Trees and after: The concept of text topology

Xuan Luong, Michel Juillard, Sylvie Mellet, Dominique Longrée

Trees and after:
The concept of text topology. Some applications to verb-form distributions in language corpora.

Xuan Luong
Université Nice-Sophia Antipolis, UMR. 6039 Bases, Corpus, Langage

Michel Juillard
Université de Nice-Sophia Antipolis, UMR 6039 Bases, Corpus, Langage

Sylvie Mellet
Université de Nice-Sophia Antipolis, CNRS, UMR 6039 Bases, Corpus, Langage

Dominique Longrée
Université de Liège, LASLA

Introduction

The favourite procedure in computational linguistics has always been to investigate linguistic data by resorting to a probabilistic model essentially based on frequency measures and the comparison of observed frequencies with theoretical models of distribution. The models can vary (e.g. normal law versus hypergeometric law), the measuring of the deviations between actually observed facts and theoretically expected facts can be more or less complex, but the underlying principle remains the same: the texts under scrutiny are assimilated to a set of unordered elements that are characterized only by their presence, their absence or their frequency. This of course brings to mind the time-honoured urn model, which has admittedly done statistical linguistics some service but labours under the drawback of considering only the paradigmatic dimension of the text, the syntagmatic dimension (with reference either to the overall structure of the text or to the other forms that make up its immediate context) being almost totally left out of account[1].

Our aim in this paper is precisely to go beyond this early model, in which texts are considered as mere sets suitable for conventional statistical processing, by endeavouring to take into account the shape of the text (in its entirety and in the parts making it up), i.e. to consider the text as a *topological space*. The topology of the text is defined as the incidence, distribution and density of individual linguistic features. The shape of this space can be analysed qualitatively or quantitatively in order to identify the relevant stylistic features associated with specific authors or genres. Readers interested in a more mathematical description of our approach are invited to refer to annex 1.

1. A model for text topology

Let us consider a linear structure made up of a set of individual linguistic events regarded as remarkable points of the textual chain. Each of these linguistic events

[1] Let us however draw attention to articles by P. Lafon on "rafales" or "bursts" and by D. Sérant and Ph. Thoiron (1988) on the "topographie des formes répétées" (topography of repeated forms), as well as to the THÈME function in É. Brunet's Hyperbase software. More recently A. Salem has also investigated what he calls textual topography and included a visualization tool in his software Lexico (2004).

(occurrence of a given word, of a group, of a grammatical category etc.) can – indeed must – be considered within its immediate context, i.e. included in a fragment of text containing a certain number of words, or, more generally, of other linguistic events situated before and after it. Such an interval, defined by an arbitrarily set number of elements preceding and following point X (i.e. the occurrence under scrutiny), makes up a *neighbourhood* of X. It is of course possible to modify the size of these intervals, so that to each point X is made to correspond a *family of neighbourhoods*.
The text is thus composed of a set of occurrences to which are associated families of neighbourhoods. It therefore appears that the concept of topological space supplies a formal framework and a mathematical model for the relatively intuitive notion of constructed object. For it should be noted that, like any scientific object, the object in text linguistics is itself also a construct. No scholar could claim to be working on strictly raw data, were it only because he selects his object among an infinity of possibilities according to his interests and the requirements of his research.

1.1 The model

A topological space -like any text of some dimension- is a complex structure which is difficult to control and apprehend in its entirety. We propose to build a model (fig. 1) making possible the local characterizations of the neighbourhoods and to supply for it a numerical function which is an image – albeit reduced – of our space. The operation of the model requires the development of a computerized procedure implementing the man-machine dialogue.

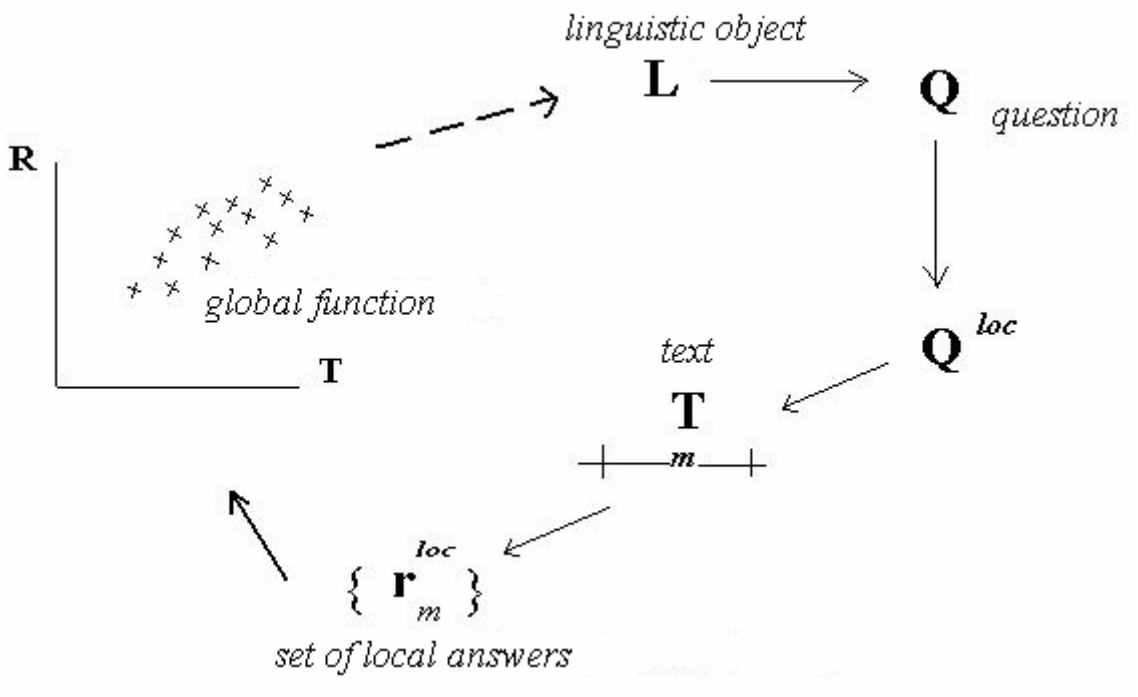


Fig. 1 The model

Let us call L the linguistic object to be studied in text T by means of one or several questions Q.

Question Q asked about L is translated into a relevant question Q^{loc}, a local as opposed to a global question, bearing on a neighbourhood of a point m of text T[2]. For example, in the study of a Latin corpus, a global question Q would typically be : « What is the normal verbal context of an occurrence of the narrative present, in other words can a single instance of the narrative present occur among the past tenses or does it have to appear as an element in a series ? ». The associated local question Q^{loc} would then become : « Given a single occurrence of the narrative present in the text, how many other instances of the narrative present and how many instances of the simple past will turn up in its neighbourhood ? ». The question is asked by means of a computer procedure which returns a numerical answer recorded as r_m^{loc} . By causing m to vary over the whole of T or a subset U of T, one obtains a set of answers $\{r_m^{loc}\}_{m\in U}$. It is then possible to produce a graph displaying a cloud of points each of which represents a local answer. This provides the scholar with the representation of a numerical function called *global function.* This figure can be used to outline and quantify certain properties of the topological space in hand. This function can then, by feedback effect, modify L and give rise to a new question Q'.

Several types of selections and constructions can be envisaged. The initial choice to perform is that of the linguistic parameters considered as characteristic of the form under scrutiny. In the case of a French language corpus it could be one or several lexical items (*immigrés*, *immigration*, *étrangers*, *migrants* ; or all the adverbs ending in *–ment*), a whole grammatical category (verbal tense for instance), a given syntactic structure (the relative clause for instance), etc[3]. Next comes the choice of the suitable "focal length" or zoom level: from the widest to the narrowest angle, this choice will result in observing alternately the distribution of relevant parameters over the whole text (even distribution as the text proceeds or on the contrary occurrences in batches or bursts, homogeneous distribution over the various parts of the text vs. occurrences figuring only in certain parts such as for example the introduction and the conclusion) or the locations in the text where special sequences occur (for instance narrative sequences made up of at least five successive occurrences of the *passé simple* or simple past). These standard examples make it necessary to specify notions of neighbourhood size and neighbourhood assessment.

1. 2. Neighbourhood assessment

Let us consider as our object of study a text reduced to the sole sequence of its verbal forms, more precisely the succession of the verbal tenses lending it structure. This corresponds to a generally significant and characteristic form. Thus, to echo the now well established distinction drawn by Emile Benveniste (1966, 1974) between discourse and history, such a form as:

---

2 The figure requires a few supplementary explanations: the method consists in the study of a linguistic object in a corpus made up of texts. According to the nature and range of the local question only certain local specificities (lexical, syntactic etc …) will be taken into account (see an illustration below with local questions on finite and non-finite verbal forms in English texts of the LOB corpus.).

3 Let us observe that the interest of text topology is greatly increased by the lemmatization and morpho-syntactic tagging of machine-readable language corpora as it is by and large grammatical categories that lend structure to the syntagmatic axis (cf. sequences of tenses, anaphoric relations, variations in enunciatory strategies, etc...).

[IMP. IMP. PQP. IMP. IMP. PS. IMP. PS. PS. PS. IMP. PQP. PS. PS. PS. PS. PS.][4]

will pertain to history and refer to an initial introductory section probably setting the frame of the events reported followed by a narrative passage in the simple past, whereas a structure restricted to the alternation between occurrences of the simple present and occurrences of the *passé composé*, interspersed with a few representatives of the imperfect or the future, such as

[PST. PST. PC. PST. PST. IMP. PST. PST. PC. PC. PST. PST. FUT. PST. PST.][5]

is obviously evocative of discourse. In the actuality of real corpora, discourse and history are often intermingled to give rise to complex forms. Within the framework of historic narrative roughly outlined in the former of the above sequences each element of the textual form under scrutiny can be characterized by a fairly narrow neighbourhood, for instance of size 5 and by a measure of this neighbourhood represented by the number of occurrences of the items in hand (here the descriptive tenses). Such a measure can remain basic by considering only occurrences of the imperfect. It can also be made more complex by taking into account the presence of occurrences of the pluperfect considered as equivalent of the imperfect for their descriptive potential, but also occurrences of participles in apposition or of certain subordinate clauses that would then have to be included in the chain and whose part in the measure could be weighted.

It appears that the properties of a neighbourhood are extremely interesting for the analysis and modelization of textual forms. Firstly, as neighbourhood size is arbitrary, hence variable, the scholar can begin working within a suitably narrow and linguistically relevant neighbourhood base (for instance the sentence) and then move on to a wider neighbourhood base, should the need arise (rarity of the phenomena investigated, necessity to handle significant units larger than the sentence, etc..). Secondly, a given neighbourhood can be characterized in two ways, both qualitatively and quantitatively. Several descriptors can be associated, so that the quality measure of the neighbourhood may depend on relatively complex properties that will ultimately account for the richness of the underlying text.

1.3. The global function

When processing is completed, each constitutive element of the textual structure finds itself associated with a numerical value representing the measure of neighbourhood. The linear chain of linguistic data can then be visualized as the sequence of the corresponding numerical values and be the object of various mathematical calculations, e.g. the automatic segmentation of the chain into different subsets, the union of subsets, automatic classification, comparison of graphs representing these numerical series, insertion into matrices for the calculation of distances, etc... Potential linguistic applications are many: automatic structuring of texts, automatic classification, comparisons for typological purposes, characterization of subsets of close texts with a view to evaluating genre, sub-genre or more simply authorial style.

---

[4] IMP = *imparfait (imperfect)*, PQP = *plus-que-parfait (pluperfect)*, PS = *passé simple (simple past)*. The sequence offered here for the needs of the demonstration is an artificial construct, but it might easily materialize in the narrative of a Latin historian or in a French classical novel, both languages making a distinction between two simple preterits, the simple past (*passé simple*) and the imperfect (*imparfait*).

[5] PST = *présent*, FUT = *futur simple*, PC = *passé composé*

But it is to be noted that the structuring and classification of texts is not the only possible object of this method of research. The topological representation of linguistic events and the model that we have just introduced also make possible an evaluation of linguistic facts themselves. Instead of being aimed at individualizing sections of the corpus, the research will then be focused on the specific distributions and the characteristic neighbourhoods of various items suitable for comparison. We supply an illustration below at paragraph 2.1.1.2.
The operative feature here is to be able to create from the same initial text as many scientific objects as is deemed necessary. The scholar is free to elaborate several global functions in order to account for the various properties of a family of neighbourhoods depending on the aims of his research. Thus, by projection on the axis of values associated to the function, he can produce several images of the topological space in hand, that can then be superposed and compared. Some of the properties selected and some of the spaces constructed may subsequently turn out to be of little interest. This is a very minor drawback easily offset by the analytical potential of the method since the procedure is by and large fully automatized and based on a modular platform for processing machine-readable, grammatically-tagged texts and easy to handle algorithmic calculations.

1.4. Computerization of the model

The model is implemented by means of a platform, developed in Sun's Java language, from which several independent modules can be operated. Each of them can be directly controled in order to perform a specific task. Here are a few examples of modules ready for use in our platform:
- storing in the central memory of the textual data, tagged or untagged, with the possibility of displaying the text on screen.
- reduction of the text to the sequence of tags pertinent to the exploitation in hand.
- use of set-theory operators: construction of a set, an intersection, a union, a difference etc...
- current statistical treatments.
- storing of results in specific files.
The questions that we ask are varied and diversified. Each of these questions is associated with a central module integrating the treatment procedures of the local neighbourhoods. We thus obtain the global function. This module is integrated into the existing environment and will require only a very limited number of computer codes and instructions.

2. A few applications

The following examples will illustrate the advantages of the method described above and give an idea of the range and variety of its fields of application. Some of them are part of more general ongoing research and have already yielded satisfactory results in the stylistic and generic characterization of texts; others have been developed here as illustrations of the method and have served no other particular purpose sofar.
With one exception, these examples all have in common to address grammatical parameters, i.e. build a topological textual structure from the distributions of various grammatical categories concerning the verb (tense, mood, etc.). This implies the use of lemmatized corpora tagged with sufficient delicacy and reliability in the field of morphosyntax. One of them is the corpus of Latin texts encoded by the LASLA (Laboratoire d’Analyse Statistique des Langues Anciennes de l’Université de Liège), another is the LOB corpus, no doubt better known to the English-speaking reader.

2.1. Local topological analysis

This initial stage of our research takes into account distribution phenomena at the micro-structural level, e.g. the recurrence of certain syntagmatic chains or the coexistence of certain linguistic items. Thus D. Longrée et X. Luong have shown that the use of sequences of variable length of verbs in the perfect or the imperfect was the ideal stylistic discriminant for each of the Latin historians represented in the LASLA corpus (Longrée & Luong 2003; Longrée 2005). Some authors, it appears, prefer to proceed by means of descriptive "bundles" of verbs in the imperfect alternating with narrative "bundles" in the perfect, while others resort to single occurrences of both tenses at regular and fairly close intervals. The corpus of English texts will be used to study the distributions and mutual relationships of respectively finite and non-finite forms of the verb in various text categories.

2.1.1 Sequences of finite and non-finite verbal forms

If one considers a text from the point of view of its syntagmatic, as opposed to its paradigmatic, dimension, it can be viewed as a sequence of entities or forms representing various grammatical categories or syntactic classes. Among these entities the verb is one of the most frequent and most interesting on account of the variety of its forms and the crucial part they play in sentence organization. Occurrences of verbs are characterized by bound morphological markers or free neighbouring morphemes coding person, number, mood, tense and aspect, as well as certain transformations (e.g. the passive). It has seemed interesting and worthwhile to apply our model of text topology to the various materializations in texts of the opposition between *finite* and *non-finite* forms of the verb, i.e. to examine the various neighbourhoods where they coexist, forming sequences or strings in the strictest sense, without any intervening foreign element as may be the case with "rafales" or bursts (see below 2.1.2). The implementation of the model will result in a topological map or chart figuring the way in which these forms are distributed within the space of the text. Before further describing our implementation of the model, it has been deemed necessary, particularly for the non-linguist reader, to introduce briefly the underlying notions of finiteness and non-finiteness.

2.1.1.1 Finite versus non-finite

Finite forms of the verb are those forms marked for tense (present, past), mood (indicative, subjunctive, imperative), person (first, second, third) and number (singular v. plural, more generally concord with the subject).
Non-finite verb forms are the forms bearing none of the marks listed above, i.e. the infinitive, the present participle, the past participle and the gerund[6].

---

[6] Modal auxiliaries (can, may, must, shall, will etc...) do not readily fit into the finite v. non-finite system since these verbal operators always figure in texts as finite elements and rarely occur alone except in cases of ellipsis. Besides, they are not marked for person and number, have no subjunctive or imperative mood and cannot undergo the passive transformation. R. Quirk lists five criteria for finite verbs, on the basis of which he suggests a gradience or scale of finiteness (Quirk, R. & al., *A Comprehensive grammar of the English language*, London, Longman, 1985, 3.52, p. 150). On the question of finiteness v. non-finiteness, see also Huddleston, R., *Introduction to the grammar of English*, Cambridge University Press, 1984, pp. 81-84 & 207-209 Huddleston, R., & Pullum, G. K. (2002) *The Cambridge Grammar of the English language*, Cambridge, Cambridge University Press, passim.

In theory, both finite and non-finite verb forms can occur individually as sole exponents of the simple verb phrase:

He *runs* daily. (finite)
Running daily will keep him fit. (non-finite)

In actual texts occurrences of finite forms are often associated with non-finite forms in complex verbal groups where the obligatory element or head[7] is typically a finite form of the verb. There is a practical limit to the theoretical number of non-finite forms that can be carried by an initial finite form. R. Quirk illustrates this limit with the following complex string (*ibid*., p. 154):

They must have been expected to have been being paid well.

The sequence of verbal forms can be symbolized as Fnnnnnnn
where F=finite and n=non-finite.

The cohesion, the consistency and the great linguistic delicacy of the principles that governed the tagging of the LOB corpus made it relatively easy to distinguish between finite (F) and non-finite (n) forms in the samples selected for this topological exploration[8].

2.1.1.2 Some results

The model makes it possible to embrace at one glance the way in which the forms under scrutiny occupy the space of the text. Figure 2 is a fraction of what we call the topological chart of the text.The whole chart gives an overall visual representation of the relative density, as single occurrences or in strings, of finite (F) and non-finite (n) forms in the section of the corpus being explored.

[7] On the notion of head, see Quirk et al., *op. cit.*, 2.27, p.61.

[8] For instance, we realized with relief that infinitive TO had been assigned a tag different from that of the homograph preposition. This judicious tagging choice greatly facilitated the distinction between infinitives and conjugated base forms.

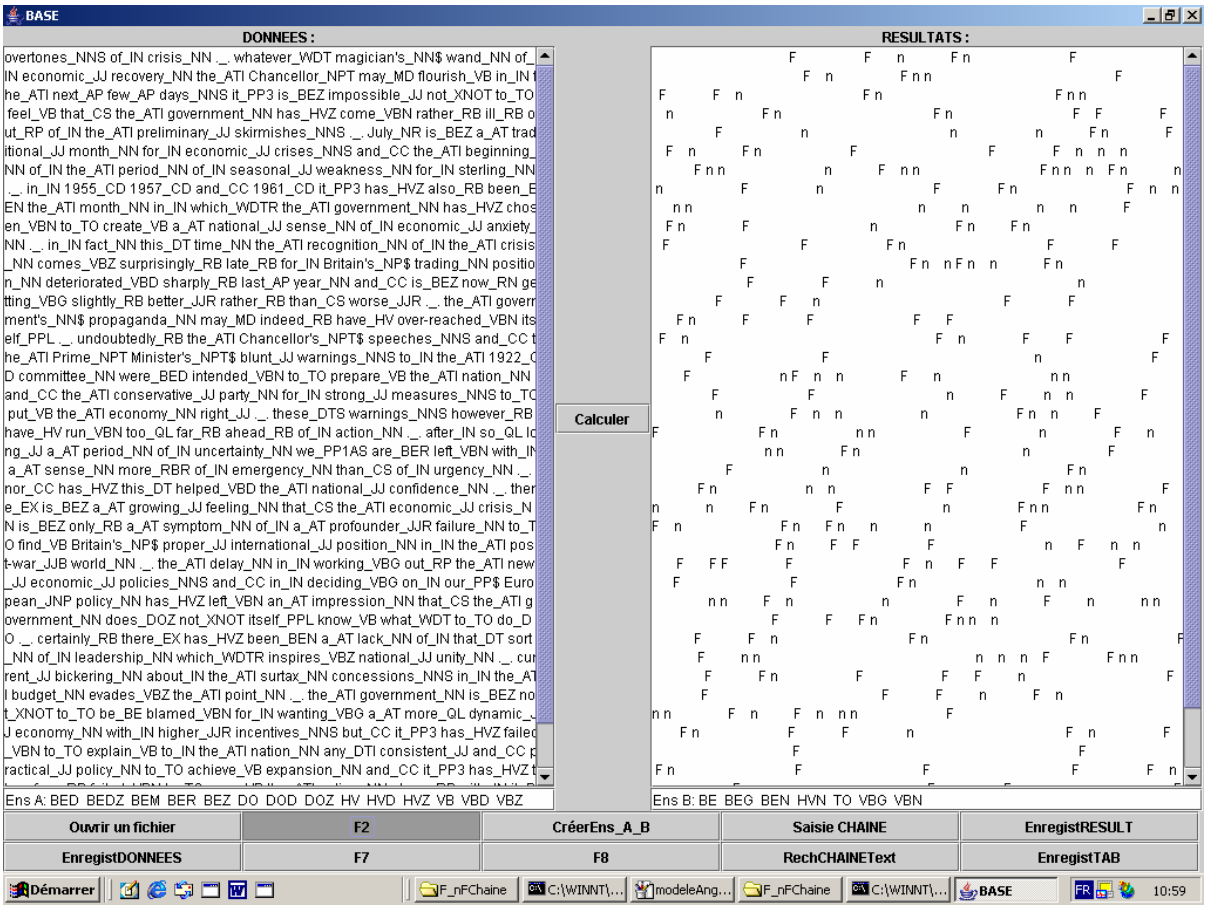


Figure 2 Topological chart (extract).

Linguists, particularly stylisticians, will favour both a more compendious and comprehensive view of the results as they are naturally interested first and foremost in the way in which all the finite and non-finite forms are combined in the verbal groups of the actual texts.

C01	C7	C8	C17	C31	C32	P01
F 1426 FF 8 FFn 1 Fn 314 FnF 5 Fnn 62 FnnF 1 Fnnn 2 n 610 nF 1 nFn 1 nn 86 nnn 5	F 1310 FF 10 FFn 3 Fn 338 FnF 1 FnFn 1 Fnn 61 FnnF 1 FnnFn 1 Fnnn 3 n 629 nF 4 nFn 1 nn 109 nnF 1 nnn 12	F 979 FF 9 FFF 1 Fn 230 FnF 1 Fnn 45 FnnF 2 Fnnn 4 n 473 nF 3 nn 65 nnF 1 nnn 9	F 1391 FF 10 FFn 4 Fn 351 FnFn 1 Fnn 51 FnnF 1 Fnnn 6 n 713 nF 8 nFn 3 nFnn 1 nn 70 nnn 3	F 1158 FF 13 FFF 1 FFn 1 Fn 316 FnF 3 Fnn 61 Fnnn 3 n 711 nF 4 nFn 1 nn 123 nnn 8	F 874 FF 6 Fn 308 FnF 3 FnFn 1 Fnn 75 Fnnn 6 n 653 nF 6 nFn 2 nFnn 2 nn 125 nnF 1 nnFnn 1 nnn 7 nnnnn 1	F 4667 FF 48 FFn 11 Fn 965 FnF 3 FnFn 1 Fnn 119 Fnnn 5 n 1338 nF 2 nn 126 nnn 12
C35	C36	C45	C53	C02	C03	C04
F 1319 FF 14	F 1079 FF 7	F 2333 FF 20	F 2290 FF 12	F 1337 FF 13	F 1455 FF 10	F 1372 FF 10

FFn 1 Fn 279 FnF 5 FnFnn 1 Fnn 51 Fnnn 1 n 569 nF 6 nn 82 nnF 1 nnFn 2 nnn 9	FFF 1 FFn 1 Fn 323 FnF 3 Fnn 43 Fnnn 3 n 585 nF 4 nFn 3 nn 96 nnF 2 nnn 4	FFn 3 Fn 486 FnF 2 FnFn 1 Fnn 80 Fnnn 2 n 592 nF 2 nn 50 nnn 2	FFn 3 FFnn 1 Fn 393 FnF 2 Fnn 52 Fnnn 1 n 479 nF 3 nn 36	FFn 1 FFnn 1 Fn 356 FnF 2 FnFn 2 FnFnn 1 Fnn 68 Fnnn 4 n 589 nF 8 nFnn 1 nn 101 nnn 8	FFn 1 FFnn 1 Fn 313 FnF 1 FnFn 1 Fnn 61 Fnnn 3 n 595 nF 6 nn 64 nnF 1 nnn 6 nnnn	FFn 1 Fn 352 FnF 3 Fnn 81 FnnF 1 Fnnn 5 n 566 nF 4 nn 66 nnn 4
D01 F 2751 FF 21 FFn 4 FFnn 1 Fn 584 FnF 5 FnFF 1 Fnn 73 FnnF 1 Fnnn 2 n 1065 nF 2 nFn 3 nn 157 nnF 1 nnn 9	E01 F 2677 FF 28 FFn 1 Fn 622 FnF 6 Fnn 91 Fnnn 4 n 1421 nF 15 nFn 4 nn 212 nnF 1 nnn 17 nnnn	A02 F 2827 FF 20 FFn 2 FFnn 1 Fn 665 FnF 4 FnFn 1 Fnn 142 FnnF 1 Fnnn 8 n 1162 nF 10 nn 130 nnF 1 nnn 10 nnnn	H11 F 2032 FF 19 FFF 1 FFFFFFFFF FFn 1 Fn 624 FnF 6 FnFn 1 Fnn 136 Fnnn 9 n 1364 nF 10 nFn 3 nFnn 2 nn 248 nnF 1 nnFnn 1 nnn 15 nnnnn	N01 F 4642 FF 45 FFF 1 FFn 3 FFnn 1 Fn 888 FnF 5 FnFn 1 Fnn 123 FnnF 1 Fnnn 4 n 1248 nF 6 nn 99 nnFn 1 nnn 6	K01 F 4481 FF 49 FFn 5 Fn 860 FnF 1 Fnn 113 FnnFF 1 Fnnn 1 n 1381 nF 9 nn 98 nnF 1 nnn 4	G51 F 2577 FF 34 FFF 1 FFn 5 Fn 593 FnF 6 FnFnnn 1 Fnn 119 FnnF 1 FnnFnnn 1 Fnnn 7 n 1287 nF 9 nn 165 nnF 2 nnFn 1 nnn 10

Table 3 Distribution of F/n strings in the corpus

The model makes it possible to identify, classify and count all the syntagmatic chains made up of any number of F and n. Table 3 supplies the identities and the number ccurrences of these various strings in subsections of 10 or 20 samples of different categories or genres in the LOB corpus[9].

The fact that no real language text of some length can exist without conjugated verbs is reflected in the relative stability of the numbers of single occurrences of F. There are however some striking variations in these absolute frequencies of single finite forms. The inflected verbal units are twice as frequent in the corpus samples of mystery and detective fiction (C45) and of romance and love story (C53). This predominance of single finite forms as exponents of the verbal group is offset by the relative rarity of long strings concatenating exponents of F and n. Both characteristics can be considered as stylistic features of the language varieties in hand. The texts in our selection showing

[9] Ten 2000-word samples for subsections relabelled C01, C02, C03 (from LOB corpus section A press reportage), C4, C7, C8 (from LOB corpus section B press editorial), C17 (from LOB corpus section F popular lore), C31 and C32 (from LOB corpus section H miscellaneous: government documents etc...), C35 and C36 (from LOB corpus section J learned and scientific writing), C45 (from LOB corpus section L mystery and detective fiction), C53 (from LOB corpus section P romance and love story), Twenty 2000-word samples for subsections P01 (romance and love story), D01 (religion), E01 (skills), A02 (press reportage), H11 (miscellaneous: government documents etc...), N01 (adventure and western fiction), K01 (general fiction), G51 (belles lettres).

the greatest complexity and variety in finite/non-finite strings are press texts, both reportage (C01,C02, C03)) and editorials (C4, C7, C8) as well as scientific texts (C35 and C36), miscellaneous: government documents and foundation or industry reports (C31 and C32). The insistent frequency of finite forms in some samples is a sign of the explicit presence of the subject in certain corpus genres or varieties where dialogue and narrative predominate. Conversely, given the constraints of language cohesion, the relative rarity of F-forms almost invariably points to strictly referential or speculative prose with low language-user involvement and a high frequency of the third person and the passive in longer than average sentences[10].

Even if our approach is not strictly probabilistic it must be owned that some of the strings are intriguing on several counts, particularly their length and the simultaneous presence of several exponents of F.

The implementation of the model makes it possible to extract and visualize the natural language counterparts of these sequences in F and n. Sequences of several finite forms of the type FFF....F correspond to enumerations of full verbs in the same tense or under the dominance of the same modal[11]. More challenging at first sight are those sequences of more than one non-finite element headed by a single F form that ultimately illustrate the dominant role of the passive voice in certain varieties or registers. As an example, here are the textual counterparts of the six occurrences of the string Fnnn in subset C17 (popular lore):

> opportunity has (F) been (n) taken (n) to revise (n) the course ...
> a man had been seen fleeing from Vauxhall station ...
> a man had been seen leaving the train at Wandsworth ...
> it has been noticed to possess another graphic name ...
> would have been used to weigh bales of wool ...
> organisers have been asked to take up the question ...

The presence of several F forms in any given string corresponds most of the time to a case of embedding. Thus the sequence FnnFn in C7 (press editorials) points to a relative clause embedded as complement of the subject of a verb in the passive[12]:

> the publicity with which the scheme has (F) been (n) launched (n) has (F) made (n) much of the gaiety ...

Such sequences, even though they are not mentioned in standard grammars, are no exceptions, as is proved by the following FnFnn examples (respectively C02, press reportage and C35, learned and scientific writing), which represent slight variations on the structure of the previous item:

---

[10] For more details on this aspect of language cohesion see: Juillard, M. (2005), *Avatars de la cohésion dans les corpus*, in Jaubert, A., ed., Cohésion et cohérence, études de linguistique textuelle, Langages, Lyon, ENS éditions, pp. 175-194.

[11] In our coding options we deliberately conflated modal and accompanying base form as one finite element.

[12] The move towards greater informality of style in press editorials detected by I. Westin had obviously not been completed by the time the LOB corpus was assembled (Westin, I. (2001), *Language change in English newspaper editorials*, Language and Computers - Studies in practical linguistics, 44, Rodopi, Amsterdam/New York).

> the new air force which president Tshombe is (F) forming (n) have (F) been (n) delivered (n) to Elizabethville ...
>
> the amount it is (F) displaced (n) is (F) said (n) to depend (n) upon the distance separating ...

These complex multiple sequences are not the preserve of highly specialized registers of contemporary English; they can be shown to occur in still more elaborate guises in texts aimed at a fairly large public, as is the case with our last two examples culled from the Belles lettres section of the corpus:

> proved that his wife from whom he was (F) separated (n) had (F) been (n) summoned (n) to give (n) evidence against him ...
> the amount of discovery and inquiry with which he has (F) been (n) surrounded (n) has (F) been (n) intended (n) to stimulate (n) his curiosity ...

These examples illustrate the reliability and versatility of our model for the topolological exploration of texts. The results are interesting in themselves, they also shed light on the structure of the verbal group in actual English texts and could ultimately serve as a basis for selecting and ordering elements of a new grammar of the English verb in use.

2.1.2 Distribution in "rafales" or bursts

If one is interested in the distribution of various items along a text, it is easy to realize that some of them evince a fairly regular distribution while others seem to occur in batches or "rafales" to borrow Pierre Lafon's felicitous image. "Rafales" are distinct from sequences or chains in that they do not imply the immediate succession of the items under consideration; thus three contiguous imperfects in the succession of the verbal tenses making up a text form a sequence whereas the presence of three, four or five imperfects that are not necessarily contiguous in a segment of text including ten verbal forms will make up a "rafale" provided the overall density of imperfects is lower. The notion of "rafale" was first used by Pierre Lafon in the study of lexical items in order to spot changes of theme within a text. When a word has a highly informative semantic load, this type of distribution is by and large linked with the thematic progression of the text. Conversely, in the case of more neutral, nondescript, words, this distributional phenomenon can point to less superficial linguistic features linked either with the author's style or with some intrisic properties of the item in hand. In both cases a close look at the distributions can yield information reaching well below the surface of the text.
Such was the hypothesis that we formulated a few years ago concerning the latin verb *coepi* "begin/start to". Philologists have as a matter of fact suspected this verb of starting to develop into an auxiliary without ever achieving the full transformation. In certain texts it does work as an inchoative auxiliary but it has left no traces in romance languages. Besides, some authors use it so frequently that it has been described as a "language tic". It has seemed to us that the notion of "rafale" might well lend scientific content to the purely intuitive one of "verbal tic" or "mannerism". The actual results of our study, published in 1995, seemed rather conclusive.
That was the reason why we thought that we could devote part of the present methodological study to a comparison of the distributional regularities or irregularities of a certain number of verbs in various Latin texts, the aims being to test the stability of

this distribution in the texts and to compare the behaviour of the different verbs in order to assess if a verb's behaviour is a reflection of its semantic status.
The method is implemented in the following fashion: the texts having been lemmatized, the textual object under study is an ordered sequence of lemmas, in other words dictionary entries as opposed to graphic forms. One then places over this linear chain a window which is 101 unit wide and centred on the verb whose distribution is being examined. The window thus has a span of 50 units before and 50 units after the hinge or pivot verb. After counting the number of occurrences of the given verb encompassed by the window, one moves the window so that it is centred on the next occurrence of the that has not yet been counted, i.e. that was not within the ambit of the previous window. After moving the window in this way over the totality of the text, the results are assessed by calculating the number of windows containing respectively one occurrence, two occurrences and three occurrences of the pivot verb. A three-term characterization of the verb is thus obtained; the third figure is weighted by a factor of value 2 as it is obvious that "rafales" or bursts of three or more occurrences are more significant and must consequently be given more importance in the analysis. The procedure is repeated for the 24 verbs in the corpus that are most frequent and so can be considered as part of its basic vocabulary. The process yielded a matrix of 24 lines by 3 columns for the first text used (Petronius' *Satyricon*). Here is a tree-analysis representing the similarities and dissimilarities of 24 verbs as well as the individual figures for the three parameters of the verbs situated on the most outlying branches. In order to keep this presentation within reasonable limits, we only supply here the most characteristic data, namely the four verbs occurring most frequently in bursts (or *rafales*) and the three verbs occurring with a more even distribution.

Facio (“ do ”): 93 – 36 – 58

Habeo (“ have ”): 80 – 26 – 50

Possum (“ can ”): 62 – 10 – 21

Inquio (“ say ”): 97 – 64 – 61

Ago (“ lead, drive”): 34 – 4 – 0

Debeo (“ must, owe”): 19 – 2 – 0

Peto (“ seek, claim”): 23 – 2 – 0

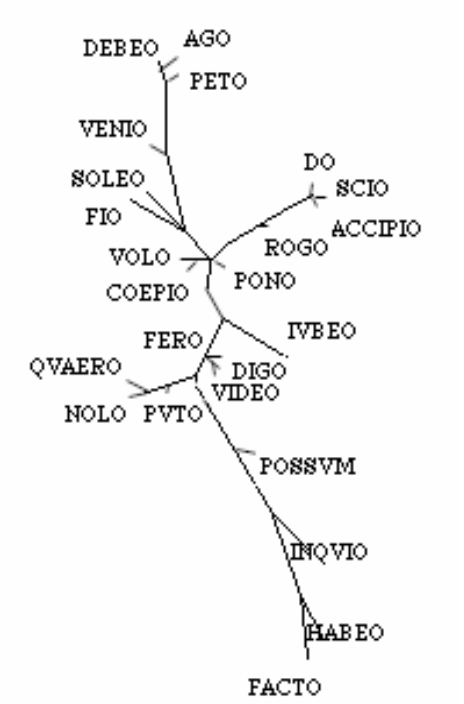


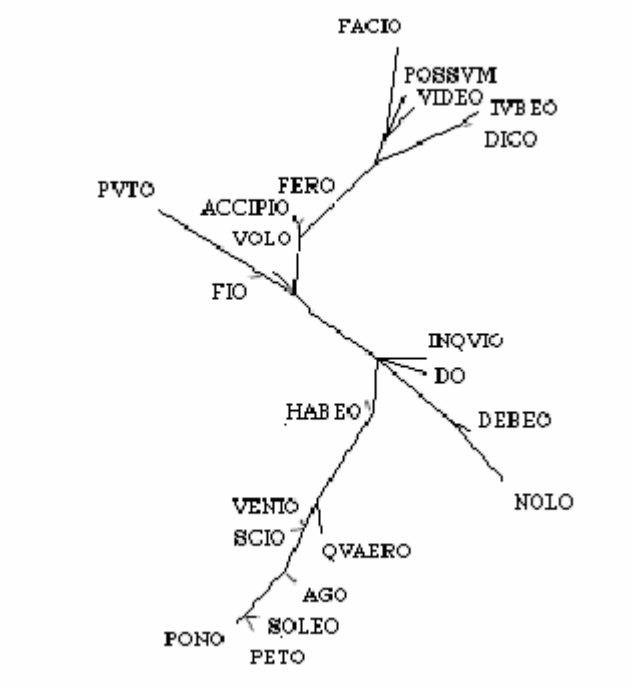
Fig. 4 PETRONIUS, *Satyricon*: occurrences in “rafales”or bursts of the 24 most frequent verbs (with weighting by 2 of the third factor)

This tree representation calls for a few comments:

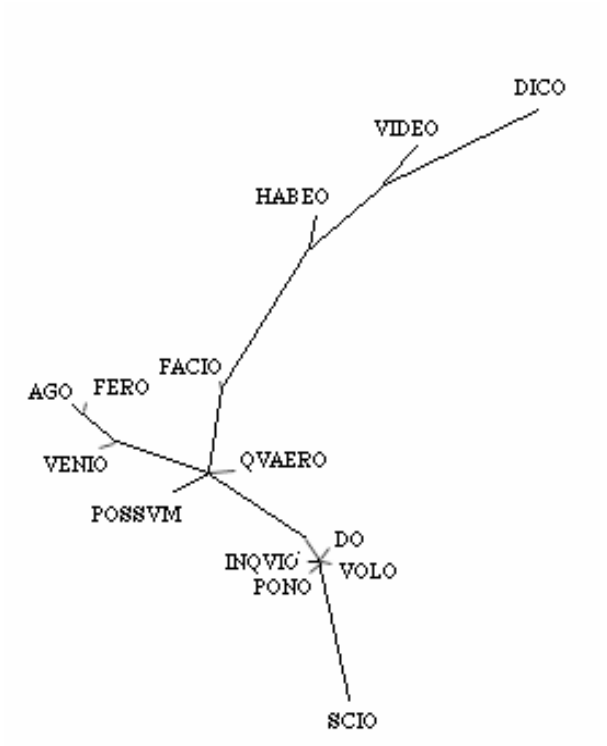
- the verbs hanging from the same branch at the bottom of the figure are characterized quantitatively by their high frequency of use (which of course goes some way to account for their occurrences in bursts or "rafales": the higher the frequency of a given item the more likely its occurring in bundles) and semantically by their relative neutrality: they are not tool-verbs properly so-called but nearly all of them can be described as prop-verbs or support verbs: *facio* “ do”, *habeo* “ have”, *inquio* “ I say, he says” (introducing reported speech), *possum* “ can”.
- The top end of the tree is occupied by verbs with more definite, albeit sometimes polysemic, lexical content: *ago* “ do, drive, lead”, *peto* “ seek, ask, claim”, *debeo* “ must, owe”, *scio* “ know”; one also notices the very close proximity of complementary lexemes: *do* “ give” and *accipio* “ receive”.
- The middle of the tree brings together a set of various verbs for which we shall offer no other comment except that this part of the tree is occupied by a few modal (*volo* “ will”, *nolo* “ will not ”) or aspectual auxiliaries (*coepio* “ start to”, *soleo* “ be wont to”) ; such verbs are not among those occurring mostly in "rafales" or bursts.

Let us now look at the distribution of the same verbs in other texts of approximately the same period (those of the previous 24 verbs with too few occurrences have been disregarded, but tree-analysis is sufficiently stable to overcome the hazard of a few missing data[13]):

[13] Tree-analysis is a classificatory topological approach (ARBORLING *Logiciel d’Analyse Arborée* – Luong’s ‘grouping method’ – CNRS UMR 6039) which begins with the most stable clusters and then moves on from cluster to cluster in order to integrate less typical elements but can be interrupted at any

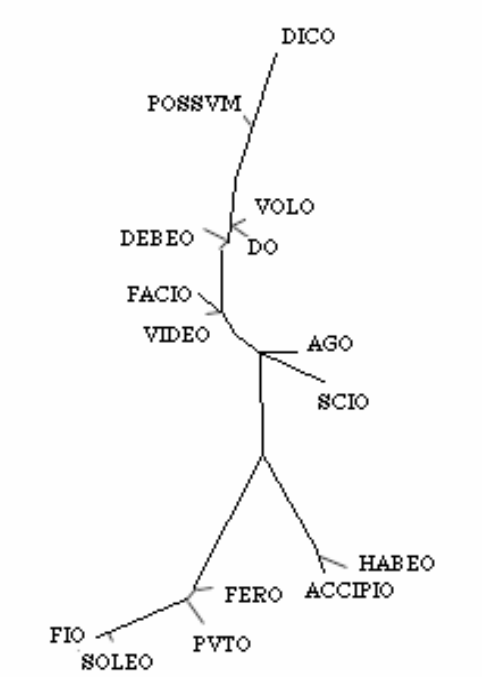


SENECA, *De Ira*



stage of the analysis without affecting the resulting classification. Thus, the analysis of the occurrences in "rafales" or bursts of the 14 most common verbs in Petronius, also figuring in Seneca's *de Vita Beata*, yields exactly the same configuration as the initial analysis of the totality of 24 verbs with a clear-cut opposition between on the one hand *ago*, *venio*, *scio*, *do* and *facio*, *habeo*, *inquio* and *possum* on the other and more sharply individualized central clusters.

SENECA, *De Vita Beata*



PLINY, *Panegyric on Trajan*

Fig. 5 The 24 verbs in Pliny, Seneca and Petronius

As the comparison between the texts proceeds, it becomes conspicuous that two of the verbs evince an invariably irregular distribution, with occurrences in bursts. They are the two basic verbs *dico* " say" and *facio* " do", closely followed by *habeo* " have" and *video* " see". These are opposed to the verbs already pointed to in Petronius: *soleo* " be wont to ", followed by *scio* " know", *ago* " lead, drive", *venio* " come" and *peto* " ask, claim".

The unvarying positioning of verb *ago* makes it necessary not to consider it as a support or prop verb, in spite of its polysemy and its ability to become part of stock phrases such as *gratias agere* " thank". That is a somewhat unexpected finding which forces one to distinguish those support verbs that are totally free, being available for frequent repetitions, and those that go into stock phrases and as a consequence become less available for undifferentiated multipurpose uses.

The next remark concerns the various auxiliaries in the list. As was already suggested by the Petronius text, these verbs do not behave in the same way: *possum* " can" is not averse to distribution in " rafales " or bursts; *debeo* " must, owe" hardly ever occurs in fits and starts, except with Pliny; *volo* " will, want" is more versatile from author to author.

Lastly, the distribution of some verbs is totally heterogeneous; this is particularly the case with *puto* " think", and with *do* " give", *fio* " become" and *fero* " carry, bear ". In this case, only by returning to the actual texts will one be able to ascertain whether or not the observed variations are due to thematic divergences between the works; one might for instance expect a philosophical treatise on generosity to give pride of place to verb *do* " give " and afford it the opportunity to appear in " rafales " or bursts in certain passages, which will be precluded in a text on another subject. And, in actual fact, the

tree-diagram of Seneca's *de Beneficiis* – albeit poorly structured and difficult to interpret – places verbs *do* “ give” and *accipio* “ receive” at the very end of the branch bearing items with occurrences in bursts, a situation where they swap places with *facio* “ do” or *dico* “ say”. These thematic motivations have already been explored by P. Lafon.
The method presented here succeeds in highlighting different behaviours for each of the verbs considered and some of these behaviours are perfectly characteristic. Another territory seems to be worth exploring: would this type of analysis, if applied to a larger body of texts, bring out generic affinities ? One may indeed think that certain types of texts will regularly favour, or on the contrary avoid, neutral, nondescript verbs such as "do" and "say" while certain authors or certain types of texts will use as a neutral verb, with possible distribution in bursts, a verb with by and large versatile behaviour like "become".

2.2. Global topological analysis

Another type of topological approach consists in widening the field of investigation in order to encompass the macro-structural phenomena that control the distributon over the whole text of the linguistic features chosen as analytical parameters. It can therefore prove interesting and productive for instance to analyze the distribution of a given grammatical category throughout a text and to assess whether its occurrences are evenly distributed or concentrated only in the introduction and the conclusion; then to wonder if the phenomenon is characteristic only of the one text or if it can be common to other texts of the same author, the same genre etc... We will introduce two methods of analysis applied to Latin texts.

2.2.1. Using segments of texts

The distribution of a linguistic feature over a text can be studied by means of graphs. Yet, such graphs are difficult to exploit (Longrée D., Luong X. & Mellet S., 2004) as there is no reliable method for comparing curves of different lengths. It is therefore necessary to revert to less qualitative methods and use tables of numerical data in order to calculate distances.
Such tables are obtained by dividing all the texts into the same number of segments. The problem of length is crucial. One cannot work with fixed spans of text as the number of columns corresponding to the profiles of the different texts would vary with their size. A text containing 300 main verbs and another containing 350 would respectively yield profiles of 30 and 35 ten-verb segments. Using "natural segments" (introduction, narrative, conclusion) cannot be envisaged either, except perhaps in the case of certain varieties (folk tales, directions for use, scientific reports) as this might introduce part of the sought answer into the very question. Besides, application to poorly structured texts would prove difficult.
The only practicable method consists in dividing the texts into the same number of contiguous segments. The method was tested with regard to the variable frequency of verbal tenses within text types, according to whether the dominant mode is Benveniste's history or discourse. In a Latin narrative, descriptive elements are typically in the imperfect or pluperfect whereas the historic framework requires the perfect or preterit. The distribution of these items in a given text should characterize the structure of its narrative, the underlying hypothesis being that, since the choice of tenses underpins the narrative structure, their overall distribution can be the mark of the in-depth organization of the text, a kind of format conditioned by the genre or sub-genre of the literary work in hand. Should this assumption be correct, one can expect that texts with

similar profiles will be brought together not only on the basis of style or period criteria but also on account of generic affinities.
This working hypothesis was tested with a corpus of Latin history texts, reduced to roughly equal (variation factor 1.2) sequences of main-clause verbs (Longrée D., Luong X. & Mellet S., 2004 et Longrée D. & Mellet S., 2006). The problem was to find the segmentation that would best serve to characterize their profiles, thus making comparisons possible.
The texts were successively divided into varying numbers of segments (20, 16, 12, 8, 6 and finally 5 segments). The aim was to find the best fit, i.e. the division yielding the best results in terms of profile definition. Each of the obtained profile matrices was Chi-square tested and graphically represented by a tree-diagram. The outcome of successive trials was that a wide-span segmentation into 5 slices yielded the most satisfactory results.

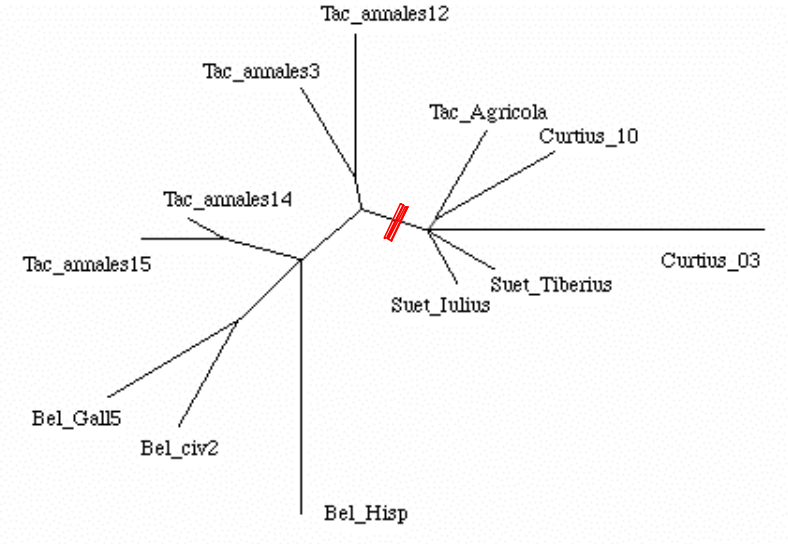


Fig. 6 Distribution of perfects in main clauses
(5-segment method)

The clusters appearing on the tree are particularly revealing in that they bring together all the texts pertaining to the sub-genre of biography which are opposed to all the other texts (see annex 2), for which the clustering is looser and somewhat expected (proximity of books 14 and 15 of Tacitus' *Annals*, fairly close proximity between Caesar's *De Bello Gallico* and *Civil War*, atypical character of *De Bello Hispanico*, distant from all others and loosely connected to the structure). The sub-genre criterion is potent enough to place Tacitus' *Life of Agricola* alongside the other biographies, genre taking precedence over the writer's style and personality.
Besides, the results are obtained with a relatively small number of segments, which means that the operative factor in terms of similarity or dissimilarity between texts is their division into large segments corresponding to the structure of the narrative. This is made very conspicuous in the histograms for the numbers of occurrences of the perfect in each of the five segments of the texts being investigated.

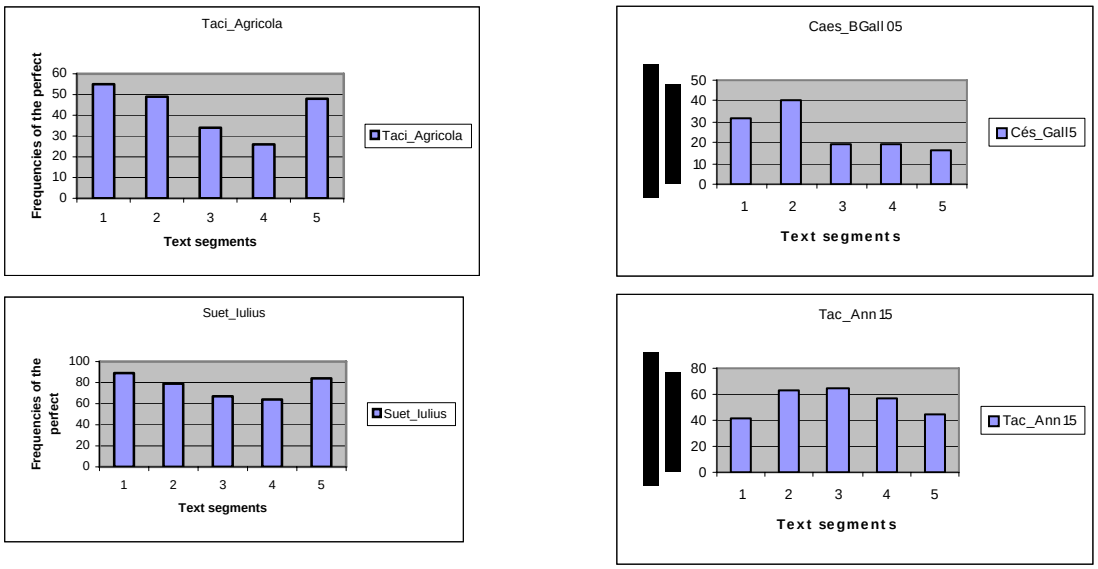


Fig. 7 Histograms for the perfect

The graphs for the two biographies of the corpus (*Tac_Agricola* and *Suet_Iulius*) show a gradual decline of the perfect in the first four parts of the text and a rise in the final part. The profile of book 15 of *Annals* is almost symmetrical, with an increase of perfects up to the middle of the text, followed by a decline in the fourth and fifth parts. The profile of Book 5 of *De Bello Gallico* is admittedly different from that of book 15 of *Annals* but it is even more different from that of the two biographies.

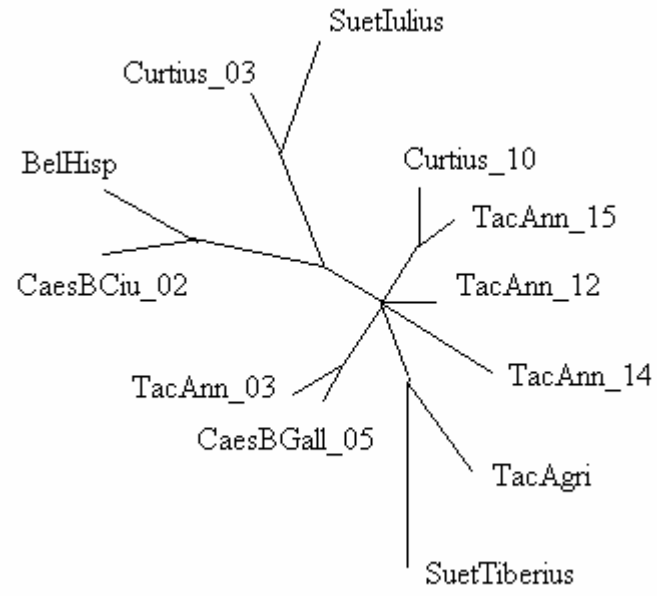


Fig. 8 Distribution of imperfects and pluperfects in main clauses
(texts divided into five segments)

Dividing the texts into five segments makes it possible to test certain linguistic parameters in terms of their relevance to the texts' overall structures. One can for instance use the same corpus and the same strings of verbal tags for a new analysis of

the same type but taking as its criteria the distribution of imperfects and pluperfects, the descriptive tenses. This distribution might be expected to be complementary to that of the perfect, the narrative tense *par excellence*. Using the same method to calculate distances between texts shows that this is not the case at all.

Not many clusters bring together works by the same author or belonging to the same genre. *Curtius_03* and *Suet_Iulius* stick together, as well as *Suet_Tiberius* and *Tac_Agricola*, but at opposite ends of the trees. *TacAnn_14* and *TacAnn_15*, the two books of Tacitus'*Annals* that were close to each other in the previous figure now hang from two relatively distant branches. The distributions of imperfects and pluperfects in the five segments of the texts then fail to account for structure and narrative organization. This state of affairs is easily accounted for: the descriptive or background elements are not concentrated solely in main clauses; they are also well represented in subordinate clauses. One should bear in mind that the distances were calculated with reference to main-clause verbs, which entails a loss of information. This loss is more important in the case of descriptive tenses than in the case of narrative tenses. Besides, it stands to reason that the profile of a story should coincide with its narrative framework and that such frameworks should be comparable in texts of the same genre, with descriptive passages more freely distributed over the corresponding structure. Other relevant linguistic features might also be investigated with the method just described.

There is of course room for improvement. One would for instance need to assess the influence of text size beyond the 1.2 variation factor enforced here. This could be achieved by adding to the corpus a number of shorter or longer texts. One would also have to question the pertinence of dividing each text into five segments. Lastly, one would have to tackle the prickly problem of representing the whole text by the sequence of its main-clause verbs. Working with additional grammatical categories, or indeed with the totality of them, might prove worthwhile and perhaps lead to revaluation of the results in hand.[14]

2.2.2. Assessing neighbourhood

The method, as we know (cf. § 1.2), consists in associating each unit of a text with a measure of its neighbourhood in terms of a property of this neighbourhood considered as relevant. As was the case for slices, the size of the neighbourhood is perfectly arbitrary. We chose to work on the sequences of tags defining the various verbal tenses in the main clauses of the texts. We opted for a neighbourhood size of 11 in which the property investigated is the presence of the indicative perfect, the basic narrative tense. For each unit of the text, a programme examines the five preceding and the five following tags counting the number of occurrences tagged for the perfect. This measure we call ***density***. The final result is a discrete topological representation of the text which is more suitable than others when it comes to accounting for its ***continuity***, since the same measure is applied to successive units. Here are the results obtained for book 2 of Caesar's *Civil War*:

[14] Longrée D., Luong X. & Mellet S., 2006.

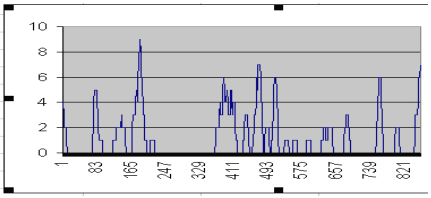
Linear chain of tense tags for all main verbs[15]:

11 12 11 11 11 11 11 11 12 12 12 12 12 12 12 12 12 12 11 11 11 11 11 15 15 12 15 15 14 11 11 15 11 11 11 11 11 11 11 15 15 15 11 12 12 15 14 12 12 12 12 12 12 14 14 14 14 11 11 14 14 12 14 14 14 14 14 14 14 14 14 14 12 12 12 12 14 14 14 14 14 14 14 14 14 14 14 15 14 14 12 12 12 12 14 14 etc.

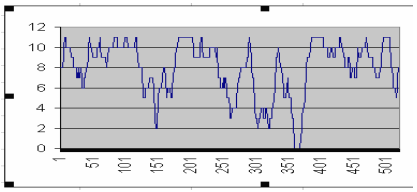
Topology based on the neighbouring measures for each unit (size of neighbourhood = 11, property of neighbourhood = number of occurrences of tag 14 coding the perfect):

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 2 3 4 5 4 4 5 6 6 7 8 8 8 8 8 9 10 10 10 10 9 8 7 7 7 7 7 7 7 8 9 10 11 11 10 10 10 9 8 7 6 6 6 6 etc.

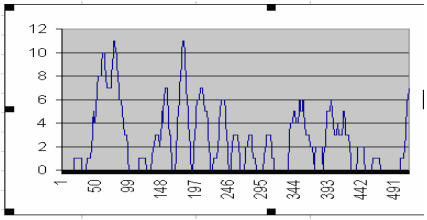
This sequence can be represented graphically. Figure 9 displays the graphs corresponding respectively to the first two books of Caesar's *Civil War* and to two books of Suetonius' *Lives of the Caesars* (Julius and Tiberius):



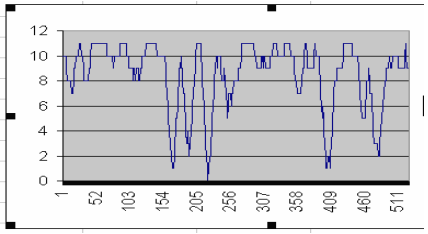
Civil War 1 (Caesar)



Life of Julius (Suetonius)



Civil War 2 (Caesar)



Life of Tiberius (Suetonius)

Fig. 9 Caesar and Suetonius

These four graphs are eloquent enough; without going into an exhaustive analysis we shall draw attention to the sharpest difference between the two authors. Caesar's two books begin in a mood which tends to exclude the perfect whereas Suetonius is not averse to using this same tense from the very first paragraphs.
These initial results then can be described as promising especially if one considers that we have sofar restricted ourselves to a rough and ready measure of neighbourhood by taking into account only one parameter. By exploiting more fully the potential of this method of neighbourhood assessment, by including more varied descriptive parameters, we can look forward to more complex and more informative findings. It will for instance be possible to complement verbal codes with parameters of sentence complexity by simultaneously taking into account the grammatical tags for subordinate clauses.

[15] Key to tags : 11=present ; 12=imperfect ; 14=perfect ; 15=pluperfect.

One problem, however, is that the exploitation of these charts is rendered difficult by the absence of a method of comparison between graphs, especially ones of different lengths. The analysis at this stage remains essentially qualitative and must be supplemented by quantitative work on equal slices of text in order to ensure a more objective exploitation of data and a more reliable calculation of distances between texts.

3. Conclusion

The power of last generation computers has made texts, culled from large language corpora syntactically tagged with coherence and delicacy, amenable to the most sophisticated mathematical models. It has also entailed a renewed challenge and incentive to linguists' creativity.

The new topological model that we have shown at work here in applications to English and Latin language corpora has yielded promising results in certain key-areas of syntactic organization and narrative structure. This analytical model, which adds a new dimension to the exclusive use of comprehensive graphic representations, has enabled us to illuminate the way in which finite and non-finite verbal forms of modern written English are combined into strings of complexity varying according to generic types and tenors and modes of communication. Applying the model to Latin texts of considerable diversity (Caesar, Curtius Quintus, Tacitus and Suetonius) has not only given fresh substance to Emile Benveniste's seminal dichotomy between discourse and history, but has also shed new light on the behaviour of a wide range of frequent verbs in literary works by authors as different from each other as Seneca, Pliny and Petronius.

It is of course necessary and desirable to move further along the lines of textual exploration initiated here and, for instance, put the model to work in subordinate clauses and in areas of the sentence beyond the verbal group. Research is in progress that will endeavour to reconcile the use of more refined analytical tools, elaborated from the same topological concepts, and the strict adherence to the fundamental quality of real language texts, the intrinsic dynamics that Ferdinand de Saussure called the linearity of language in use.

Annex 1

The mathematical formulation of linguistic concepts was part of René Thom's pioneering work in the seventies[16]. In his topological model – the so-called "catastrophe theory"[17] – he considers the possibility of giving a geometric interpretation to certain concepts that can be placed within the framework of "differentiable manifolds"[18] where they appear as "singularities". One of the a priori most interesting aspects of Thom's theory, often referred to as bordism or cobordism, is the importance assumed by border areas. It is well known that linguists, among other scholars, also have to accommodate margin, border, transition phenomena, e.g. the

---

16 Mathematician and philosopher (1924-2002), awarded the Fields medal for 1958.

17 These catastrophes, as the mathematician once humorously remarked, have never killed anybody. The object of Thom's so-called "catastrophe theory" is to study the way in which very diverse systems can vary when one parameter is altered. According to R. Thom, there are seven forms which can develop to reach a point of no-return: the fold, the dove-tail, the flounce, the butterfly, the elliptical umbilic (e.g. the point of a needle), the parabolic umbilic (e.g. the cap of a mushroom), the hyperbolic umbilic (e. g. the crest of a wave breaking). For further information see Thom R., *Stabilité structurelle et morphogénèse*, Interéditions, 1977.

18 For instance, the curves and surfaces in R^3, the three-dimensional Euclidean space, which are continuous and with continuous derivatives.

fluctuations of a given form between two or several grammatical categories. It should therefore come as no surprise that this ground-breaking mathematical theory opened promising perspectives not only to other scientists, particularly physicists and above all biologists like Zeeman[19], psycholinguists such as J. Petitot[20], but also to artists[21].
The text as a whole considered as a topological space is however situated at a much less constraint-dependent level[22]. It is a discrete structure[23] for the exploration of which there is a restricted number of mathematical tools. It is therefore necessary to look for a better-suited model, bearing in mind that any text does represent a topological space in the form of an ordered linear chain of words characterized by the neighbourhoods of each of these words and by their specific order. Whatever the number of these occurrences, there are computer procedures enabling the scholar to individualize each of them.

Annex 2: Latin texts codes, with their generic characterization (section 2.2.1)

Annals and commentaries :

Tac_Ann_03 (or Tac_annales3) : Book 3 of Tacitus *Annals*
Tac_Ann_12 (or Tac_annales12) : Book 12 of Tacitus *Annals*
Tac_Ann_14 (or Tac_annales14) : Book 14 of Tacitus *Annals*
Tac_Ann_15 (or Tac_annales15) : Book 15 of Tacitus *Annals*

CaesBGall_05 (or Bel_Gal5) : Book 5 of Caesar's *de Bello Gallico* (*Gallic War*)
CaesBCiv_02 (or Bel_civ2) : Book 2 of Caesar's *de Bello Ciuile* (*Civil War*)
Bel_Hisp : *de Bello Hispanico* (*Spanish War*) written by an anonymous officer of Caesar's

Biographies :

Tac_Agricola (or TacAgri) : Tacitus *Life of Agricola*
Suet_Iulius : Suetonius *Life of Iulius*
Suet_Tiberius : Suetonius *Life of Tiberius*
Curtius_03 : Book 3 of Curtius Rufus Quintus *Life of Alexander the Great*
Curtius_10 : Book 10 of Curtius Rufus Quintus *Life of Alexander the Great*

---

19 Zeeman C. (1977) *Catastrophe Theory: Selected papers 1972-1977*, New York, Addison-Wesley.

20 Petitot, J. (1991) "Syntaxe topologique et grammaire cognitive", *Langages*,103, 97-128.

21 In 1983 Salvador Dali painted a canvas as a tribute to the Fields medal laureate and the composer Pascal Dusapin dedicated one of his works to him in 1996.

22 Structural levels in mathematics go from the least constrained to the most constrained, each of them being associated with a specific class of functions: at the lowest level one finds set-theory structures (points independent of each other, ordinary functions), at level one topological structures (points "sticking" together through qualitative neighbourhood relations, continuous functions), at level two differentiable structures, etc...

23 A topology on a set E is said to be "discrete" if the totality of its parts P[E] can be considered as neighbourhoods for the elements of E.

References

Barthélemy, J.P. & Luong, X. (1987) "Sur la topologie d'un arbre phylogénétique: aspects théoriques, algorithmiques et applications à l'analyse de données textuelles", *Math. et Sciences Humaines* , n° 100, pp 57-80.

Benveniste, E. (1966 & 1974): *Problèmes de linguistique générale*, vols. 1 & 2, Paris, Gallimard.

Culler Jonathan (1975) *Structuralist poetics*, Routledge.

Huddleston, R. (1984): *Introduction to the grammar of English*, Cambridge, Cambridge University Press.

Huddleston, R. & Pullum, G. K. (2002): *The Cambridge Grammar of the English language*, Cambridge, Cambridge University Press.

Juillard, M. & Luong, X. (1989): "Unrooted Tree Revisited: Topology and Poetic Data", *Computers and the Humanities* 23, pp. 215-225.

Juillard, M. (1997): "Les voisinages de *sure* et *certain* dans les textes", *Numéro spécial lexicologie et linguistique, Recherches en linguistique étrangère***,** 29, hors série, Besançon, pp. 277-292.

Juillard, M. (1998): "Les lexèmes dans l'espace du texte: analyses arborées et bases de voisinage", *communication au congrès de la SAES de 1997, Cycnos 15*, Nice, numéro spécial, 1998, pp. 57-75.

Juillard, M. (2005): "Avatars de la cohésion dans les corpus", in A. Jaubert (ed.), *Cohésion et cohérence, études de linguistique textuelle*, Lyon, ENS Editions (coll. Langages), pp. 175-194.

Labbé, C & Labbé, D (2005) « A Tool for Literary Studies : Intertextual Distance and Tree Classification », *Literacy and Linguistic Computing*

Lafon, P. (1981): " Statistique des localisations des formes d'un texte", *Mots* 2, pp. 157-187.

Lamalle, C. & Salem, A. (2002): "Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels" in A. Morin & P. Sébillot (eds.), *JADT 2002, 6èmes Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, Irisa et Inria, vol. 2, pp. 529-538.

Longrée, D. (2003): "Temps verbaux et spécificités stylistiques chez les historiens latins", in G. Calboli *et al*. (eds.), *Papers in Grammar* IX, 1, Rome: Herder editrice, pp. 863-875.

Longrée, D. & Luong, X. (2003): "Temps verbaux et linéarité du texte: recherches sur les distances dans un corpus de textes latins lemmatisés", *Corpus 2*, pp. 119-140.

Longrée, D., Luong, X. & Mellet, S. (2004): "Temps verbaux, axe syntagmatique, topologie textuelle: analyse d'un corpus lemmatisé", in G. Purnelle, C. Fairon & A. Dister (eds.), *Le poids des mots*, Louvain: Presses universitaires de Louvain, vol. 2, pp. 743-752.

Longrée, D. & Mellet, S., (2006): "Temps verbaux et prose historique latine: à la recherche de nouvelles méthodes d'analyse statistique", *in* J. Denooz & G. Purnelle, *Ordre des mots et cohérence en latin*, Liège: Presses université de Liège

Longrée, D., Luong, X. & Mellet, S. (2006): "Le classement des textes d'après leur structure: méthodes de découpage et analyses arborées", in *Actes des JADT06*, Besançon, Avril 2006.

Luong, X. & Mellet, S. (1995): "Les calculs multidimensionnels au service de l'analyse syntaxique diachronique" in S. Bolasco, L. Lebart & A. Salem (eds.) *Analisi statistica dei dati testuali*, Rome: CISU, vol. II, pp. 281-288.

Luong, X. & Mellet, S. (2003): "Mesures de distance grammaticale entre les textes", *Corpus* 2, pp. 141-166.

Petitot, J. (1991): "Syntaxe, topologie et grammaire cognitive", *Langages*, 103, pp. 97-128.

Quirk, R. *et al*. (1985): *A Comprehensive grammar of the English language*, London, Longman.

Salem, A. (2002): "Topographie textuelle dans l'analyse quantitative des textes", *Actas del segundo Seminario de la Escuela interlatina de altos Estudios en Lingüística aplicada, Matemáticas y Tratamiento de Corpus (s.n.),* San Millán de la Cogolla, Logroño, pp. 53-59.

Salem, A. (2004): "Introduction à la résonance textuelle"" in G. Purnelle, C. Fairon & A. Dister (eds.) *Le poids des mots*, Louvain, Presses universitaires de Louvain, vol. 2, pp. 986-992.

Saussure, F. de (first edition 1916) *Cours de linguistique générale*, Paris, Payot.

Sérant, D. & Thoiron, Ph. (1988): "Topographie des formes répétées", *Revue Informatique et Statistique dans les Sciences humaines* 24 , pp. 333-343 (*Le nombre et le texte: Hommage à Étienne Évrard*).

Thom, R. (1977): *Stabilité structurelle et morphogénèse*, Paris, Benjamin.

Thom, R.(1989): *Entretiens avec des mathématiciens*, J. Nimier, éditions de l'Institut de recherche sur l'enseignement des mathématiques de Lyon.

Zeeman, C. (1977) *Catastrophe Theory: Selected papers 1972-1977*, New York, Addison-Wesley.

Westin, I. (2002), *Language change in English newspaper editorials*, Language and Computers – Studies in practical linguistics, 44, Rodopi, Amsterdam/New York.

Wittgenstein, L, (1953), *Philosophische Untersuchungen*, Oxford, Blackwell.