Mathématiques et sciences humaines
Mathematics and social sciences

187 | Automne 2009
Journée 2007 de la Société Francophone de Classification

Représentations du texte pour la classification arborée et l'analyse automatique de corpus. Application à un corpus d'historiens latins

Text formal reductions for tree analysis and automatic classification. Application to a latin historians corpus

Sylvie Mellet, Nguyen Xuan Luong, Dominique Longrée et Jean-Pierre Barthelemy

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/msh/11152
DOI : 10.4000/msh.11152
ISSN : 1950-6821

Éditeur
Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée
Date de publication : 30 décembre 2009
Pagination : 107-121
ISSN : 0987-6936

Référence électronique
Sylvie Mellet, Nguyen Xuan Luong, Dominique Longrée et Jean-Pierre Barthelemy, « Représentations du texte pour la classification arborée et l'analyse automatique de corpus. Application à un corpus d'historiens latins », *Mathématiques et sciences humaines* [En ligne], 187 | Automne 2009, mis en ligne le 15 décembre 2009, consulté le 21 avril 2019. URL : http://journals.openedition.org/msh/11152 ; DOI : 10.4000/msh.11152

Math. Sci. hum / Mathematics and Social Sciences (47e année, n° 187, *2009*(3), p. 107-121)

REPRÉSENTATIONS DU TEXTE POUR LA CLASSIFICATION ARBORÉE ET L'ANALYSE AUTOMATIQUE DE CORPUS. APPLICATION À UN CORPUS D'HISTORIENS LATINS

Sylvie MELLET[1], Xuan LUONG[2], Dominique LONGRÉE[3], Jean-Pierre BARTHÉLEMY[4]

RÉSUMÉ – *Nous exposons ici différentes méthodes de classification automatique des textes littéraires et nous en comparons les performances, notamment en ce qui concerne leur aptitude à traduire les structurations génériques du corpus. Nous montrons qu'une approche topologique des textes, qui prend en compte leur linéarité fondamentale, c'est-à-dire l'ordre macro- et micro-structurel de leurs différentes unités constitutives, permet d'obtenir de meilleurs résultats classificatoires que les méthodes traditionnelles qui tendent à négliger cette structure linéaire.*

MOTS-CLÉS – Analyse arborée, Classification générique, Motif, Structures linéaires, Topologie textuelle, Treillis, Voisinage.

SUMMARY – Text formal reductions for tree analysis and automatic classification. Application to a latin historians corpus

In this paper, we present different methods of automatic classification applied to a corpus of literary texts and we compare their different results; in particular, we evaluate how each of them is suitable for exhibiting the generic classification of the corpus. We demonstrate that a topological approach of the texts which takes into account their linearity, i.e. the order of their micro- and macro-structures, results in better clustering than traditional quantitative methods which leave generally out of count this linear structure.

KEYWORDS – Generic classification, Lattice, Linear textual structures, 'Motif' (pattern), Neighbourhood, Texts topological approach, Tree analysis

1. INTRODUCTION

La recherche présentée dans cet article est le fruit d'une collaboration interdisciplinaire entre laboratoires : elle associe un mathématicien (Jean-Pierre Barthélemy), un informaticien (Xuan Luong) et deux linguistes (Dominique Longrée et Sylvie Mellet). Son objectif est de préciser et d'évaluer l'intérêt d'une approche topologique des textes

[1] Laboratoire BCL, Université Nice Sophia-Antipolis, CNRS ; MSH de Nice, 98 bd E. Herriot, F-06200 Nice ; mellet@unice.fr

[2] Laboratoire BCL, Université Nice Sophia-Antipolis, CNRS ; MSH de Nice, 98 bd E. Herriot, F-06200 Nice ; luong@unice.fr

[3] Laboratoire LASLA, Université de Liège, Quai Roosevelt 1B, B-4000 Liège et Laboratoire BCL, Nice ; Dominique.Longree@ulg.ac.be

[4] Laboratoire TAMCIC, ENST Bretagne, CNRS, BP 832, F-29285 Brest Cédex ; jp.barthelemy@enst-bretagne.fr

en vue de leur classification. Nous avons justifié ailleurs[5] la notion de « topologie textuelle » en montrant qu'elle ne reposait pas sur un simple transfert métaphorique d'un domaine à l'autre, mais qu'elle avait une véritable légitimité épistémologique. Nous reviendrons bien sûr sur ce point. Mais nous insisterons surtout ici sur l'intérêt d'évaluer et de comparer les diverses représentations de la chaîne textuelle que nous sollicitons pour mettre au point des méthodes de classification automatique des textes littéraires qui soient propres à rendre compte des différentes structurations latentes d'un corpus, c'est-à-dire susceptibles de superposer à la banale classification par auteur et à la classification chronologique, une classification par genres et, si possible, par sous-genres.

2. LE TEXTE : UN ESPACE ORDONNÉ

Très vite, les méthodes de classification issues de la lexicométrie se sont heurtées au fait qu'un texte n'est pas un sac dans lequel seraient rassemblées en vrac ses unités constitutives. Un texte est à tout le moins une chaîne linéaire, donc un espace ordonné. Même lorsque, pour les besoins de l'analyse, on réduit le texte à certaines de ses unités constitutives, celles-ci restent intégrées à des structures, locales ou globales, qui leur donnent sens. L'examen des deux extraits suivants en donnera une illustration ; l'un est emprunté à Maupassant, l'autre à Hugo :

> *Je n'*hésitais *point. Je* connaissais *une hôtellerie d'amour non loin de ma demeure, et j'y* courus*, et j'y* entrai *comme font ces gens qui se jettent à l'eau pour voir s'ils savent encore nager. Je* nageais*, et fort bien. Et je* demeurai *là longtemps* (...)*. Puis je me* trouvai *dans la rue* (...)*. Je me* sentais *maintenant calme et sûr de moi* (...)*, et prêt encore, me* semblait*-il, pour des prouesses*. [Maupassant, *Toine*]

[Imp. Imp. PS. PS. Imp. PS. PS. Imp. Imp.]

> *Toute cette cavalerie, étendards et trompettes au vent,* (...) descendit *la colline de la Belle-Alliance,* s'enfonça *dans le fond redoutable* (...)*, et* disparut *dans la fumée, puis, sortant de cette ombre,* reparut *de l'autre côté du vallon* (...)*. Ils* montaient*, graves, menaçants, imperturbables; dans les intervalles de la mousqueterie et de l'artillerie, on* entendait *ce piétinement colossal. Etant deux divisions, ils* étaient *deux colonnes : la division Wathier* avait *la droite, la division Delord* avait *la gauche.* [Hugo, *Les misérables*]

[PS. PS. PS. PS. Imp. Imp. Imp. Imp. Imp.]

Les deux textes contiennent le même nombre d'occurrences de deux temps verbaux : chacun présente 4 passés simples et 5 imparfaits ; si donc on s'intéresse aux temps de la narration dans ces textes comme paramètre d'analyse susceptible de caractériser l'écriture de chacun d'eux, un pur dénombrement d'occurrences ne mettra en évidence aucun écart entre les deux extraits ; or, comme on le sait, les styles des deux auteurs et les genres de leurs œuvres sont bien distincts. En revanche, la différence se marquera du moment où l'on prendra en compte la répartition des occurrences dans la chaîne linéaire que constitue le texte : chez Maupassant, dans le premier texte, on relève une alternance régulière entre imparfaits et passés simples, alors que chez Hugo, à un récit constitué par la suite des 4 passés simples, succède une description contenant les 5

[5] Cf. [Mellet, Barthélemy, 2007].

imparfaits : les occurrences de deux temps s'y présentent ainsi totalement groupées et opposent les deux parties de l'extrait.

La reconnaissance de la linéarité des textes a donc conduit au développement de travaux qu'on pourrait regrouper sous la bannière « Beyond Bag-of-Words »[6] et dont l'impact a été particulièrement important en stylométrie[7]. Pour notre part, nous partons du postulat qu'un texte est d'abord un ensemble (E) d'unités linguistiques qui ne sont pas indépendantes les unes des autres, et qui est muni d'une structure ou, plus exactement, de plusieurs structures imbriquées dont l'union constitue cet ensemble. La prise en compte de l'existence dans les textes de ces micro-structures ordonnées et récurrentes nous a conduits à développer une réflexion sur les propriétés topologiques du texte et sur les différentes transformations ou réductions qui permettent de lui appliquer des traitements quantitatifs à des fins classificatoires.

3. LES MÉTHODES TRADITIONNELLES : CLASSIFICATION SUR DÉNOMBREMENT D'OCCURRENCES

Avant d'aller plus avant dans la présentation des méthodes classificatoires que nous avons mises au point, il ne semble pas inutile de préciser ici rapidement les avantages, ainsi que les limites des méthodes reposant sur le schéma d'urne et sur des dénombrements d'occurrences[8]. Dans le cadre de ces méthodes, les textes sont ramenés à un ensemble d'unités, par exemple, l'ensemble des formes graphiques qui les constituent, ou l'ensemble des lemmes, c'est-à-dire l'ensemble des formes du dictionnaire auxquelles ces textes font appel, ou encore l'ensemble des catégories grammaticales qui y sont représentées. Dans l'exemple présenté ici et portant sur un corpus de 12 textes d'historiens latins[9], nous avons choisi comme unités d'analyse les temps narratifs de l'indicatif. Notre corpus est en effet lemmatisé et étiqueté[10]. On peut donc automatiquement dénombrer les occurrences des temps verbaux, parmi lesquels nous avons retenus ceux de la narration (parfait – équivalent du passé simple français –, imparfait, plus-que-parfait). Ces dénombrements ont permis de constituer un classique tableau de contingence à partir duquel nous avons réalisé une classification non hiérarchique : nous avons choisi l'analyse arborée[11] pour représenter les distances entre les textes et leurs nœuds de regroupement (cf. Figure 1).

On voit que la méthode traditionnelle adoptée ici permet une première classification par auteur : les œuvres de César et de Quinte-Curce se regroupent dans le bas de la figure, les deux livres de Suétone sont à l'autre extrémité de l'arbre et ceux de Tacite sont au centre. L'auteur de la *Guerre d'Espagne*, un imitateur de César, apparaît isolé de son modèle. La méthode du schéma d'urne permet donc ici globalement d'assez bons résultats, mais elle ne permet pas de faire émerger, par exemple, des classifications

---

[6] Du nom du Workshop de la 28e Conférence annuelle internationale ACM SIGIR : voir http://www.cs.cmu.edu/~pbennett/pubs.html

[7] Cf. [Holmes 1998]. Cf. aussi la notion de « stylome » chez [Van Halteren *et alii*, 2005].

[8] Cf. [Longrée, Luong, 2003].

[9] Ces textes sont les suivants : livre 2 de la *Guerre Civile* et livre 5 de la *Guerre des Gaules* de César (resp. Civ2 et Gall5) ; *Guerre d'Espagne* (Hisp) par un imitateur anonyme de César ; livres 3 et 4, puis 9 et 10 de la *Vie d'Alexandre* racontée par Quinte-Curce (resp. QC3_4 et QC9_10) ; livres 3, 12, 14 et 15 des *Annales* de Tacite (Ann3, Ann12, Ann14 et Ann15) et enfin la *Vie de Jules César* et la *Vie de Tibère* racontées par Suétone (resp. Iul et Tib).

[10] C'est-à-dire que chaque forme du texte est rapportée à son entrée de dictionnaire et affectée d'une étiquette ou code alphanumérique qui résume ses propriétés morpho-syntaxiques.

[11] Cf. [Barthélemy, Luong, 1987, 1998].

d'ordre générique, c'est-à-dire liées au genre ou au sous-genre dont relève chaque texte. Ainsi, les œuvres de Suétone et Quinte-Curce sont à l'opposé sur l'arbre, alors qu'il s'agit, dans les deux cas, d'œuvres à composante biographique. La *Vie d'Agricola* de Tacite reste groupée avec les autres œuvres de ce même auteur, œuvres qui, elles, n'ont rien de biographique. Comme le suggérait l'exemple un peu simpliste de Hugo et de Maupassant, la prise en compte de la linéarité et de la structuration interne des textes pourrait permettre d'aller plus loin, ce qui nous a incités à développer des outils de classification complémentaires.

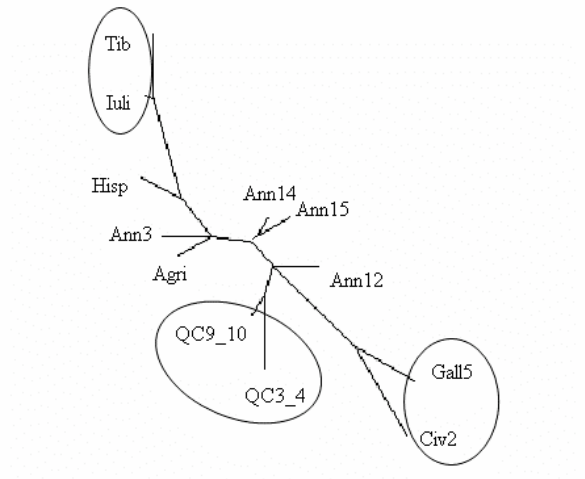


FIGURE 1. Classification par analyse arborée de 12 œuvres d'historiens latins en fonction du nombre global d'occurrences des divers temps narratifs dans chacune d'elles

4. LE DÉCOUPAGE EN TRANCHES : UNE PREMIÈRE PRISE EN COMPTE DE LA LINÉARITE DU TEXTE

Une première méthode permettant de prendre en compte la structure linéaire des textes est leur découpage en tranches[12]. Rejetant un découpage « naturel » basé sur le contenu du texte, nous avons choisi de découper arbitrairement chaque texte en un nombre fixe de tranches contiguës, de taille égale, et un travail précédent nous a permis de constater empiriquement qu'un découpage à large empan, en cinq ou six tranches, était celui qui, sur notre corpus, donnait les meilleurs résultats. Nous avons retenu ici un découpage en cinq tranches. Dans chacune des tranches, nous avons dénombré le même paramètre caractérisant, à savoir le nombre d'occurrences d'indicatifs parfaits en proposition principale, ce qui permet d'obtenir un profil de la distribution de ce temps au fil des 5 tranches, profil dont on peut penser qu'il rend compte, au moins en partie, de la dynamique narrative de chaque texte. Si nous recourons toujours au dénombrement d'occurrences, nous l'appliquons donc désormais à un texte réduit à la succession de ses prédicats principaux et séquencé en 5 parties égales. Pour le même corpus de 12 textes historiques, nous avons un tableau de contingence à 5 colonnes et la représentation arborée que nous obtenons à partir de ce tableau est bien différente de la précédente.

[12] Cf. [Longrée, Luong, Mellet, 2004], [Longrée, Mellet, 2008] et [Longrée, Mellet, Luong, 2006].

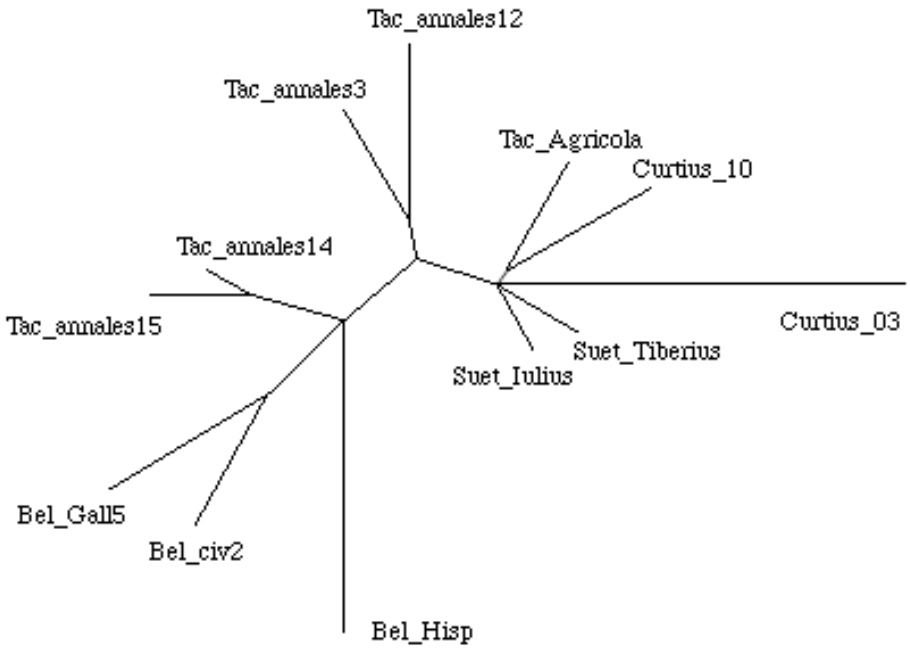


FIGURE 2. Classification par analyse arborée de 12 œuvres d'historiens latins en fonction de la distribution du parfait au fil d'un découpage des textes en 5 tranches

La bipartition la plus immédiate et intuitive de l'arbre fait apparaître, sur la droite de la figure, un premier groupement qui rassemble, à partir d'un même embranchement, tous les textes relevant de la biographie ou, du moins, à forte composante biographique : on trouve là les deux *Vies* de Suétone, très proches l'une de l'autre, auxquelles se rattachent d'une part les branches des deux livres de Quinte-Curce et d'autre part, la branche portant la *Vie d'Agricola*, qui, cette fois, s'écarte des autres œuvres de Tacite. Le facteur générique l'emporte ici sur le style et la personnalité de l'écrivain. La méthode utilisée permet donc bien de prendre en compte des différences sous-génériques, ignorées par les méthodes traditionnelles. Cette méthode de découpage en tranches a pourtant ses limites : elle suppose notamment que les textes considérés présentent entre eux des variations de taille assez réduites. D'autres méthodes ont donc dû être développées, en parallèle à une réflexion méthodologique sur la nature même de l'objet « texte ».

5. LE TEXTE COMME OBJET À STRUCTURE TOPOLOGIQUE

Soit le texte conçu comme une structure linéaire constituée d'un ensemble d'événements linguistiques[13]. Considérant chacun de ces événements linguistiques, par exemple, l'occurrence d'un mot ou d'une catégorie grammaticale, comme un point remarquable de la chaîne textuelle, l'analyste ne saurait détacher l'unité observée de son contexte immédiat, c'est-à-dire de la portion textuelle jugée pertinente pour l'analyse et

[13] Cette définition n'exclut pas la possibilité de considérer aussi le texte dans sa dimension « réticulaire », c'est-à-dire comme constitué de réseaux thématiques enchevêtrés dont certains sont purement internes et dont d'autres font écho à d'autres textes antérieurs. Voir sur ce point les travaux de [Viprey, 2000 et 2002].

qui comprend un certain nombre d'autres événements linguistiques situés avant et après lui. L'exemple liminaire des extraits de Hugo et Maupassant montrait de manière presque simpliste qu'une forme de passé simple devait être étudiée en fonction de son entourage temporel et que son analyse dépendrait, partiellement mais nécessairement, de la qualité homogène ou hétérogène de ce contexte. Par ailleurs la pertinence de la taille de ce contexte varie selon les faits étudiés, la taille globale du texte, l'objectif de la recherche, etc. L'empan contextuel doit donc être défini par le linguiste en fonction de ces différents paramètres, éventuellement de manière arbitraire. Cette façon d'appréhender la notion de contexte vaut quelle que soit l'unité linguistique sous étude.

Résumons-nous : nous sélectionnons donc une occurrence particulière qui est un élément x du texte E, situé à un point particulier de sa chaîne linéaire. Nous délimitons autour de x un empan textuel constitué par un certain nombre d'éléments x_i du texte E entourant le point x ; nous avons donc construit un *voisinage* de x. La taille des voisinages pouvant varier au gré de l'analyste, nous obtenons par suite une *famille de voisinages*. L'opération peut être répétée pour chaque point x de E et l'ensemble de tous ces voisinages fonctionne dès lors comme un système T, système de sous-ensembles non ordonnés de E.

Le texte E, muni de ce système de sous-ensembles T, constitue bien un espace topologique, puisque ce système de sous-ensembles T vérifie les propriétés définitoires suivantes : l'ensemble vide et E appartiennent à T, l'intersection de deux éléments de T (*i.e.* de deux voisinages) est bien un élément de T, c'est-à-dire un voisinage ; et l'union de deux voisinages est aussi un voisinage, élément de T.

5.1. LES VOISINAGES : ILLUSTRATION

Un cas concret permettra de mieux faire comprendre ce que nous entendons par voisinage au sein d'un texte. Dans l'exemple présenté ci-dessous, tiré du début du livre 2 de la *Guerre Civile* de César, nous avons d'abord réduit le texte à la chaîne des codes numériques représentant les temps de verbes des propositions principales[14].

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 12 12 12 12 14 (14 14 14 14 14 **14** 14 14 14 14 14)

Nous avons ensuite défini arbitrairement une taille de voisinage de 11 unités (symbolisée par les parenthèses) et, à chacune des unités constitutives du texte, nous avons affecté une mesure de son voisinage. Cette mesure, que nous appellerons densité, a été ici calculée en fonction du nombre d'occurrences du code 14 présente dans ce voisinage, le code 14 étant affecté au parfait de l'indicatif, temps fondamental de la narration. On obtient ainsi la suite numérique des densités correspondant à chaque voisinage[15] : ainsi au premier des deux voisinages mis en évidence par le parenthésage dans la chaîne ci-dessus, correspond une densité de 1 (un seul code 14 dans la parenthèse), au deuxième, une densité de 11 (onze codes 14 dans la parenthèse).

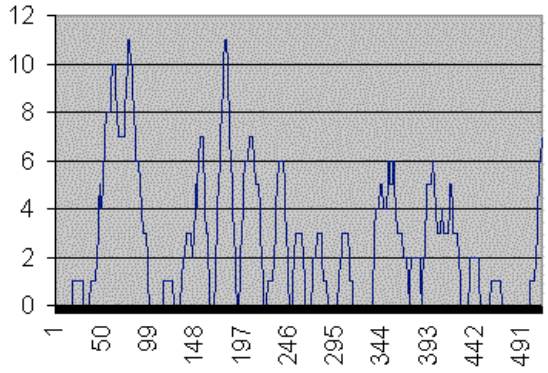
X X X X X 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 **1** 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 2 3 4 5 4 4 5 6 6 7 8 8 8 8 8 9 10 10 10 10 9 8 7 7 7 7 7 7 7 8 9 10 11 **11**

[14] Le code 11 représente le présent de l'indicatif, le code 12 l'imparfait, le code 14 le parfait, le code 15 le plus-que-parfait.

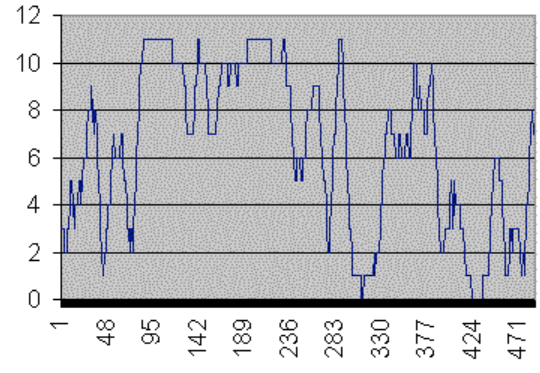
[15] Les premières unités textuelles ne permettent pas de construire ce voisinage de taille 11 et sont donc représentées par un X.

5.2. LES VOISINAGES : RÉSULTATS

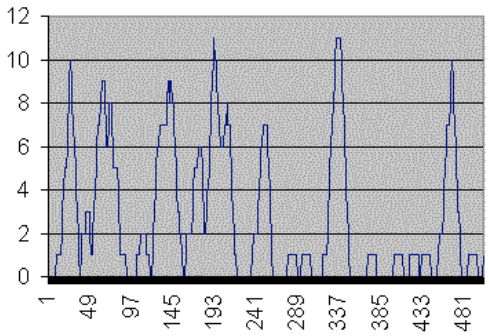
La suite numérique des densités peut donner lieu à une représentation graphique sous forme de courbes. Voici celles de la *Guerre Civile* - livre 2, de la *Guerre des Gaules* - livre 5, de la *Guerre d'Espagne* et de la *Vie de Tibère*.



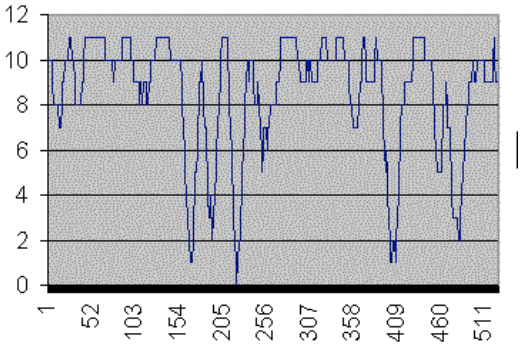
Guerre Civile 2 (César)



Guerre d'Espagne (imitateur de César)



Guerre des Gaules 5 (César)



Vie de Tibère (Suétone)

Les profils de textes obtenus sont assez parlants : sans entrer dans une analyse approfondie, remarquons simplement la différence très nette entre les courbes des deux livres de César, à gauche, opposées, à droite, à celles la *Guerre d'Espagne*, due à un continuateur, et de la *Vie de Tibère* de Suétone. Ces profils confirment la classification initiale de la Figure 2 où ces textes s'opposaient également à partir d'un simple dénombrement d'occurrences, mais ils permettent en outre de rendre compte de la dynamique interne des textes : on constate ainsi que les livres de César commencent sur un mode qui tend à exclure le parfait alors que *Vie de Tibère* accepte ce temps verbal dès les premiers paragraphes. Cependant ces courbes restent actuellement peu exploitables en l'absence de méthode fiable pour comparer des courbes de longueurs différentes. Il s'agit là d'un chantier à approfondir pour lequel nous sommes ouverts à toute collaboration.

Outre sa capacité à fonder le texte comme espace topologique, la notion de voisinage présente l'avantage majeur d'autoriser les analyses multidimensionnelles[16] : on peut ainsi travailler sur des voisinages complexes prenant en compte divers paramètres et permettant de conjoindre approche qualitative et approche quantitative.

[16] Cf. [Luong, Juillard, Mellet, Longrée, 2007].

En revanche, la notion de voisinage n'induit aucune contrainte sur l'ordre des éléments étudiés au sein du voisinage, d'où notre définition des voisinages comme des sous-ensembles non ordonnés. Cette notion de voisinage reste donc insuffisante pour rendre compte de la structure textuelle où l'ordre des unités simples aussi bien que de certaines structures complexes et la succession ordonnée des différentes zones textuelles jouent bien évidemment un rôle majeur. C'est pourquoi nous proposons de fonder l'approche topologique des textes sur un autre concept, celui de « motif », lequel pourrait aussi devenir un élément de classification endogène des corpus étudiés.

6. LES MOTIFS

6.1. DÉFINITION

On appelle ici « motif » l'association récurrente de n éléments de l'ensemble (E) muni de sa structure linéaire qui donne une pertinence aux relations de successivité et de contiguïté. Ainsi, si l'ensemble (E) est composé de x occurrences des éléments A, B, C, D, E, F, un premier motif pourra être la récurrence du groupe linéairement ordonné ABD, un autre motif pourra être la récurrence du groupe AA. Apparemment simple, cette définition soulève pourtant un certain nombre de difficultés théoriques touchant aux limites à donner à cette définition[17]. Signalons simplement ici que cette définition peut englober les « segments répétés » de Salem, mais se veut plus large en raison, notamment, de deux propriétés fondamentales des motifs : d'une part le motif a vocation à être hétérogène (c'est-à-dire à associer des éléments de nature lexicale et grammaticale), d'autre part il est susceptible d'accueillir une variable en son sein.

À l'instar des voisinages, les motifs sont eux aussi de bons candidats à fonder une structure topologique des textes[18]. Les résultats, quoique encore incomplets, donnent à penser que les motifs, sans satisfaire à toutes les propriétés topologiques auxquelles répondent les voisinages, en satisfont cependant suffisamment pour conférer au texte au moins une structure de « treillis ».

Et quoi qu'il en soit, les motifs fournissent de très bons paramètres d'analyse pour la classification, comme nous allons le voir maintenant.

6.2. ILLUSTRATION À PARTIR DE MOTIFS SYNTACTICO-STYLISTIQUES

Pour évaluer l'intérêt de la notion de motifs pour une classification des textes, nous travaillerons sur un exemple qui met en jeu les différents types de propositions – principales et subordonnées – de divers textes narratifs latins, en admettant que la phrase est une unité pertinente pour la détection des motifs liés à ce type d'unités linguistiques (cadre syntaxique phrastique classique). L'objectif est d'obtenir une classification des textes du corpus en fonction, d'une part, des motifs dominants dans chacun des textes, d'autre part, de leur combinaison mutuelle et de leur distribution au fil de chaque texte.

Les propositions qui nous intéressent sont d'une part la proposition principale, d'autre part certaines subordonnées qui contribuent à poser un cadre circonstanciel à l'action, tels les ablatifs absolus (c'est-à-dire une forme de proposition participiale) et

[17] Cf. [Mellet, Barthélemy, 2007] ; [Longrée, Luong, Mellet, 2008].
[18] *Ibid*.

les subordonnées en *cum* + subjonctif (c'est-à-dire l'équivalent approximatif des propositions en *alors que*, *comme*). L'organisation de toutes ces propositions structure en effet la narration et les stylisticiens ont depuis longtemps montré qu'elles contribuaient à caractériser le style d'un auteur : certains en effet les utilisent avec parcimonie, d'autres ne reculent pas devant leur accumulation. Certains préfèrent placer les subordonnées en début de phrase, pour poser d'abord le cadre de l'action principale : on peut alors parler de « motifs cadratifs ». D'autres, au contraire, pratiquent volontiers la « relance syntaxique » en fin de phrase, rajoutant après coup des éléments informatifs secondaires, que l'on qualifiera de « motifs de rallonge ».

Pour illustrer notre propos, voici un exemple de texte (*Guerre des Gaules* 2) accumulant les motifs « cadratifs » en tête d'une phrase dont le début est ici symbolisé par D. Avant la principale (symbolisée par p), César en pose en effet les circonstances par le biais de deux ablatifs absolus et d'une proposition en *cum*, les deux structures étant symbolisées indifféremment par E, ce qui nous donne un motif du type DEEEp :

> *(D) Cuius aduentu, spe inlata (E) militibus ac redintegrato (E) animo, cum pro se quisque in conspectu imperatoris etiam extremis suis rebus operam nauare cuperet (E), paulum hostium impetus tardatus est (p). (César, De bello gallico, II, 25, 3)*
>
> *(D) A l'arrivée de celui-ci, l'espoir ayant été rendu (E) aux soldats et leur moral ayant été retrouvé (E), alors que chacun pour lui-même désirait (E) apporter sa contribution sous le regard du général en chef, même si sa situation était la plus critique, l'assaut des ennemis fut un peu ralenti (p).*

À l'opposé Tacite préfère, lui, rejeter après la proposition principale p, une série d'ablatifs absolus et de propositions en *cum,* série qui se place juste avant la fin de la phrase, symbolisée ici par F. Ces propositions postposées, symbolisées par E, apparaissent ici sous forme de « rallonges ». On obtient dans ce cas un motif du type pEEEEF.

> [...] *insurgere paulatim, munia senatus magistratuum legum in se trahere (p), nullo aduersante (E), cum ferocissimi per acies aut proscriptione cecidissent (E), ceteri nobilium, quanto quis seruitio promptior, opibus et honoribus extollerentur (E) ac nouis ex rebus aucti tuta et praesentia quam uetera et periculosa mallent (F). (Tacite, Annales, I, 2, 1)*
>
> [...] *il commença à s'élever peu à peu, à attirer (p) à lui l'autorité du sénat, des magistrats, des lois, nul ne lui résistant (E), alors que les plus acharnés avaient péri (E) par la guerre ou la proscription, que tous les autres nobles, en fonction de l'empressement de chacun à servir, s'élevaient (E) en richesses et honneurs et que, ayant gagné par ce changement politique, ils préféraient (E) la sécurité présente aux périls passés (F).*

Chez les auteurs latins, on observe bien sûr des usages mixtes, combinant l'une et l'autre des structures. Les différents motifs, leur fréquence respective et leur combinaison caractérisent donc le style d'un auteur ; leur distribution pourrait aussi caractériser les différentes parties d'une œuvre (les parties introductives ou de commentaires entre les passages narratifs offrant, par exemple, un cadre plus accueillant aux accumulations de subordonnées).

Chaque texte du corpus se trouve donc caractérisé par un ensemble de descripteurs qui rendent compte, sous forme numérique, de l'usage (en fréquence et en distribution locale et globale) des motifs au sein de chaque texte du corpus. L'ensemble de ces descripteurs donne le profil de chaque texte eu égard aux paramètres d'analyse retenus et ces profils sont ensuite intégrés à une matrice rectangulaire dont les textes fournissent l'intitulé des lignes et dont les descripteurs de profil fournissent les

colonnes. À partir de là peuvent être appliquées les méthodes classiques de calcul de distance ; pour notre part nous adoptons à nouveau le calcul de distance et la classification hiérarchique de l'analyse arborée proposée par Xuan Luong & Jean-Pierre Barthélémy[19], analyse qui a l'avantage de constituer des classes tout en imposant simultanément une contrainte de proximité.

Pour évaluer la pertinence classificatoire de la méthode, nous avons travaillé sur 38 textes du corpus historique latin[20] et, d'abord sur 18 motifs, c'est-à-dire sur un ensemble comportant 12 motifs « cadratifs » et 6 motifs de « rallonge ». Voici les premiers résultats :

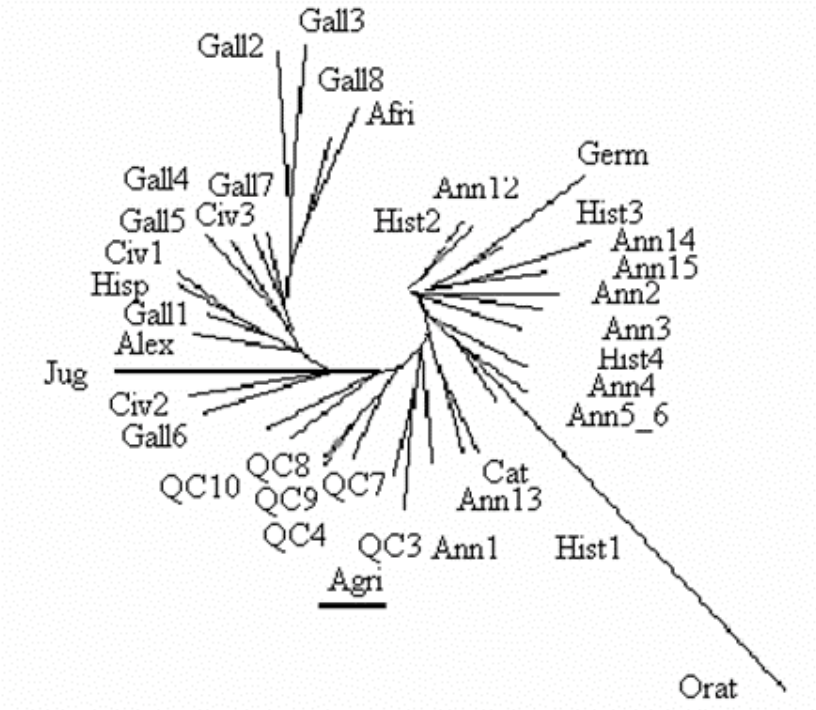


FIGURE 3. Classification par analyse arborée de 38 textes d'historiens latins en fonction de l'emploi de 18 motifs syntactico-stylistiques

Les regroupements obtenus sont remarquables, car ils tiennent compte non seulement des auteurs, mais aussi des différences sous-génériques entre textes : ainsi, les œuvres de Tacite s'opposent à une extrémité de l'arbre à l'ensemble du corpus césarien à l'autre extrémité de ce même arbre ; le corpus à caractère biographique se regroupe au centre de l'arbre, et l'on voit que la *Vie d'Agricola* (soulignée) se détache des autres œuvres de Tacite qui, elles, relèvent de l'histoire annalistique ; une autre œuvre de Tacite d'un tout autre genre, un traité de rhétorique appelé le *Dialogue des Orateurs*, se détache très nettement de toutes les autres œuvres. On peut aussi remarquer un regroupement de premiers livres, Ann1, Hist1 et Ann13, dont nous avons pu

[19] Cf. [Barthélemy, Guénoche, 1988], [Barthélemy, Luong, 1987, 1998]. Pour une discussion sur diverses méthodes de représentations graphiques des données textuelles, cf. aussi [Lebart, 2004].

[20] César : Gall1, Gall2, Gall3, Gall4, Gall5, Gall6, Gall7, Civ1, Civ2, Civ3 ; Imitateurs de César: Gall8, Alex, Afri, Hisp ; Salluste: Jug et Cat ; Quinte-Curce : QC3, QC4, QC7, QC8, QC9, QC10 ; Tacite : Germ, Orat, Agri, Hist1, Hist2, Hist3, Hist4, Ann1, Ann2, Ann3, Ann4, Ann5_6, Ann12, Ann13, Ann14, Ann15.

souligner ailleurs[21] la position souvent marginale. Notons enfin que la zone d'intermédiarité de l'arbre accueille aussi les œuvres qui, chronologiquement, se situent entre les deux extrêmes de la période couverte par le corpus.

6.3. PERTINENCE ET STABILITÉ DE LA MÉTHODE

L'arborée des motifs ne manque pas, elle non plus, d'intérêt et confirme la pertinence de leur sélection sur des critères externes (hérités de la tradition philologique) :

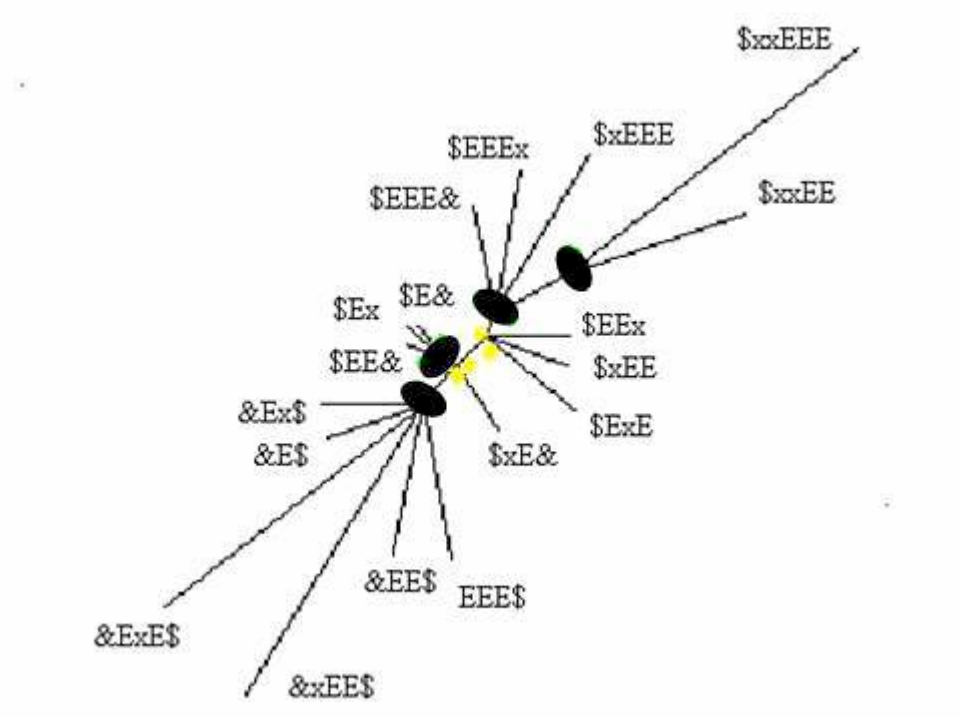


FIGURE 4. Classification par analyse arborée des 18 motifs retenus pour l'analyse des textes

Dans cet arbre (où $ symbolise le début ou la fin de phrase, où & marque la proposition principale et x une proposition quelconque), la bipartition la plus intuitive et immédiate de l'arbre voit se regrouper les motifs de « rallonge » d'un côté, et les motifs « cadratifs » de l'autre. Les regroupements sont pour la plupart très stables (ellipses noires) ; si l'on tient compte de l'historique de la construction de l'arbre, on voit que les deux premiers regroupements sont d'une part le regroupement de deux motifs « cadratifs » longs, $xxEE et $xxEEE, et d'autre part le regroupement des 6 motifs de « rallonge » (terminés par $). Ces deux regroupements sont de même niveau.

La constitution de ces nœuds paraît cohérente, puisqu'elle permet de regrouper des motifs ayant une partie commune. D'où l'idée que nous avons eue de sommer les occurrences de ces motifs proches et de réduire la liste à 13 motifs seulement, de manière notamment à évaluer la stabilité de la méthode.

[21] Cf. [Longrée, Mellet, 2008].

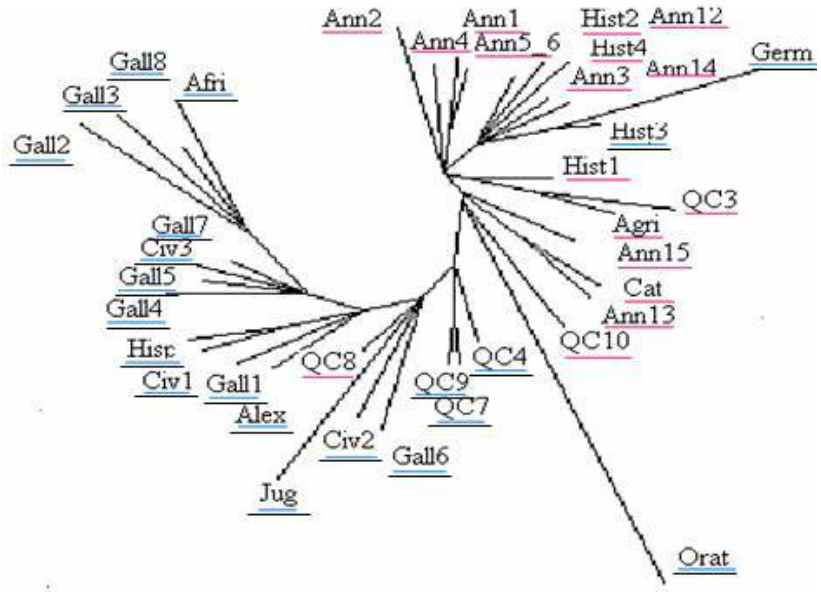


FIGURE 5. Classification par analyse arborée de 38 textes d'historiens latins en fonction de l'emploi de 13 motifs syntactico-stylistiques

En comparant la Figure 3 et la Figure 5, on observe une stabilité de la forme générale des arborées absolument remarquable : les regroupements les plus stables sont ici soulignés d'un double trait. C'est Tacite qui semble le plus sensible à l'opération de regroupement des motifs : ainsi la *Vie d'Agricola* (Agri), reste proche de QC3, mais se rapproche avec QC3 du reste de l'œuvre tacitéenne. Cela pourrait vouloir dire que, chez Tacite, chaque motif, bien séparé des autres et bien identifié, a son rôle propre à jouer dans l'écriture de l'auteur et n'est donc pas interchangeable avec les autres. À moins que le regroupement n'accentue le poids des motifs vraiment importants chez lui.

La comparaison des arborées des motifs s'avère elle aussi très instructive : les regroupements suggérés par le premier arbre de la Figure 4 ne modifient que fort peu la représentation.

On note simplement que la branche regroupant les motifs de « rallonge » EEE$, xEE$ et ExE$, en bas, à gauche de l'arbre se détache légèrement des trois autres motifs de « rallonge » (&Ex$, &E$, &EE$), décomposant ainsi un peu l'étoile que formaient les branches de ces divers motifs sur la représentation arborée à 18 motifs (fig. 4). Cette légère décomposition est sans doute liée au fait que le poids de cette branche devient un peu lourd à la suite du regroupement des motifs EEE$, xEE$ et ExE$. Mais globalement, la comparaison des deux arbres confirme la forte pertinence des motifs choisis, ainsi que leur effet structurant. Elle conforte aussi la stabilité et la fiabilité de l'outil de classification mis au point. Pour tester la méthode, on pourrait par ailleurs aussi faire varier le nombre des textes : les premières recherches effectuées en ce sens semblent indiquer que plus le nombre de textes est important, plus les regroupements et donc les classifications restent stables, avec une structuration d'autant plus grande des familles de motifs.

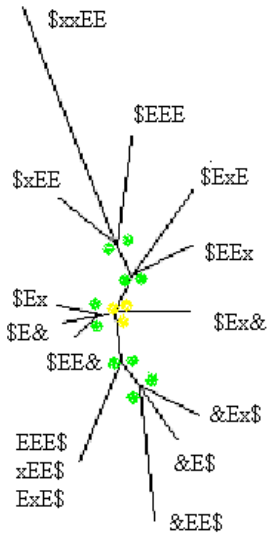


FIGURE 6. Classification par analyse arborée des 13 motifs obtenus après regroupements sur la liste initiale

7. CONCLUSIONS ET PERSPECTIVES

Au nombre des points positifs des méthodes proposées ici, il faut noter que l'analyse s'appuie sur des paramètres qui confèrent aux structures textuelles des propriétés modélisables. En outre, aussi bien les voisinages que les motifs fournissent des bases de classifications stables. Il est également à souligner que ces méthodes sont, à toutes les étapes, multidimensionnelles. Le profil affecté à chaque texte et intégré à la matrice de calcul des distances est un profil complexe qui répercute la diversité des paramètres nécessaires pour rendre compte d'un texte dans sa richesse et sa singularité.

Ces méthodes ont toutefois leurs limites : on a vu ainsi que, pour les voisinages, se posait la question des méthodes à utiliser pour comparer des courbes de longueurs différentes. Par ailleurs, à ce stade, l'aspect générique n'est que peu mis en évidence dans les dernières analyses par motifs. Nous comptons tenter de récupérer cette partie de l'information en faisant progresser les méthodes dans deux directions. Il s'agira tout d'abord d'accentuer le caractère multidimensionnel des analyses : pour appréhender un texte dans toute sa complexité on ne saurait se contenter de travailler sur une seule de ses dimensions; le linguiste et/ou le stylisticien doit donc essayer de détecter plusieurs des dimensions pertinentes constitutives de la structure du texte ou des textes sous étude, de caractériser chacune de ces dimensions par la fréquence et la distribution des motifs afférents et de voir ensuite comment ces différentes classes de motifs s'articulent entre elles pour donner une image multidimensionnelle de la structure textuelle. Pour améliorer nos méthodes, il s'agira ensuite d'en combiner les avantages : nous pensons ainsi étudier la distribution des motifs au fil des textes par la reprise du découpage de ceux-ci en un nombre défini de parties, ou encore dénombrer les motifs apparaissant dans le cadre de voisinages. Nous espérons ainsi arriver à des méthodes de classifications plus performantes, permettant d'obtenir, sinon une représentation de la topologie interne des textes, à tout le moins, une topologie externe, image de la structuration, – ou plutôt d'une structuration possible – du corpus.

Mais l'intérêt principal de cette étude nous semble résider dans le fait que les méthodes proposées appliquent les outils somme toute classiques et bien rôdés que sont le dénombrement d'occurrences et la classification arborée à des représentations du texte originales, dans des manipulations contrôlées à la fois par le cadre théorique emprunté à la topologie et par la qualité des résultats obtenus. Ce faisant, elles obligent à approfondir le concept même de textualité.

BIBLIOGRAPHIE

BARTHÉLEMY J.-P., GUÉNOCHE A., *Les arbres et les représentations des proximités*, Paris, Masson, 1988.

BARTHÉLEMY J.P., LUONG X., « Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmiques et applications à l'analyse de données textuelles », *Mathématiques et Sciences humaines* 100, 1987, p. 57-80.

BARTHÉLEMY J.-P., LUONG X., « Représenter les données textuelles par des arbres », in JADT 1998, *Actes des 4e Journées Internationales d'analyse de données textuelles,* Université de Nice, UMR 6039, 1998, p. 49-70.

BAKER L.D, MCCALLUM A.K, "Distributional Clustering of Words for Text Classification", *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, Melbourne, 1998, p. 96–103.

HOLMES D.I., "The Evolution of Stylometry in Humanities Scholarship", *Literary and Linguistic Computing*, vol. 13, n°3, 1998, p. 111-117.

LEBART L., « Validité des visualisations de données textuelles », in G. Purnelle, C. Fairon, A. Dister (éds), JADT 2004, *7e Journées internationales d'Analyse statistique des Données Textuelles*, Université Catholique de Louvain, Presses universitaires de Louvain, vol. 2, 2004, p. 708-715.

LONGRÉE D., LUONG X., « Temps verbaux et linéarité du texte : recherches sur les distances dans un corpus de textes latins lemmatisés », *Corpus 2*, 2003, p. 119-140.

LONGRÉE D., LUONG X., MELLET S., « Temps verbaux, axe syntagmatique, topologie textuelle : analyse d'un corpus lemmatisé », in G. Purnelle, C. Fairon, A. Dister (éds), JADT 2004, *7e Journées internationales d'Analyse statistique des Données Textuelles*, Université Catholique de Louvain, Presses universitaires de Louvain, vol. 2, 2004, p. 743-752.

LONGRÉE D., LUONG X., MELLET S., « Les motifs : un outil pour la caractérisation topologique des textes », communication soumise aux JADT 2008, *9e Journées internationales d'Analyse statistique des Données Textuelles*, Lyon, mars 2008.

LONGRÉE D., MELLET S., « Temps verbaux et prose historique latine : à la recherche de nouvelles méthodes d'analyse statistique », in G. Purnelle et J. Denooz (éds), *Ordre et cohérence en latin*, Genève, Droz, 2008, p. 117-128.

LONGRÉE D., MELLET S., LUONG X., « Distance intertextuelle et classement des textes d'après leur structure : méthodes de découpage et analyses arborées », in J.M. Viprey (éd.), JADT 06, *8e Journées internationales d'Analyse statistique des Données Textuelles*, Besançon, Presses universitaires de Franche-Comté, vol. 2, 2006, p. 643-654.

LUONG X., JUILLARD M., MELLET S., LONGRÉE D., "Trees and after: The Concept of Text Topology. Some applications to Verb-Form Distributions in Language Corpora", *Literary and Linguistic Computing* 22, 2, 2007, p. 167-186.

MELLET S., BARTHÉLEMY J.-P., « La topologie textuelle : légitimation d'une notion émergente », *Lexicometrica 7 « Topographie et topologie textuelles »*, 2007, [cf. http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf].

VIPREY J.-M., « Hypertexte de corpus littéraire : cartographie et statistique multidimensionnelle », in M. Rajman, J.C. Chappelier (éds), JADT 2000, *5e Journées internationales d'Analyse statistique des Données Textuelles*, École Polytechnique Fédérale de Lausanne, vol. 2, 2000, p. 535-539.

VIPREY J.-M., « Dynamisation de l'analyse micro-distributionnelle des corpus textuels », in A. Morin, P. Sébillot (éds), JADT 2002, *6e Journées internationales d'Analyse statistique des Données Textuelles*, Saint-Malo, IRISA / INRIA, vol. 2, 2002, p. 779-790.

VAN HALTEREN H., BAAYEN H., TWEEDIE F., HAVERKORT M., NEIJT A., "New Machine Learning Methods Demonstrate the Existence of a Human Stylome", *Journal of Quantitative Linguistics*, vol. 12, fasc. 1, 2005, p. 65-77.