of an "appendage motif"**

This classification method (Figure 11) succeeds in regrouping all of Tacitus's works, including the *Life of Agricola*, vs. all Caesarian works including Hirtius's 8th book of the *Gallic War*.

These opposite uses of the "appendage motif" can be made obvious by comparing the linear distribution of this motif, for instance, across Caesar's *Civilian War* 2 and Tacitus's *Annales 12*. We will make use here of the "neighbourhood method" (Mellet and Barthélemy,

2007; Luong, Julliard, Mellet and Longrée, 2007; Barthélemy, Longrée, Luong, and Mellet 2009), a method we borrowed from topology. We reduce the text to a chain of codes symbolizing two types of sentence: with or without an appended element, respectively code 1 and code 0. Then, we determine a contextual sliding span of an arbitrary size, here of size 11. In each sliding span, called a neighbourhood, we count the number of codes 1. This number corresponds to the density of the motif in the span and can be graphically represented. In Figure 12, we observe that the maximal density in Caesar's book is 2 and that it is 6 in Tacitus's book.[15] In addition, the global number of "appended element sentences" is far greater in Tacitus's book than in Caesar's, and is also far more regular, whereas appended elements are completely absent in Caesar's book between sentences 92 and 169.

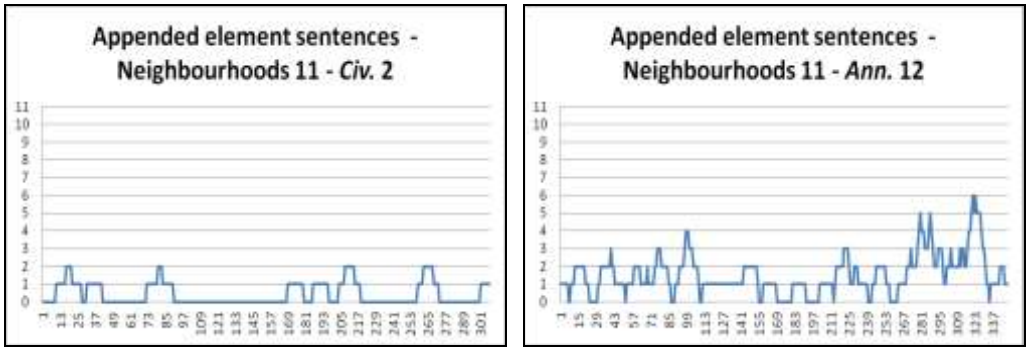


**Figure 12 – Linear distribution of the "appended element sentences" throughout Caesar's *Civilian War* 2 and Tacitus' *Annales* 12**

The same method can be used to study the linear distribution of the "framing motif sentences".

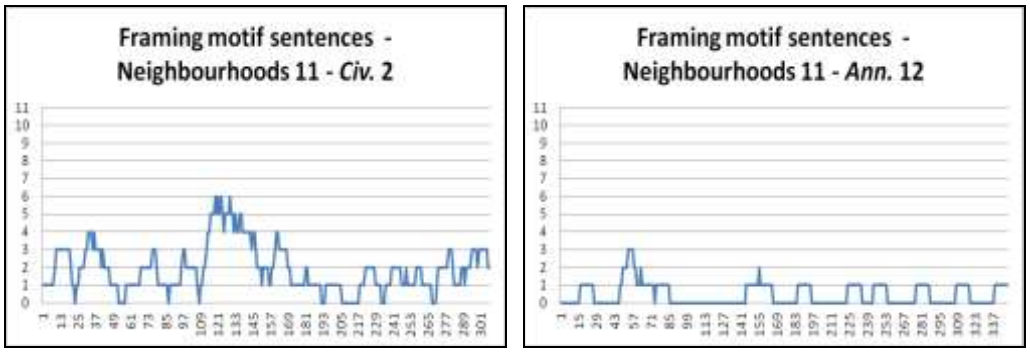


**Figure 13 – Linear distribution of the "framing motif sentences" throughout Caesar's *Civilian War* 2 and Tacitus' *Annales* 12**

Figure 13 shows a situation opposite to that present in Figure 12: the maximal density in Caesar's book is 6, and in Tacitus's book it is only 3. The global number of "framing motif sentences" is far greater in Caesar's book than in Tacitus's, and is also far more regular, while framing motifs are completely absent in Caesar's book between sentences 85 and 141.

[15] Please note that, for graphic reasons, the scale of theY-Axis has been increased tenfold.

This study highlights the capacity of our motifs to distinguish between and to characterize two different writing styles within the same literary genre. The method takes into account both a micro-syntagmatic dimension (the structure of the motifs themselves) and a macro-syntagmatic one (the overall dynamics of the text). However, it also adds a paradigmatic dimension by way of the set of so-called "motif variations": for instance, our framing motif includes seven different patterns, BCCCM, BCCM, BCCx, BCM, BCx, BxCM and BCxC, where B stands for "beginning of the sentence", C for "circumstantial element, *ablativus absolutus* or *cum*-clause", M for main clause and x for any "intruding" element. We create in this way a new closed class of syntactic structures and thereby define a new type of paradigmatic list that includes a syntagmatic dimension and that is able to characterize the grammar of an author's style. In doing this, it is truly possible to go beyond the opposition between the paradigmatic and syntagmatic approaches.

## 4. Conclusion

The various analyses proposed here raise a methodological question: because they include syntagmatic strings at various linguistic levels, the motifs that we have studied imply that the texts have been reduced to various schematic layouts, such as POS-tags or clause-type tags … We have to wonder to what extent text analysis may deconstruct and reconstruct its object.

The consistency of the results obtained through this crossed research based on various criteria at least partially legitimates this kind of deconstruction / reconstruction process. Going back to the real text brings an additional guarantee.

This methodological exploration results in the following assessment: a topological approach makes it possible to go beyond a purely sequential and localized approach to the text that is incapable of globally capturing the grammar of genres and styles. It allows the detection and mining of repeated and characteristic textual structures, as well as their grouping into paradigmatic lists of syntagmatic patterns that are able to characterize a genre or a style. This far more effective method therefore leads to an association of both the syntagmatic and paradigmatic approaches in a cross-fertilization process that goes beyond the initially observed complementarity of the two approaches.

## References


Aumont, Jacques. 1996. *Métrique et stylistique des clausules dans la prose latine. De Cicéron à Pline le Jeune et de César à Florus*. Paris: Champion.

Barthélemy, Jean-Pierre, Dominique Longrée, Xuan Luong & Sylvie Mellet. 2009. Représentations du texte pour la classification arborée et l'analyse automatique de corpus: application à un corpus d'historiens latins. *Mathematics and Social Sciences* 187 (3). 107-121.

Biber, Douglas. 1988. *Variation across language and writing*. Cambridge: Cambridge University Press.

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

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and wrinting. *IJCL* 14 (3). 275-311.

Chausserie-Laprée, Jean-Pierre. 1969. *L'expression narrative chez les historiens latins. Histoire d'un style*. Paris: de Boccard.

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

Gledhill, Christopher & Pierre Frath. 2007. Collocation, phrasème, dénomination: vers une théorie de la créativité phraséologique. *La Linguistique* 43 (1). 63-88.

Kohl, Alfred. 1959. *Der Satznachtrag bei Tacitus*, Diss., Wurtzbourg

Legallois, Dominique. 2006. Des phrases entre elles à l'unité réticulaire du texte. *Langages* 164. 56-70.

Legallois, Dominique, Solen Quiniou, Peggy Cellier & Thierry Charnois. 2012. What About Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics ? In *Lecture Notes in Computer Science*. Berlin, Heidelberg, Dordrecht: Springer.

~~Longrée, Dominique. 2005. « Temps verbaux et spécificités stylistiques chez les historiens latins : sur les méthodes d'analyse statistique d'un corpus lemmatisé ». In Gualtiero Calboli (éd.), *Papers on Grammar, IX, 2, Latina Lingua !, Proceedings of the Twelfth International Colloquium on Latin Linguistics*, 863-875. Roma: Herder.~~

Longrée, Dominique & Xuan Luong. 2003. Temps verbaux et linéarité du texte: recherches sur les distances dans un corpus de textes latins lemmatisés. *Corpus* 2. 119-140. http://corpus.revues.org/33 (accessed 18 May 2017).

Longrée, Dominique & Xuan Luong. 2005. Spécificités stylistiques et distributions temporelles chez les historiens latins: sur les méthodes d'analyse quantitative d'un corpus lemmatisé. In Geoffrey Williams (ed.), *La Linguistique de Corpus* (Rivages Linguistiques), 141-152. Rennes: Presses Universitaires de Rennes.

Longrée, Dominique, Xuan Luong & Sylvie Mellet. 2004. Temps verbaux, axe syntagmatique, topologie textuelle: analyses d'un corpus lemmatisé. In Gérald Purnelle, Cédric Fairon & Anne Dister (eds.), *JADT 2004, Le poids de mots, Actes des 7e Journées internationales d'Analyse statistique des données textuelles*, 743-752. Louvain-la-Neuve http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_071.pdf (accessed 18 May 2017).

Longrée, Dominique, Xuan Luong & Sylvie Mellet. 2006. Distance intertextuelle et classement des textes d'après leur structure: méthodes de découpage et analyses arborées. In Jean-Marie Viprey, Claude Condé, Alain Lelu & Max Silberztein (eds.), *JADT 2006,Actes des 8èmes Journées internationales d'Analyse statistique des Données Textuelles,* 643-654. Besançon: Presses universitaires de Franche-Comté. http:// lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-057.pdf (accessed 18 May 2017).

Longrée, Dominique & Sylvie Mellet. 2007. Temps verbaux et prose historique latine: à la recherche de nouvelles méthodes d'analyse statistique. In Gérald Purnelle, Joseph Denooz (eds.), *Ordre et cohérence en latin*, 117-128. Genève: Droz.

Longrée, Dominique & Sylvie Mellet. 2013. Le motif : une unité phraséologique englobante ? Etendre le champ de la phraséologie de la langue au discours. *Langages* 189. 65-79.

Longrée, Dominique & Sylvie Mellet. 2014. Les variantes des motifs chez les prosateurs latins, Entre récurrence générique et spécificité d'auteur, des formes révélatrices et caractérisantes. In Dominique Longrée, Sabine Fialon & Paul Pietquin (eds.), *Langues anciennes et analyse statistique : cinquante ans après – Distances textuelles et intertextualités = Les Etudes Classiques*, 82. 65-88.

Luong Xuan, Marcel Juillard, Sylvie Mellet & Dominique Longrée. 2007. Trees and after: The Concept of Text Topology. Some applications to Verb-Form Distributions in Language Corpora, *Literary and Linguistic Computing*, 22, 2, 167-186.

Longrée, Dominique, Xuan Luong & Sylvie Mellet. 2008. Les motifs: un outil pour la caractérisation topologique des textes. In Serge Heiden, Bénédicte Pincemin (eds.), *JADT 2008, Actes des 9èmes Journées internationales d'Analyse statistique des Données Textuelles*, 733-744. Lyon: Presses de l'ENS. http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/longree-luong-mellet.pdf (accessed 18 May 2017)..

Longrée, Dominique, Sylvie Mellet & Céline Poudat. 2010. Les taggers, auxiliaires heuristiques en ADT ? In Sergio Bolasco (ed.), *Actes des 10èmes Journées internationales en Analyse statistique des Données Textuelles*, 1195-1206. Milan: LED. http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-1195-1206_027-Longree.pdf (accessed 18 May 2017).

Magri, Véronique & Gérald Purnelle. 2012. Mot à mot, brin par brin: les suites [Nom préposition Nom] comme motifs. In Anne Dister, Dominique Longrée et Gérald Purnelle (eds.), *JADT 2012, Actes des 11èmes Journées internationales d'analyse statistique des données textuelles*, 659-673. Liège: Université de Liège. http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm (accessed 18 May 2017).

Mellet, Sylvie & Jean-Pierre Barthélemy, 2007. La topologie textuelle : légitimation d'une notion émergente. *Lexicometrica* 7. http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Magri,%20Veronique%20et%20al.%20-%20Mot%20a%20mot,%20brin%20par%20brin.pdf (accessed 18 May 2017).

Mellet, Sylvie & Dominique Longrée. 2009. Syntactical Motifs and Textual Structures. *Belgian Journal of Linguistics* 23 (New Approaches in Textual Linguistics). 161-173.

Mellet, Sylvie & Dominique Longrée. 2012. Légitimité d'une unité textométrique: le motif. In Anne Dister, Dominique Longrée & Gérald Purnelle (eds.) *JADT 2012, Actes des 11èmes Journées internationales d'Analyse statistique des Données Textuelles*, 716-728. Liège: Université de Liège. http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Mellet,%20Sylvie%20et%20al.%20-%20Legitimite%20d'une%20unite%20textometrique.pdf (accessed 18 May 2017).

Mellet, Sylvie & Dominique Longrée. 2016. A Text Structure Indicator and two Topological Methods: New Ways for Studying Latin Historic Narratives. *Digital Scholarship in Humanities*. Oxford: Oxford University Press. doi: 10.1093/llc/fqw021.

http://dsh.oxfordjournals.org/content/early/2016/04/27/llc.fqw021.full?ijkey=wDyZkoG1iV8aqRa&keytype=ref (accessed 18 May 2017).

Mellet, Sylvie. 1980. Le présent "historique" ou de "narration". Quelques remarques à propos de César, *Guerre des Gaules VII* et Charles de Gaulle, *Mémoires de Guerre*. *L'Information grammaticale* 4. 6-11.

Pawlowski, Adam. 1999. Language in the Line vs. Language in the Mass: On the Efficiency of Sequential Modelling in the Analysis of Rhythm. *Journal of Quantitative Linguistics* 6 (1). 70-77.

Quiniou, Solen, Peggy Cellier, Thierry Charnois & Dominique Legallois. 2012. Fouille de données pour la stylistique: cas des motifs séquentiels émergents. In Anne Dister, Dominique Longrée & Gérald Purnelle, *JADT 2012, Actes des 11èmes Journées internationales d'analyse statistique des données textuelles*, 821-833. Liège: Université de Liège.
http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Quiniou,%20Solen%20et%20al.%20-%20Fouille%20de%20donnees%20pour%20la%20stylistique.pdf (accessed 18 May 2017).

Seitz, Konrad.1958. *Studien zur Stilentwicklung und zur Satzstruktur innerhalb der Annalen des Tacitus*. Diss., Marbourg

Stubbs, Michael & Isabel Barth. 2013. Using recurrent phrases as text-type discriminators. *Functions of Language* 10 (1). 65-108.

Viprey, Jean-Marie. 1997. *Dynamique du vocabulaire des* Fleurs du Mal. Paris: Champion.

Viprey, Jean-Marie. 2002a. *Analyses textuelles et hypertextuelles des Fleurs du mal*. Paris: Champion.

Viprey, Jean-Marie. 2002b. Dynamisation de l'analyse micro-distributionnelle des corpus textuels. In Annie Morin, Pascale Sébillot (eds.), *JADT 2002, Actes des 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, 779-790. Saint-Malo: IRISA/INRIA.
http://lexicometrica.univ-paris3.fr/jadt/jadt2002/PDF-2002/viprey.pdf (accessed 18 May 2017).