The contribution of Latin to French-language quantitative linguistics: from lemmatisation to grammaticometry and textual topology

Dominique Longrée, Sylvie Mellet

The contribution of Latin to French-language quantitative linguistics: from lemmatisation to grammaticometry and textual topology

LONGREE Dominique, MELLET Sylvie
* Univ. de Liège, LASLA, Belgique
** Univ. Côte d'Azur, CNRS, BCL, France

0. Introduction

In this overview of quantitative linguistics in France, we focus on works involving Latin corpora. Our contribution points out that statistical handling of digitized Latin texts is an original and important addition to quantitative linguistics studies, and we investigate the epistemological foundations of this addition. To this end, we go beyond the boundaries of France and look to Belgium, because the development of quantitative studies devoted to Latin texts is a Franco-Belgian achievement, and is based almost entirely on resources produced, beginning in 1961, by the Laboratory for the Statistical Analysis of Ancient Languages (LASLA) at the University of Liège[1]. We first emphasize the role of lemmatisation, and show how this simple operation of abstraction and regrouping allows other more or less complex analysis units to emerge. We then discuss the importance that variability of word order in Latin has assumed with regard to research issues and approaches; finally we discuss software advances and certain necessary adaptations involving digital research methods and quantitative handling made necessary by specific approaches to Latin corpora.

1. From lexicometry to grammaticometry

Preparing Latin corpora for textual data analysis or quantitative linguistics is a particular operation, since the researcher is immediately confronted with the problem of lemmatisation. In the first place, Latin is an inflectional language for which a lexicometry based on graphic forms is problematic. This does not mean that the question of lemmatisation has not been the object of a lively debate about methods for analyzing French or other living languages that are not inflectional (or not as inflectional) [2]. But in Latin, the other alternative – the one that consists

[1] http://www.cipl.ulg.ac.be/Lasla/

[2] We recall especially the revealing title of an article by Etienne Brunet: "Qui dit lemme, dilemme attise". The explicit rejection of lemmatisation by M. Tournier ("La

in focusing on graphic forms – appears at first glance to be more limiting and restrictive, even paralyzing. The workaround consisting in using chains of characters in order to collect all the forms of one lexeme is quite ineffective, since inflection has a much greater impact on the variability of forms: as regards verb conjugation, nominal inflection includes 6 cases and 3 gender types. Most importantly, inflection can have a considerable effect on the forms of radicals. [3]

The automatic production of indexes – the objective of the Latin lexicometry pioneers – was directly in line with the philological tradition[4] and presupposed a particular form of the organization of data. Entries in the index were lemmas, arranged alphabetically, and under each lemma, forms were arranged in a fixed morphological order. Here is an example with the lemma DICO2 (for the verb *dico* / *dicere* of the third conjugation, differentiated from the verb *dico* / *dicare* of the first conjugation by the index 2). This example says that there are 183 occurrences of this lemma in the text among which one form *dico* (first singular person) in the 8th place of the 19th sentence of the 12th chapter of the 5th book of the work; and 3 occurrences of the form *dicis* (second singular person) with their references according to the same reference system as previously:

183 DICO2
dico
5, 12, 19, 8
dicis
3, 12, 19, 8
3, 15, 13, 6
3, 21, 2, 8
…

The efforts at lemmatisation made indispensable by such a conception of indexes were therefore ahead of their time, and were inevitably accompanied by morphosyntactic analysis, the results of which could usefully, and with little additional work, be recorded in order to be directed toward other purposes. It should be noted that the granularity of grammatical labelling used by LASLA is very fine indeed. This is another direct result of the principles that governed the compilation of indexes; since it is only from a precise and complete description

---

lemmatisation ne résout rien et empire tout") dates from 1985, but the effort toward lemmatisation of Latin texts by LASLA goes back to the early 1960s.

[3] Cf. Mellet 1996; Purnelle, 1996; Mellet 2002a; Mellet, Sylvie & Purnelle, Gérald, 2002.

[4] You can see a list of LASLA publications at: http://www.cipl.ulg.ac.be/Lasla/publications.html. But see also the early publications by Etienne Brunet and the entire collection, "Travaux de Linguistique Quantitative" from Slatkine, which collects mainly vocabularies and indexes.

of the form that we can determine, automatically, its position under the lemma that is also its entry in the index.

Since this morphological information was already in computer memory and easily available, why not use it for other purposes? The first case of this was pedagogical. But researchers also found very interesting new units of analysis in this information: why not study their frequency and distribution? Thus it was that, beginning in the mid-1960s, Étienne Évrard, one of the founder of LASLA, began to ask questions about the stability of grammatical categories in Latin texts[5] and about the possibility (or lack thereof) of transposing, in respect of these categories, general laws discovered regarding the distribution of vocabulary. And about ten years later, studies based on grammatical categories began to appear, whether it was a matter of characterizing the writings of an author or a particular work,[6] or of studying the use, distribution and characteristics of a grammatical construction, or more broadly a complete sub-system such as the system of subordinates in Latin.[7]

Otherwise, owing to the inflectional character of the language, word order is less dispositive in Latin as regards the identification of the syntactic functions of various syntagms that make up a sentence. It appears to be more "flexible" than it is in languages such as French or English, but it is no less significant at other linguistic levels (semantic, pragmatic, stylistic, etc.). Thus Latinists early on took an interest in this question, and proceeded to work up counts of various configurations, although in the beginning these remained intuitive and only approximate. In the pioneer work of J. Marouzeau, *L'ordre des mots dans la phrase latine*,[8] there are many vague expressions of this sort: "in many cases", "quite a few examples", "most examples", etc. The first systematic counts appeared with the thesis of F. Charpin, published in 1977.[9] They had as much to do with the order of syntactical constituents as with sequences of identical endings or with chains of pre-accentual sequences. The utility of such counts became apparent to Latinists, and beginning in 1978, J. Perrot[10] emphasized the necessity of statistical enquiries concerning the "norms" for the arrangement of "meaningful material", "comparable to those that have been produced for phonic

[5] Évrard 1966. It is true that in the year in which this communication was delivered (conference in 1964), Robert Martin and Charles Muller published "Syntaxe et analyse statistique. La concurrence entre le passé antérieur et le plus-que-parfait dans *La Mort le Roi Artu*"; but most of the counts were obtained manually.

[6] See Fleury 1978; based on observation of a positive specific deviation for the verb *dico* "to say" in the *Satires* of Persius, this paper looked at tenses, modes and persons in terms of which the verb was conjugated, and at its most frequent constructions. See also Delatte 1979, which observes the use of grammatical categories in Ovid's *Héroïdes* and introduces in this context the notion of binary chains of two labels – a notion which we refer to below.

[7] See Delatte, Govaerts & Denooz 1978.

[8] Marouzeau, 4 vol. 1922-1953.

[9] Charpin 1977.

[10] Perrot 1978.

material": these enquiries were made really practical only through the existence of computerized data bases.[11]

An equally important contribution made in studies on word order in Latin can be found in research on recurrent "formulas". Recognizing such formulas is of particular interest in the linguistic or stylistic characterization of literary works. In an article written in 1989, G. Purnelle[12] presented research concerning "verbal groups [that are] syntactically homogeneous and repeated, whose constituent elements are contiguous or nearly so in the text". This study did indeed take into account the work of A. Salem and the Saint-Cloud Laboratory concerning repeated segments, but differed in two respects. First, G. Purnelle distinguishes "recurrent verbal groups" whose identification does not at all depend on literary, stylistic or semantic considerations, from "formulas" that corresponded to "a still more homogeneous group which makes up a true fixed expression, if not in terms of the entire language, then at least in terms of the author language or of the genre of the work itself, and which functions as a single semantic entity". In addition, in Latin, a formula can have not only inflectional variations, but also inversions of the order of its constituents, or insertions of terms that can, naturally, be expansions of the formula, but which can also have nothing to do with it. Under these conditions the notion of "repeated segments" can serve as a model, but the analysis must eventually go beyond it. G. Purnelle thus offers a method that is based on the annotations contained in LASLA files, which aims at taking into account variations in morphology and word order; Purnelle suggests the possibility of developments that would take into account the distance in the text separating each occurrence of a given formula, in order to "distinguish actual formulas from what is only a simple repetition of an expression recently employed in the text, which is borne in mind by both the author and the reader". Thus the way was opened for studies of textual dynamics. The programme of research proposed in the article was not immediately carried out by the author, but would be carried forward by others a few years later.

2. Evolution and adaptation of tools and methods

2.1 Software tools

In order to reach these objectives which had been quite specific for them for a long time (constitution of lemmatized alphabetical indexes, study of morpho-syntactic categories, research on word order), Latinists had to create or adapt tools and methods that would later benefit the larger community of researchers in textual data analysis.

Well before the appearance of the first automatic taggers, LASLA designed a semi-automatic lemmatiser that took apart each textual form, comparing it with a lexicon of radicals and affixes, and then provided the philologist with a list of all possible analyses (assignment to a lemma and complete morpho-

[11] Charpin 1989a and 1989b in which the author completes the counts presented in his thesis.

[12] Purnelle 1989.

syntactical description). The philologist then had to choose the correct analysis. This produced files in which each textual form was associated with several different kinds of information:

1. its lemma, such as it appears in the reference dictionary;
2. an index allowing people to distinguish between different homograph lemmas, or to mark proper nouns and the adjectives derived from them;
3. the precise reference, and accordingly the position of the form in the text;
4. a complete morphological analysis in an alphanumeric format;
5. for verbs, syntactical information allowing researchers to distinguish between the predicates of a main clause or of a subordinate clause, and in the latter case, to connect the predicate to its subordinating word.

The first files of LASLA were thus presented as follows:

Lemma	**Index**	**Form**	**Reference**	**Morphological Tag**
LIBERTAS		LIBERTATEM	41 001 0001 007 007	13C00
ET	2	ET	41 001 0001 008 008	81000
CONSVLATVS		CONSULATUM	41 001 0001 009 009	14C00
LVCIVS	N	L.	41 001 0001 010 010	12A00
BRVTVS	N	BRUTUS	41 001 0001 011 011	12A00
INSTITVO		INSTITUIT	41 001 0001 012 012	53C14

In this table is pictured an excerpt of the LASLA file corresponding to the following sentence: *Libertatem et consulatum L. Brutus instituit*. To each form of the sentence (third column) is first associated the lemma, with an index which removes a possible ambiguity (for example ET2 = coordinating conjunction "and" while ET1 = adverb "too"). The 4th column gives the reference of this form in the book: here *libertatem* appears in the chapter 41, in the first paragraph of the chapter, in the first sentence of the paragraph; it is the 7th word of the paragraph and the 7th word of the sentence. *Et* is the 8th word of the same sentence and the same paragraph, *consulatum* is the 9th one and so on. Finally, the last column gives an alphanumeric tag: for *libertatem* 13C00 means substantive of the third declination, singular accusative; for *instituit* 53C14 means verb of the third conjugation, third singular person, perfect indicative.

In some cases, the number of data points associated with a single form could go as high as ten. Thus, for a participle such as *regnante*, the following data are given: reference, lemma, part of speech, conjugation type, voice, case, number,

mode, tense and gender. Finally, one would eventually be able to determine if the form is the predicate of an ablative absolute (participial proposition).

With the development of personal computers in the early 1990s, the utility of creating software tools that could manipulate all this information became apparent. In the beginning, it was a matter of concordance programmes that could produce not only alphabetical lists of all instances of forms, but also all the forms of a lemma or all the forms associated with one or more given grammatical categories, or with a particular syntactical annotation. In a short time, software developed in order to manipulate the data in the files of LASLA, Estela and Opera Latina first, and Hyperbase-Latin following, allowed the creation of concordances on the basis of a complex search combining the research of a lemma associated with one or more grammatical or syntactic categories. For example, the concordance-maker can provide a contextualized list of the occurrences of the verb *uincere* "to win" only in passive forms in a relative subordinated clause.

Thanks to the way the data are prepared and structured, and to the resulting enrichment of texts, these concordance makers were able to take into account the multidimensionality of textual data, well before S. Fleury and A. Salem perfected their concept of a "Trameur".

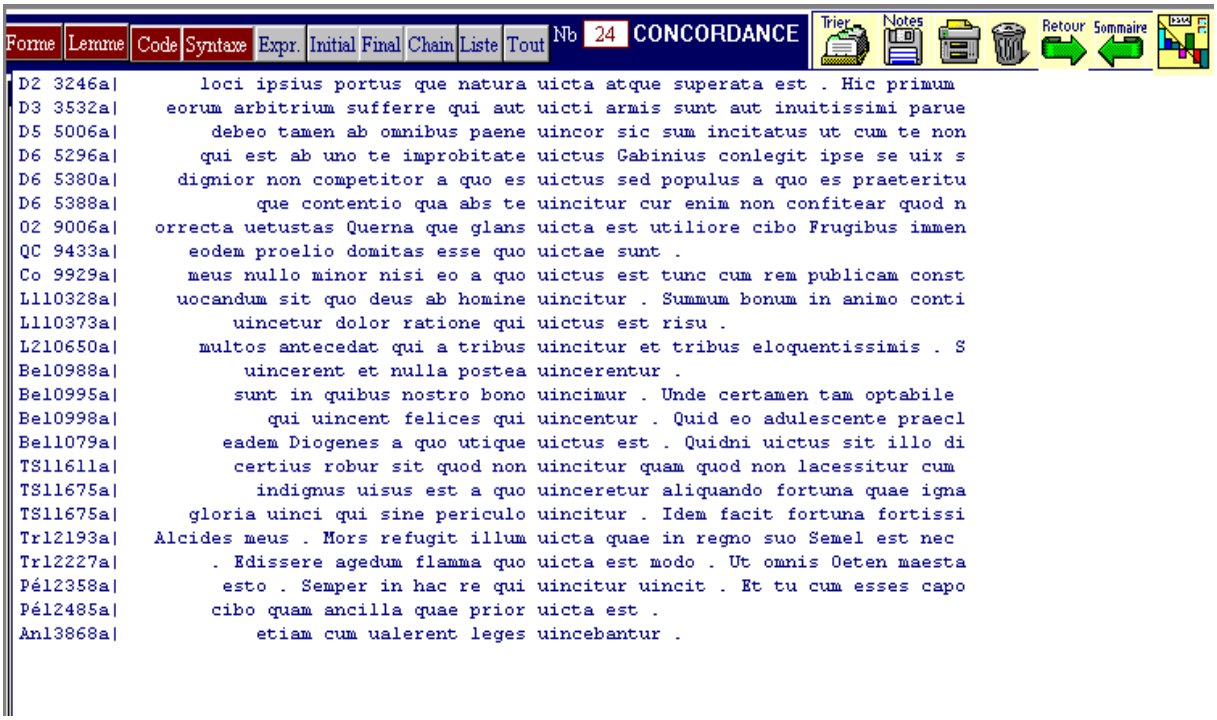

Figure 1. Concordance of passive forms in the lemma *uincere* "to win" in a relative subordinated clause, in the whole Latin corpus of LASLA

An illustration of this premature concern for multidimensionality can be found in the simultaneous display, by Hyperbase, of a single textual sequence of both forms and lemmas corresponding to them, or of both forms and morpho-syntactic codes associated with them. Such an illustration can be found as well in

the simultaneous display, in the dictionary, of forms, lemmas, and morpho-syntactic codes.

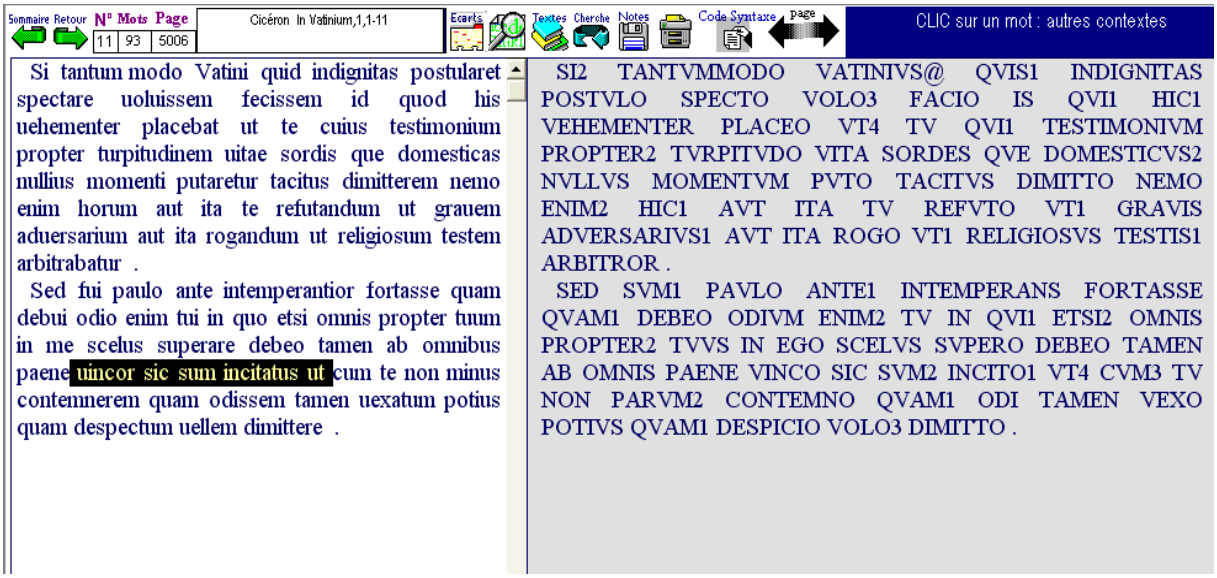

Figure 2:.Parallel reading of a text excerpt, set up as a string of forms and a string of lemmas

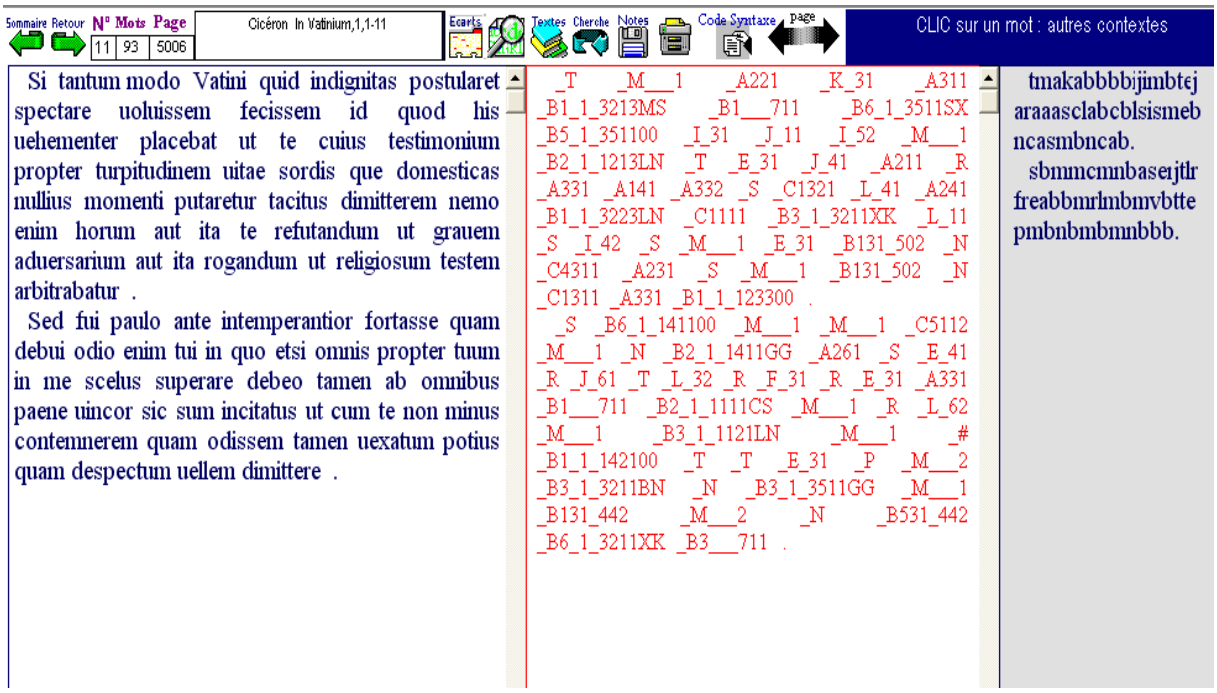

Figure 3. Parallel reading of a text excerpt set up as a string of forms and as a string of morphosyntactic codes

Latinists' interest in this multidimensional approach to Latin texts explains, for the most part, why Hyperbase, which was adapted to Latin at a relatively early date, later became one of the first software programmes that si-

multaneously took into account lemmas and grammatical categories for corpora of French, English or Portuguese texts: it was only necessary to wait until trainable automatic taggers were able to furnish dependable morphosyntactical information for these languages.

Apart from concordance makers, software for manipulating Latin textual data rapidly came to include functionality based on statistical calculations such as the calculation of chi-square, reduced variations, specificities. These functions have proven particularly useful in order to characterize the texts in terms of their use of grammatical categories, and also in order to gain better understanding of the function of these categories, which from that point on could be grasped in terms of the specificity of their distribution in context.[13] This was one of the objectives of S. Mellet in her thesis devoted to the imperfect indicative in Latin.[14] Other examples can be found in the article by C. Bertrand on verbal forms and the structure of phrases in the *Historia Augusta*[15] or in the article by D. Renard on the parts of discourse used by various characters in the *Satyricon* of Petronius.[16]

2.2 Statistical methods

In a first approach, the statistical treatment of grammatical categories seems to be able to use the methods and calculations that are applied to lexicons (forms or lemmas). As we have just noted, one can easily and accurately calculate the grammatical specifications of a text, or of any other part of a corpus. One can go further still in taking account of the effect of grammatical functions in the calculation of the specific co-occurrents of a keyword.[17] Finally, one may integrate grammatical categories into all methods of multidimensional calculation, and on this grammatical basis handle data matrices in order to extract from them a graphic representation that is then submitted for interpretation to a linguist. Factorial analyses of correspondences or tree analyses made on the basis of the distribution of grammatical categories in different areas of a corpus are quite expressive and often succeed in corroborating certain classifications (according to authors, to genres, to a chronology, etc.) that in turn appear in the results of lexicometric treatments. But they bring to this a healthy independence, in relation to the thematics of works. They also offer complementary elements of analysis that make more subtle classifications possible, as we were able to demonstrate as early as 1987, and several times thereafter.[18] Another pioneer in this area has been D. Biber,[19] who experimented in English with "grammaticometric" tech-

[13] See Evrard & Mellet 1998.
[14] Mellet 1987.
[15] Bertrand 1982.
[16] Renard 2000.
[17] Longrée & Mellet, 2012
[18] Mellet, 1987; Mellet 1998; Mellet 2002b; Longrée 2004; Longrée 2005.
[19] Biber, 1988.

niques, and was able to demonstrate their interest for linguistics. These methods were applied to different kinds of corpora in different languages.[20] However Latinists have retained the distinction of working with very fine grammatical categories thanks to the initial material they had to work on. The contribution of the quantitative analysis of grammatical categories is particularly valuable when a corpus groups together works that share a general theme, a vocabulary and certain conventional motifs: in fact the use of grammatical categories is more likely to be independent from this thematic and semantic framework, and to escape the control of the writer: therefore they give access to deep and intrinsic characteristics of the author's style. This is one of the benefits used in the thesis of Caroline Philippart de Foy, devoted to an *Étude d'un corpus de traductions médiolatines d'origine grecque*,[21] the analyses of which have allowed us to globally characterize different groups of translations, initially defined on the basis of historical and philological sources, to confirm the pertinence of this classification without masking the heterogeneous aspects of each group, and to suggest some definite attributions to particular authors, or at least to suggest a school of translation, in the case of orphaned works. Used for purposes having to do with the characterization of works and not just for classification, these multidimensional methods provide solid complementary information, amounting to significant added value when complementing lexical analysis.

Just the same, "grammaticometricians" are quickly led to question the pertinence of applying the statistical tools of classical statistical linguistics to grammatical categories. Grammatical categories have distributional specificities that force us to rethink using the statistical methods used in classical lexicometry. On one hand it is rare for a major grammatical category to be completely absent in a text. We cannot calculate intertextual distances according to models that are set up in terms of the presence or absence of a variable in different texts that are compared. It is necessary to work with frequencies, and to develop new algorithms for this purpose. Also, the enumeration of this new type of variables produces matrices that sometimes have columns that contain very few data points, but which nonetheless contain important information in the eyes of the philologist or the expert in stylistics (for example, the use of the infinitive of narration used by historians). It is thus necessary to recover this information in the calculation of distance, even though the number of instances is much too low to allow for its being analyzed in terms of classical statistical analysis. One of the adaptations suggested by S. Mellet and X. Luong[22] was to make the computation depend upon numerical values not corresponding to the number of the instances of each category in each text of a corpus, but to a numerical ordering according to this number of instances: the matrix of initial data is converted so that it assigns to each text a number in an order that represents its rank with respect to the use it makes of various grammatical categories being examined. This classific-

[20] See for example Kastberg 2006; Loiseau, Poudat & Ablali 2006.

[21] Philippart de Foy 2008.

[22] Luong & Mellet 2003.

ation produces first a pre-order when it gives rise to cases of equal standing; this pre-order can be transformed into a classification of "middle ranks". In every case, we are working with homogeneous data distributed over a reduced scale, which can be submitted to a simple Euclidean calculation of distance (all forms of weighting are useless here). Results obtained in terms of works and classifications are very satisfying, in that they reveal groupings that are coherent but not completely obvious in terms of philological knowledge. It should be noted that this method does not appear to have been used by others since it was published.

Thus we see that lemmatisation and tagging of Latin texts allows us to escape from a illusory naturalness of data, and to create new analysis through a double process of abstraction and construction of the object studied. This approach reaches its highest point thanks to the conceptualization of a complex object – the motif – in the new epistemological framework of textual topology.

2.3 Textual topology

In this process of construction of the object of study, it appears that just counting the appearances of a form, a lemma or a code, taken in isolation, is not enough to give an account of the specific textual dynamics of a work. The recurrent succession of certain sequences of items and the configuration of morpho-syntactical sequences belonging to certain works appear as particularly pertinent elements of analysis. Thus D. Longrée and X. Luong in 2003 published an initial article on sequences of verb tenses: identified after a reduction of the text to a chain of the morphological codes of predicates of main clauses, these sequences have been chosen as a parameter for the characterization of Latin historians' writings.[23] This first attempt at integrating the ordered linearity of a text through a quantitative treatment foreshadowed the later development of methods, which under the aegis of the famous "beyond the bag of words" (2005-2006),[24] were no longer content to apply to texts the traditional statistical method of the Polya urn scheme. In fact, when we work on an author's style or on the structure of a work, we quickly see the necessity of taking account – even in the context of a quantitative treatment – of the organization of a syntagmatic axis grasped at one and the same time in terms of short range (in repetitive sequences of a single form or of a single grammatical structure, and in the breaks in these sequences) and of long range (in the distribution of the studied units across the different parts of a text).[25] Such a manner of apprehending textual structure led J.P. Barthelémy, D. Longrée, X. Luong to S. Mellet to explore the possibility of a topological

[23] Longrée & Luong 2003, and also Longrée & Luong 2005; Longrée & Mellet 2007.

[24] From the name of the Workshop of the 28th Annual International Conférence of ACM SIGIR.

[25] It is interesting to note that this context is also one in which methodological work was developed on co-occurrences, whose methodological point of view is not unrelated to our purpose. Cf. Mayaffre 2008a and 2008b.

modelling of texts.[26] Over short ranges, the aptitude of a grammatical category to be systematically associated with other categories from a syntagmatic point of view or to favour certain collocations can be apprehended through studies of *voisinages* (neighbourhoods).[27] Over long ranges, the distribution of a sequence according to parts of the text (introduction, narration, commentary, conclusion, etc.) can be analyzed through a method of cutting the texts into fixed or variable sections,[28] and its rhythm of appearance can be analyzed through the method of the calculation of "bursts" ("rafales").[29]

Taken together, these approaches allow us to get beyond the stage at which texts are considered as simple ensemble-type structures[30] and to take into account the form of the text as a whole and in terms of its parts. The notion of topological space applied to texts has been theorized: the text becomes an ensemble of points, each of which has a family of neighbourhoods, and can thus be studied through the concepts and tools we borrow from mathematical topology – more precisely, discrete topology.

As we have deepened our study of neighbourhood structures, we have become aware that some of these have properties which make them textual objects that are particularly worthy of study: they are multidimensional (they associate lexical and grammatical constraints), ordered and recurrent, and they possess a textual function (structuring or characterizing). We have given the name of "***motifs***" to these structures, which constitute elements of the modelling of a text as a topological space [31] and that formalize in a more systematic manner the properties that Gérald Purnelle had attributed to "formulas". The property of recurrence makes them good candidates for treatment by means of textometric tools. Motifs allow us to automatically characterize either the various parts of a text, or the different texts of a corpus.[32] Otherwise, thanks to the articulation between its basic schematic form and its textual functionality (which contributes in an important way to the stability of its recognition), a "motif" can feature variants (permutation of two elements; commutation within a paradigmatic series; insertion, expansion or erasure; inflectional variation). Thus the study of motifs returns us to the problem of word order, which is a guiding thread for quantitative linguistics in Latin. Subsuming the notions of repeated segments, collocations and colligations, it permits an enlargement of the domain of phraseology and constitutes a contribution – relatively unexpected – from Latin to the disciplinary field that is generally devoted to the terminology of living lan-

[26] Mellet & Barthélemy 2007; Barthélemy, Longrée, Luong & Mellet 2009.

[27] In the mathematical sense of the word; see Longrée, Luong & Mellet 2004

[28] Longrée, Luong & Mellet 2004 and 2006

[29] Lafon 1981. For an application, see for example Lenoble 2006, especially pp. 479-493.

[30] Longrée, Luong, Juillard & Mellet 2007.

[31] Longrée, Mellet & Luong 2008; Mellet & Longrée 2009; Mellet & Longrée 2012.

[32] Gohy & Martin Leon 2012; Magri & Purnelle 2012.

guages.[33] Another interesting relationship involving motifs takes place in connection with another domain of Natural Language Processsing (NLP), the domain of *data mining*, as soon as this begins to integrate sequential constraints into its methodology.[34] Finally, the notion of motif also allows openings toward psychology, inasmuch as psychologists might use it as a tool for analyzing the verbal production of subjects under examination, or insofar as it might function as a particularly complete representation of lexical associations whose cognitive functioning is thus modelled. In this area, work is underway.

Naturally, the constitution of a new unit of analysis, on the one hand, the taking into account of the topological dimension of texts, on the other hand, have led to new software developments, most often in collaboration with Latinists. The functionality of Hyperbase-Latin, and also that of Hyperbase-français have been considerably enriched in recent years. TXM has developed, especially under the influence of the reflection engaged in by B. Pincemin on the necessary modelling of texts.[35] And the designers of textometric software have discussed and collaborated with specialists in NLP in order to develop non-supervised research tools for motifs, for example, the online program for the extraction of sequential motifs, that is, SDMC, "Sequential Data Mining under Constraints". [36]

This state of the art and this assessment of the research progress show that the contribution of classical languages to quantitative linguistics is based first on the relative anteriority of computerized and tagged corpora for these languages, and consequently on the lines of questioning they prematurely supported. This longevity of a line of research in textual data analysis has been accompanied by a strong methodological consideration, linked to the specific quality of Latin data. The pathways opened up have not always been followed up or investigated by others, but in a certain number of cases, convergences have given rise to particularly fruitful collaborations, especially in order to comprehend the textual structure from a global point of view and to develop the software necessary for following up this global approach.

References

Barthélemy, Jean-Pierre ; Longrée, Dominique ; Luong Xuan ; Mellet Sylvie (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* (47ème année) *187, 3 : 107-121*.

[33] Longrée & Mellet 2013. This study contributes, once again, to the research in this area the strong multi-dimensionality and precise labelling of numerical Latin data, and also enhances the approaches developed, for example, in Biber 2009 or Grezka & Poudat 2012.

[34] Quiniou, Cellier, Charmois & Le Gallois 2012.

[35] Pincemin 2008; Pincemin, Heiden, Lay, Leblanc & Viprey 2010; Heiden, Magué & Pincemin 2010.

[36] Béchet, Cellier, Charnois, Crémilleux & Quiniou 2013.

Béchet, Nicolas; Cellier, Peggy; Charnois, Thierry; Crémilleux, Bruno; Quiniou, Solen (2013). SDMC: un outil en ligne d'extraction de motifs séquentiels pour la fouille de textes. In : *Actes de la Conférence Francophone sur l'Extraction et la Gestion des Connaissances (EGC'13)*, Toulouse 2013 [see HAL web-site: http://hal.archives-ouvertes.fr/hal-00817074].

Bertrand, Cécile (1982), « L'Histoire Auguste : formes verbales et structure des phrases dans la *Vita Hadriani* et la *Vita Heliogabali* », *RELO, 18, 59-79.* [http://promethee.philo.ulg.ac.be/RISSHpdf/annee1982/CBertrand.pdf]

Biber, Douglas (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

Brunet, Étienne (2000). Qui lemmatise, dilemme attise. *Lexicometrica* 2. [see *Lexicometrica* web-site: http://lexicometrica.univ-paris3.fr/article/numero2/brunet2000.PDF].

Charpin, François (1977). *L'idée de phrase grammaticale et son expression en latin*. Lille – Paris: H. Champion.

Charpin, François (1989a). Étude de syntaxe énonciative: l'ordre des mots et la phrase. In: G. Calboli (ed.), *Subordination and other topics in Latin, Proceedings of the third Colloquium on Latin linguistics, Bologna, 1-5 April 1985, 503-520* (Studies in Language Companion Series, 17), Amsterdam – Philadelphia: John Benjamins.

Charpin, François (1989b). Les finales homonymes dans le discours latin. *Revue, Informatique et Statistique dans les Sciences humaines, 25, 65-108*. [http://promethee.philo.ulg.ac.be/RISSHpdf/Annee1989/Articles/FCharpin.pdf].

Delatte, Louis (1979), « Recherches statistiques sur les *Héroïdes* XVI et XVII d'Ovide », *RELO, 14 (2), 1-61.* [http://promethee.philo.ulg.ac.be/RISSHpdf/annee1979/02/LDelatte.pdf/].

Delatte, Louis; Govaerts, Suzanne; Denooz, Joseph (1978). *L'ordinateur et le latin. Techniques et méthodes, morphologie, syntaxe, lexicologie, stylistique*, Liège, LASLA [(http://promethee.philo.ulg.ac.be/LASLApdf/Lordinateuretlelatin.pdf].

Évrard, Étienne (1966). La fréquence des phénomènes grammaticaux est-elle constante? In: *Actes du premier colloque international de linguistique appliquée (Nancy, 26-31 octobre 1964).* Nancy: PUN (« Annales de l'Est »), *157-162*.

Évrard, Étienne; Mellet, Sylvie (1998). Les méthodes quantitatives en langues anciennes. *LALIES 18 : 111-155*.

Fleury, Philippe (1978). « Essai d'exploitation de données fournies par des moyens informatiques sur les *Satires* de Perse », *RELO, 14 (3), 45-70.* [http://promethee.philo.ulg.ac.be/RISSHpdf/annee1978/03/PFleury.pdf]

Gohy, Stéphanie; Martin, Leon Benjamin (2012). Détection automatique des textes épistolaires du corpus néo-égyptien: méthodes exploitant la récurrence de motifs discriminants. In: Anne Dister, Dominique Longrée, Gérald

Purnelle, *JADT 2012, Actes des 11e Journées internationales d'analyse statistique des données textuelles*, Liège, *487-500*. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm)

Grezka, Aude; Poudat, Céline (2012). Building a database of French frozen adverbial phrases », *in Proceedings of LREC 2012*, *685-692*. [see LREC web-site: http://www.lrec-conf.org/proceedings/lrec2012/pdf/1020_ Paper.pdf].

Heiden, Serge; Magué, Jean-Philippe; Pincemin, Bénédicte (2010) TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement. In: Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds.), *JADT 2010, Statistical Analysis of Textual Data - Proceedings of 10th International Conference*. Rome Edizioni Universitarie di Lettere Economia Diritto.
[see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2010/ allegati/ JADT-2010-1021-1032_025-Heiden.pdf].

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

Lafon, Pierre (1981). Statistiques des localisations des formes d'un texte. *Mots 2, 157-187.*

Lenoble, Muriel (2006). *Le passif impersonnel du type uenitur chez les historiens latins (César, Salluste et Tacite). Essai méthodologique, quantitatif et descriptif*. Unpublished dissertation, Facultés Saint-Louis, Bruxelles.

Loiseau, Sylvain; Poudat, Céline; Ablali, Driss (2006). Exploration contrastive de trois corpus de sciences humaines. In: Jean-Marie Viprey (ed.), *JADT 2006, 8èmes Journées internationales d'Analyse statistique des Données Textuelles*, Besançon. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-056.pdf]

Longrée, Dominique (2004). Une approche statistique de la concurrence entre démonstratifs chez les historiens latins (César, Salluste, Tacite). In: C. Bodelot (éd.), *Anaphore, cataphore et corrélation en latin.* Clermont: Presses Universitaires Blaise Pascal, (Collection « Erga », Recherches sur l'Antiquité, 6), *157-178*.

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

Longrée, Dominique; Luong Xuan (2003). Temps verbaux et linéarité du texte: recherches sur les distances dans un corpus de textes latins lemmatisés. *Corpus* 2 (« La distance intertextuelle »), *119-140*. [see : Revues.org: http://corpus.revues.org/33].

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

Longrée, Dominique ; Luong Xuan; Juillard, Michel; Mellet, Sylvie, (2007). The concept of Text Topology. Some applications to Verb-Form Distributions in Language Corpora. *Literary and Linguistic Computing 22(2), 167-186.*

Longrée, Dominique; Luong Xuan ; Mellet, Sylvie (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.* Louvain-la-Neuve, *743-752*. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_071.pdf].

Longrée, Dominique; Luong Xuan; Mellet, Sylvie (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, 8èmes Journées internationales d'Analyse statistique des Données Textuelles: 643-654*. Besançon : Presses universitaires de Franche-Comté [see *Lexicometrica*, web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2006/PDF/II-057.pdf].

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

Longrée, Dominique; Mellet, Sylvie (2012). Asymétrie de la cooccurrence et contextualisation. Le rôle de la flexion casuelle dans la structuration des réseaux cooccurrentiels d'un mot-pôle en latin. *Corpus 11, 91-128.* [see Revues.org : http://corpus.revues.org/2230].

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

Longrée, Dominique; Mellet, Sylvie; Luong Xuan (2008). Les motifs: un outil pour la caractérisation topologique des textes. In: *JADT 2008, Actes des 9èmes Journées internationales d'Analyse statistique des Données Textuelles*. Vol. *2, 733-744*. Lyon: Presses de l'ENS. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/longree-luong-mellet.pdf].

Luong Xuan; Mellet, Sylvie (2003). Mesures de distance grammaticale entre les textes. *Corpus* 2 (« Les distances intertextuelles »), *141-166.* [see Revues.org: http://corpus.revues.org/34].

Magri, Véronique; Purnelle, Gérald (2012). Mot à mot, brin par brin: les suites [Nom préposition Nom] comme motifs. In: Anne Dister, Dominique Longrée, Gérald Purnelle (eds.), *JADT 2012, Actes des 11e Journées internationales d'analyse statistique des données textuelles: 659-673*, Liège. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm].

Mayaffre, Damon (2008a). Quand 'travail', 'famille', 'patrie' co-occurrent dans le discours de Nicolas Sarkozy. Etude de cas et réflexion théorique sur

la co-occurrence. In: Serge Heiden, Bénédicte Pincemin (eds.), *JADT 2008, 9es journées internationales d'analyse statistique des données textuelles: vol. 2, 811-822*. Lyon: Pul, [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/mayaffre.pdf].

Mayaffre, Damon (2008b). De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. *Sémantique & Syntaxe 9, 53-72*.

Marouzeau, Jean (1922). *L'ordre des mots dans la phrase latine.* Vol. I. *Les groupes nominaux*. Paris: Champion.

Marouzeau, Jean (1938). *L'ordre des mots dans la phrase latine.* Vol. II. *Le verbe*. Paris: Champion.

Marouzeau, Jean (1949). *L'ordre des mots dans la phrase latine: Les articulations de l'énoncé*. Paris: Champion.

Marouzeau, Jean (1953). *L'ordre des mots en latin*. Volume complémentaire. Paris: Champion.

Martin, Robert; Muller, Charles (1964). Syntaxe et analyse statistique. La concurrence entre le passé antérieur et le plus-que-parfait dans *La Mort le Roi Artu. Travaux de Linguistique et de Littérature* 2: 1-27.

Mellet, Sylvie (1987). *L'imparfait de l'indicatif en latin.* Louvain – Paris: Peeters.

Mellet, Sylvie (1994). Logiciels d'exploitation de la banque de données de textes latins du L.A.S.L.A. *Revue, Informatique et Statistique dans les Sciences humaines 30, 91-108*. [http://promethee.philo.ulg.ac.be/RISSHpdf/Annee1994/Articles/SMellet.pdf]

Mellet, Sylvie (1996). Les atouts de la lemmatisation. In : G. Moracchini (ed.) *Actes du Colloque international «Bases de données linguistiques: conceptions, réalisations, exploitations»: 309-316* (Corte 11-13 octobre 1995), Univ. de Corse / Univ. Nice Sophia Antipolis.

Mellet, Sylvie (1998). Les tragédies de Sénèque vues à travers Hyperbase. In: S. Mellet ; M. Vuillaume (eds.), *Mots chiffrés et déchiffrés, Mélanges offerts à Étienne Brunet: 255-271*. Paris: Champion.

Mellet, Sylvie (2002a). Lemmatisation et encodage grammatical: un luxe inutile? *Lexicometrica*, [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/thema/thema1/spec1-texte2.pdf].

Mellet, Sylvie (2002b). La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte. *Médiévales* 42 (« Le latin dans les textes »), *13-26.*

Mellet, Sylvie; Barthélemy, Jean-Pierre (2007). La topologie textuelle: légitimation d'une notion émergente. *Lexicometrica* n°spécial, 12 pages. [see *Lexicometrica* web-site: http://lexicometrica.univ-paris3.fr/numspeciaux/special9/mellet.pdf].

Mellet, Sylvie; Longrée, Dominique (2009). Syntactical motifs and textual structures. *Belgian Journal of Linguistics 23, 161-173* (« New Approaches in Textual Linguistics »).

Mellet, Sylvie; Longrée, Dominique (2012). Légitimité d'une unité textométrique: le motif. In: Anne Dister, Dominique Longrée, Gérald Purnelle (eds), *JADT 2012, Actes des 11e Journées internationales d'analyse statistique des données textuelles: 715-728.* Liège [see : *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Mellet,%20Sylvie%20et%20al.%20-%20Legitimite%20d%27une%20unite%20textometrique.pdf]

Mellet, Sylvie; Purnelle, Gérald (2002). Les atouts multiples de la lemmatisation: l'exemple du latin. In: A. Morin; P. Sébillot (eds.), *JADT 2002(*2), *529-538, 6èmes Journées internationales d'Analyse statistique des Données Textuelles,* Saint-Malo: Irisa et Inria.

Perrot, Jean (1978). Ordre des mots et structures linguistiques. *Langages 50, 17-26.*

Pincemin, Bénédicte (2008). Modélisation textométrique des textes. In: Serge Heiden, Bénédicte Pincemin (eds), *JADT 2008, Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles, vol. II, 949-960.* Lyon: Presses Universitaires de Lyon, [see *Lexicometrica* web-site*:* http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/pincemin.pdf].

Pincemin, Bénédicte; Heiden, Serge; Lay, Marie-Hélène; Leblanc, Jean-Marc; Viprey, Jean-Marie (2010). Fonctionnalités textométriques: Proposition de typologie selon un point de vue utilisateur. In: Sergio Bolasco, Isabella Chiari, Luca Giuliano (eds.), *JADT 2010, Statistical Analysis of Textual Data -Proceedings of 10th International Conference*. Rome: Edizioni Universitarie di Lettere Economia Diritto. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-0341-0354_023-Pincemin.pdf].

Philippart de Foy, Caroline (2008). *Hagiographie et statistique linguistique: étude d'un corpus de traductions médiolatines d'origine grecque*, thèse non publiée de l'Université Nice Sophia Antipolis.

Purnelle, Gérald (1989). Recherche automatique de groupes verbaux récurrents et de formules dans les fichiers latins lemmatisés. *Revue, Informatique et Statistique dans les Sciences humaines, 25, 157-191*. [http://promethee.philo.ulg.ac.be/RISSHpdf/Annee1989/Articles/GPurnelle.pdf]

Purnelle, Gérald (1996). Utilisation d'une banque de données des textes latins lemmatisés et analysés. Problèmes spécifiques aux données linguistiques. In: G. Moracchini (ed.) *Actes du Colloque international «Bases de données linguistiques : conceptions, réalisations, exploitations», 295-307* (Corte 11-13 octobre 1995), Univ. de Corse / Univ. Nice Sophia Antipolis.

Quiniou, Solen; Cellier, Peggy; Charnois, Thierry; Legallois, Dominique (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. In: Anne Dister, Dominique Longrée, Gérald Purnelle, *JADT 2012, Actes des 11e Journées internationales d'analyse statistique des données textuelles: 821-833.* Liège. [see *Lexicometrica* web-site*:* http:// lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm].

Renard, Denis (2000). Les parties du discours chez les personnages du *Satiricon*. In: Martin Rajman, Marie Decrauzat, Jean-Cédric Chappelier (eds.), *JADT 2000, 5èmes Journées internationales d'Analyse statistique des Données Textuelles.* Lausanne. [see *Lexicometrica* web-site*:* http://lexicometrica.univ-paris3.fr/jadt/jadt2000/pdf/55/55.pdf].