

Les motifs : un outil pour la caractérisation topologique des textes

Dominique Longrée¹, Xuan Luong², Sylvie Mellet²

¹LASLA, Université de Liège, Quai Roosevelt, 1B - Bat. A4, B-4000 Liège – Belgique

²Laboratoire BCL, Université Nice Sophia-Antipolis, CNRS
MSH de Nice ; 98 bd E. Herriot, 06 200 NICE

Abstract

In this paper, we will describe how “motifs” (“patterns”) can be used as a new conceptual tool for a topological characterization of texts. We will first summarize the stages of our approach of textual topology, point out the main arguments in favour of such a formal representation of texts and stress the relevance of the concept of “neighborhood”. We will then show how the concept of “pattern” proves to be very useful to take into account, - within the frame of textual topology-, the order of textual micro-structures (a main difference with other types of tools). At last, we will describe two applications based on syntactic “motifs” selected in a corpus of historical Latin texts.

Résumé

Dans cette présentation, nous proposons d'appuyer la caractérisation topologique des textes sur la notion de « motif ». Nous commencerons par rappeler les étapes de notre réflexion sur la topologie textuelle, ainsi que quelques arguments en faveur d'une telle représentation du texte, notamment à partir de la notion de « voisinage ». Puis nous montrerons que la notion de « motif », tout en répondant aux propriétés définitoires de la topologie, permet d'introduire aussi la notion de structure ordonnée (ce que ne permet pas celle de « voisinage »). Nous terminerons par deux exemples d'applications à partir de motifs syntaxiques repérés dans un corpus d'historiens latins.

Mots-clés : motifs, topologie textuelle, voisinages, classification, analyses arborées, historiens latins, phrases narratives.

1. Introduction

L'objectif de cette présentation¹ est de proposer et tenter de valider un nouvel outil pour la caractérisation topologique des textes : le « motif ». Celui-ci sera défini comme un élément structurant de la texture discursive et la pertinence de son statut comme outil d'analyse sera vérifiée d'un double point de vue :

- théorique d'abord : il s'agira de fonder épistémologiquement la notion et de voir comment elle s'intègre à une représentation véritablement topologique des textes ;

¹ Ce travail a été réalisé dans le cadre d'un programme de collaboration franco-belge CNRS/FNRS soutenu par la DRI (Direction des Relations Internationales) du CNRS et le CGRI (Commissariat général aux relations internationales) de la Communauté française de Belgique.

- méthodologique ensuite : il s'agira de tester l'utilité heuristique de l'outil et de voir ce qu'il apporte à l'étude de la structuration interne des textes et à leur caractérisation au sein d'un corpus contrastif.

Le corpus d'étude pour ces premiers tests de validation sera le corpus d'historiens latins que nous avons déjà utilisé et présenté dans plusieurs travaux antérieurs (voir références bibliographiques), notamment dans notre présentation aux JADT 2004. Outre les atouts déjà exposés précédemment (corpus lemmatisé et étiqueté selon les normes du LASLA ; disponible et outillé sous Hyperbase ; homogène quant au genre, mais offrant des variations chronologiques et sous-génériques intéressantes ; documenté par des siècles de critique philologique ; etc.), la reprise de ce corpus va nous permettre de comparer les résultats obtenus par diverses méthodes d'analyse et donc de mieux évaluer l'apport spécifique du nouvel outil proposé.

L'exposé sera organisé en trois temps :

- des rappels sur le cadre théorique dans lequel se situe notre proposition ; en effet celle-ci participe d'une réflexion générale sur la possible structure topologique des textes et la notion de « motif » est une pierre apportée à l'édifice en cours de construction ;
- une définition de ce que nous appelons « motifs » et de leurs propriétés ;
- une application à des faits de syntaxe et de style dans le corpus retenu.

2. Rappels sur une élaboration théorique en cours

2.1. *Le texte comme espace topologique*

Notre conception du texte comme espace topologique a d'abord été assez intuitive et nous utilisons le terme « topologique » de manière métaphorique, par un transfert quelque peu cavalier de la notion d'un champ disciplinaire à l'autre. Mais la récurrence de cette métaphore dans nos propos nous a obligés à en préciser le sens exact et à en assurer les fondements épistémologiques, ce que nous avons commencé à faire dans trois publications récentes² grâce, notamment, à un dialogue très fructueux avec J.-P. Barthélemy.

Cette réflexion nous a conduits à constater qu'un texte, conçu comme une structure linéaire, pouvait répondre à la définition d'un espace topologique au sens strict du terme ; en mathématiques en effet un espace topologique est un ensemble E , muni d'un système T de sous-ensembles de E qui vérifie les propriétés suivantes :

- l'ensemble vide et E appartiennent à ce système T ;
- une intersection d'éléments de T est dans T ;
- une union d'éléments de T est aussi dans T ;

2.2. *La notion de voisinage*

C'est la notion de « voisinage », utilisée notamment dans Longrée *et al.* 2004, qui fournit au texte la structuration nécessaire pour répondre à cette définition. En effet, dans une perspective textuelle, un voisinage est un empan de texte autour d'une occurrence spécifiée,

² Luong X., Juillard M., Mellet S. et Longrée D., 2007 ; Mellet S. et Barthélemy J.-P., 2007 ; Longrée D., Luong X. et Mellet S., 2007.

empan dont la taille est arbitrairement définie et qui est caractérisé à partir de divers paramètres ; il représente donc un nombre fixé d'éléments x_i de (E) entourant un point x ; il formalise la notion de contexte textuel et l'on peut, selon les besoins de l'étude, faire varier la taille des contextes ainsi définis, si bien que chaque point x se trouve muni d'une famille de voisinages. Or l'ensemble des voisinages constitue un système T de sous-ensembles **non ordonnés** de E , vérifiant les propriétés topologiques déjà signalées :

- le voisinage zéro appartient à T ;
- le texte entier conçu comme voisinage appartient à T ;
- l'intersection de 2 voisinages est un voisinage ;
- l'union de 2 voisinages est un voisinage.

Les voisinages confèrent donc bien au texte une structure d'espace topologique.

Nous avons montré ailleurs les avantages concrets, pour l'analyse textuelle, du recours aux voisinages qui permettent de travailler simultanément sur des structures locales et des structures globales, de prendre en compte la continuité du texte et d'en avoir une approche multidimensionnelle. Mais nous nous sommes heurtés aussi à quelques-unes de leurs limites, notamment qu'ils ne rendent pas compte des structures ordonnées récurrentes du texte ni de l'organisation micro-structurelle de celui-ci.

La notion de « motif » que nous introduisons aujourd'hui a donc pour objectif de permettre la prise en compte de l'ordre des éléments à l'intérieur des sous-ensembles de (E) et de travailler sur la récurrence des structures ordonnées.

3. La notion de « motif »

3.1. Définition

On définit le motif comme un sous-ensemble ordonné de (E) formé par l'association récurrente de n éléments de l'ensemble (E) muni de sa structure linéaire. Ainsi, si le texte est formé d'un certain nombre d'occurrences des éléments A, B, C, D, E , un motif pourra être la micro-structure récurrente ACD ou bien encore AA , etc., sans qu'on préjuge ici de la nature des éléments A, B, C, D, E en question : il s'agit de faire du motif le cadre – ou l'enveloppe – accueillant un ensemble de paramètres à définir et susceptibles de caractériser les divers textes d'un corpus, voire les différentes parties d'un texte.

3.2. Questions

La définition du motif paraît simple ; elle ne va pas toutefois sans soulever un certain nombre de questions auxquelles seules des réflexions plus approfondies et de nombreuses expérimentations permettront de répondre.

- Un motif peut-il être hétérogène, c'est-à-dire accueillir des éléments de nature différente ?
- Peut-on considérer qu'une séquence contrainte (telle que les lexies composées, ou – plus subtil – des syntagmes régis ou toute autre structure de dépendance imposée par la syntaxe) constitue légitimement un motif ?

- Un motif peut-il « enjamber » un séparateur de phrase ? Et de façon plus générale, doit-on ou non tenir compte de la ponctuation dans la reconnaissance du motif (par exemple ABD est-il équivalent de AB, D ?)
- Un motif tolère-t-il la présence d'une variable en son sein ? $ABxD$ est-il un motif et, si oui, est-il comparable à $ABxxD$?

Autant de questions auxquelles devront répondre les études ultérieures.

3.3. Propriétés topologiques

Les motifs sont-ils de bons candidats pour structurer le texte en espace topologique ?

Ils constituent assurément un système T de sous-ensembles ordonnés de E qui vérifie les propriétés formelles suivantes :

- l'ensemble vide appartient à ce système T ;
- l'intersection de deux motifs est un motif ;

En revanche l'union de deux motifs n'est un motif que si les deux motifs en question sont consécutifs. Il manque donc là une propriété cruciale.

Mais on peut observer cependant que pour une série de motifs apparentés il existe toujours une borne inférieure et une borne supérieure³ ; par exemple, les motifs CAB , $CAAB$, $CAAAB$, $CxAB$, $CAxAB$, $CAAxAB$ ont pour borne supérieure le motif $CxxxxB$ et pour borne inférieure le motif CA ; nous avons donc une structure d'ordre partiel, qui correspond à ce qu'en topologie on appelle un « treillis ».

4. Applications

4.1. Le choix des motifs syntaxiques

Comme nous l'avons dit, le motif fournit un cadre formel accueillant à toutes sortes d'unités linguistiques et textuelles ; par exemple, dans le domaine lexical, il pourrait englober ce qu'A. Salem a reconnu sous le nom de « segments répétés » ou ce que, dans un autre domaine d'application, on appelle les « unités phraséologiques », ou encore – dans la mesure où le motif admet la présence d'une variable en son sein – ce qu'en milieu anglo-saxon on appelle « collocations »⁴. Mais le motif pourrait aussi rassembler des structures phonologiques, ou des structures métriques. Pour notre part, nous avons choisi de travailler ici sur des structures syntaxiques récurrentes, traditionnellement reconnues pour caractériser divers types de phrases narratives chez les historiens latins.

En effet, les stylisticiens considèrent qu'il existe, dans les œuvres historiques, deux manières fondamentales d'organiser la phrase narrative. Dans la première, l'auteur pose d'emblée le *cadre* circonstanciel de l'événement narré : il accumule donc en début de phrase des propositions subordonnées et, en particulier, des propositions participiales (ablatifs absolus) et des conjonctives circonstancielles qu'on traduit volontiers en français par « alors que, comme » (*cum* + subjonctif). La phrase se termine par la proposition principale relatant le fait

³ Voir notamment Mellet et Barthélemy, 2007.

⁴ Sur toutes ces notions relevant finalement de la cooccurrence régulière d'éléments lexicaux, voir ici même l'exposé de Damon Mayaffre.

de premier plan. Dans le second cas, l'historien rajoute, après la principale, des éléments circonstanciels supplémentaires qui viennent donc en *rallonge* d'une phrase déjà complexe et qui paraissait terminée. Selon les auteurs, la préférence va tantôt à l'un, tantôt à l'autre de ces procédés.

Les deux types de phrases admettent, bien sûr, des types mixtes, mais ils fournissent les deux pôles extrêmes caractéristiques de deux manières d'écrire bien différentes et ils sont donc à la source de motifs syntaxiques récurrents. Si, en effet, on symbolise par D le début de phrase, par F la fin de phrase, par p la proposition principale, par E l'ensemble des circonstanciels qui nous intéressent (ablatifs absolus et *cum* + subjonctif) et par x tout autre type de proposition, on détecte dans les textes deux séries de motifs opposés :

DEEEp	pEEEF
DEExP	pEExF
DExEp	pExEF
DEEp	pEEF
DExp	pxEF
DEp	pEF
etc.	

dont on trouve une illustration dans les deux extraits suivants. On notera au passage que chacune de ces séries possède un plus grand dénominateur commun, *i.e.* une borne inférieure qui légitime le statut du motif dans la représentation topologique du texte.

Premier extrait : César, *De bello gallico*, II, 25, 3 ; **Motif D E E E p**

(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).

« (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). »

Second extrait : Tacite, *Annales*, I, 2, 1 ; **Motif p E E E E F**

[...] *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).

« [...] 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 préféreraient (E) la sécurité présente aux périls passés (F). »

Pour travailler sur de tels motifs nous avons donc dû réduire les textes à la succession des étiquettes représentant les diverses propositions qui en forment les chaînes phrastiques ; il s'agit de codes alphanumériques à quatre caractères dont les deux premiers donnent la nature de la proposition et les deux derniers donnent le temps et le mode du verbe de la proposition. Les quatre \$ marquent les frontières de phrase. Dans un premier temps nous obtenons donc des séquences comme celle-ci,

Guerre d'Espagne :

\$\$\$\$ AD44 AD44 LN35 BN35 BN35 BN35 CD11 &&14 \$\$\$\$ &&14 \$\$\$\$ NN32 &&14
 \$\$\$\$ AD44 \$\$\$\$ &&12 &&12 \$\$\$\$ SX15 BN35 &&12 AD44 XK32 \$\$\$\$ [...]

où AD44 code un ablatif absolu au participe parfait,

LN35 code une proposition subordonnée relative au subjonctif plus-que-parfait,

&&14 code une proposition principale à l'indicatif parfait, etc.

Cette symbolisation, directement issue des étiquettes affectées aux textes dans la base de données, est évidemment peu maniable. Nous l'avons donc convertie pour la simplifier et faire apparaître plus clairement les motifs qui nous intéressent ; nous avons gardé l'esperluette pour les principales, le E pour les circonstancielles sous étude, un simple point pour toutes les autres propositions. Ainsi, dans la représentation ci-dessous d'un extrait de la Guerre d'Espagne et d'un extrait de la Vie de Tibère, voit-on bien apparaître, pour le premier extrait, le motif EE en début de phrase et, pour le second, le motif &E en fin de phrase (parfois augmenté en &EE) :

Guerre d'Espagne :

/EE.EEE.&/&/.&/E/&&/E&E./&/.&/&/EE.E..E...E/&/.&.../.
 &.&/&.../&&/EE&.../E&.../E&.&.../&/.&.&/E&.../&&/EE&/&
 E&/E&.../E.&&/.&&/E.&&.../E&/EE&&/E&/&.../E&/.&/&
 ./&./EE.&&/&&/&.../E...E&.../EE&&...

Vie de Tibère :

&E.E/E&E.../&E.../&.../&/EE/E&/&.../.../&/&...&.../&
 &.../&.../&&.../&E/.&.../&&/&.../E.&&.../&.&E/&/&/&
 &E/EE&.../&.../&/&.../&E/&.../&/&.../E.../&.../E.&/E&...&
 .../E&&/E.&.../E&...&E&/&.../.../&&.../.../E.E&.../EE...&/
 &.../&E/E...&.../.....

4.2. L'opposition binaire entre les motifs choisis

La première chose que nous avons vérifiée est la pertinence des motifs choisis et de leur rôle éventuel dans la bipartition des types de phrases narratives chez les différents auteurs du corpus. Nous avons donc sélectionné dans un premier temps 18 motifs susceptibles d'être caractéristiques et nous avons calculé quels étaient les proximités et les éloignements entre chacun des motifs en fonction de leur emploi dans chacun des 15 textes du corpus ; ce calcul fournit une classification des 18 motifs représentée dans l'arborée ci-dessous, où l'on voit très nettement se regrouper en haut de l'arbre les motifs de rallonge⁵ ; en bas de l'arbre se trouve les motifs cadratifs lourds (comprenant de trois à cinq propositions subordonnées initiales) tandis que la zone d'intermédiarité de l'arbre accueille les cadratifs plus légers et, de ce fait, plus banals.

⁵ Rappelons que dans cette figure, comme dans les suivantes, D code le début de phrase, F la fin de phrase, p une principale, E un ablatif absolu ou une circonstancielle en *cum* + subjonctif et x toute autre proposition.

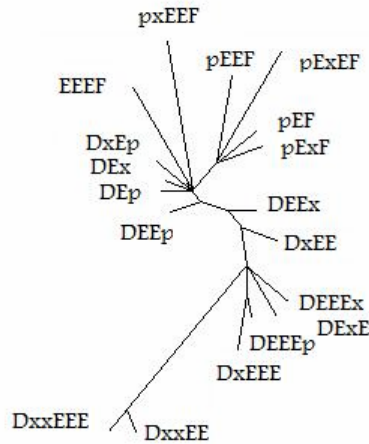


Figure 1 : classification par analyse arborée de 18 motifs syntaxiques selon leur distribution dans un corpus de 15 textes historiques latins

Une telle classification⁶, qui légitime la bipartition des motifs ainsi que leur rôle discriminant dans la définition du style narratif de chaque texte, est confirmée par une application à un corpus plus vaste. Comme on le voit sur la figure ci-dessous, on obtient une classification assez comparable des 18 motifs à partir d'un corpus de 38 textes :

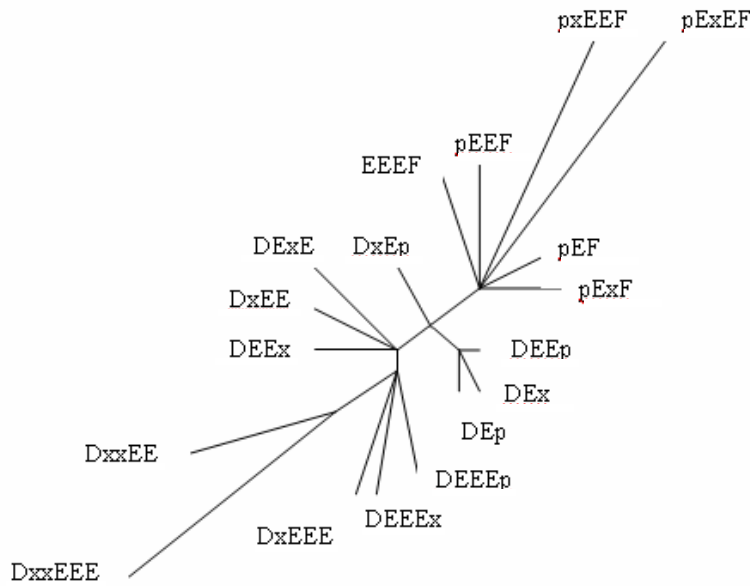


Figure 2 : classification par analyse arborée de 18 motifs syntaxiques selon leur distribution dans un corpus de 38 textes historiques latins

Linguistiquement, cette deuxième classification est encore meilleure que la précédente, avec le très fort regroupement, sur un même noeud, de tous les motifs de rallonge (terminés par F).

⁶ Dont les indices d'agrégations sont tous supérieurs à 0,84.

La stabilité de cette classification et de certains de ses regroupements nous a donc incités à nous interroger sur la possibilité de rassembler certains motifs proches sous une seule entité qu'on pourrait appeler « supra-motif ». Ce regroupement aurait une triple légitimation :

- linguistique : à l'évidence les motifs susceptibles d'être regroupés en un seul ont des structures syntaxiques très proches et jouent un rôle fonctionnel dans la phrase, ainsi qu'un rôle stylistique comparables ;
- théorique : la reconnaissance de ces « supra-motifs » manifeste concrètement la propriété topologique de borne inférieure de la structure ;
- classificatoire : la stabilité des nœuds dans les arborées ci-dessus montre que, quel que soit le corpus choisi, les regroupements se font de manière semblable.

Nous proposons donc une liste de 13 motifs en lieu et place des 18 motifs initiaux, dont les quatre « supra-motifs » suivants :

$$DEEE = DEEE_x + DEEE_p$$

$$D_{xx}EE = D_{xx}EE + D_{xx}EEE$$

$$D_xEE = D_xEE + D_xEEE$$

$$EEE_xEE_ExEF = pEEEF + pxEEF + pExEF$$

Il reste cependant à vérifier que ces regroupements ne modifient pas trop sensiblement la classification des motifs ainsi réorganisés au sein d'une nouvelle liste plus synthétique. Voici cette nouvelle classification que l'on comparera à celle de la figure 1 :

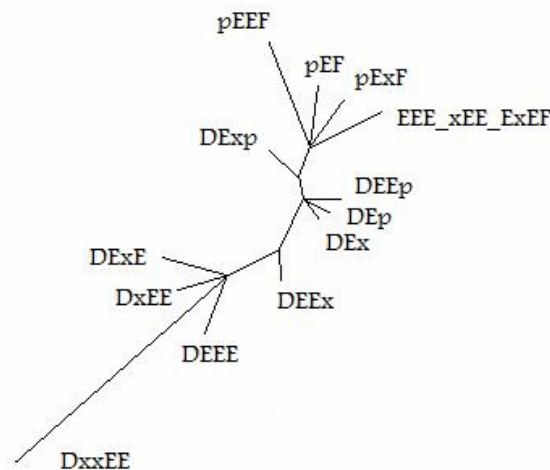


Figure 3 : classification par analyse arborée de 13 motifs ou « supra-motifs » syntaxiques selon leur distribution dans un corpus de 15 textes historiques latins

Cette comparaison montre une légère amélioration de la classification quand on travaille sur un regroupement en 13 motifs : la zone d'intermédiarité subsiste, ce qui est normal, mais les deux pôles opposés se structurent de manière plus homogène autour de deux nœuds principaux ; et en particulier les motifs de rallonge EEEF et pxEEF ne viennent plus s'adjoindre à un nœud portant aussi des motifs cadratifs.

4.3. Valeur opératoire des motifs syntaxiques pour caractériser la structure d'un texte et pour comparer entre eux les textes du corpus

Dans l'état actuel de notre recherche nous avons testé deux applications possibles de la notion de motif : l'une a donné des résultats décevants, l'autre au contraire des résultats très satisfaisants.

4.3.1. Caractérisation de la structure interne des textes

Nous avons d'abord voulu voir si la distribution des motifs au fil de chaque texte pouvait caractériser la structure globale du récit et sa dynamique interne ; en somme, les motifs étaient-ils inégalement répartis dans chaque texte et certains auteurs les réservaient-ils plutôt à l'introduction et à la conclusion, d'autres au corps même de la narration ? Nous avons en effet déjà observé cet effet structurant dans la distribution des formes de parfait, temps de la narration par excellence, en découpant arbitrairement chaque texte en cinq tranches successives et en dénombrant les occurrences de ce temps verbal dans chacune des tranches. Le corpus était légèrement différent de celui que nous étudions ici et rassemblait des textes de longueurs plus homogènes ; dans ces conditions, les résultats permettaient d'obtenir, en fonction des différents profils de distribution du parfait, une classification générique très fine⁷. Ici le corpus rassemble des textes dont les écarts de taille sont sensiblement plus importants : notre expérience tend à prouver que la classification est alors moins bonne. Et de fait, la représentation arborée ci-dessous n'est pas aussi pertinente que celle à paraître dans [Mellet et Longrée 2008] ; elle affiche néanmoins quelques bons regroupements sous-génériques, tels que la proximité de la *Vie d'Agricola* (Tacite) avec la *Vie de Tibère* (Suétone) ou même celle de *Domitien* (Suétone) :

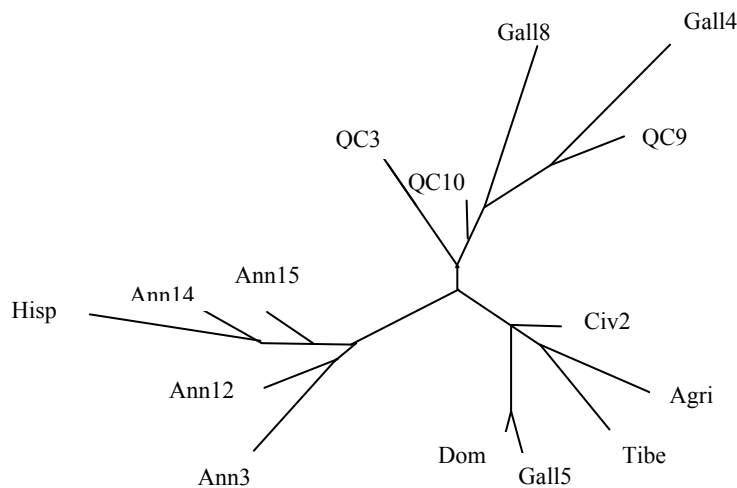


Figure 4 : classification par analyse arborée de 15 textes historiques latins en fonction de la distribution, au fil des textes, du parfait de l'indicatif en principale

Nous avons donc appliqué la même méthode pour la distribution des motifs, mais comme on le voit sur l'arborée de la figure 5, les résultats sont décevants ; et il en va de même quel que soit le motif retenu.

⁷ Voir Longrée, Luong et Mellet 2004 ; Longrée et Mellet, 2008.

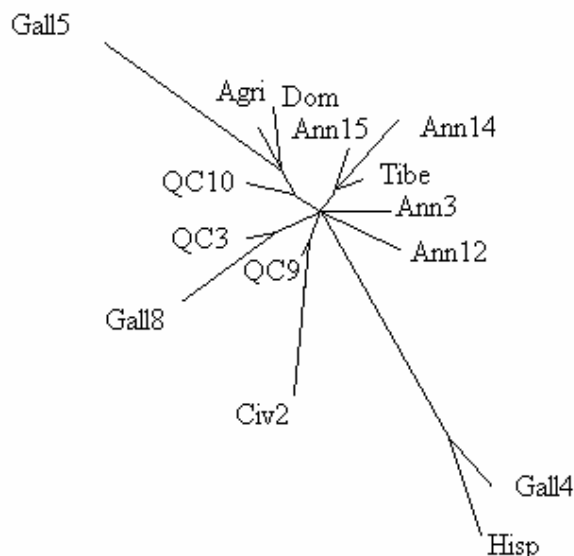


Figure 5 : classification par analyse arborée de 15 textes historiques latins en fonction de la distribution, au fil des textes, du motif de rallonge EF

L'arbre n'est pas très bien structuré et met surtout en évidence la forte opposition entre le livre 5 de la *Guerre des Gaules* d'un côté, le livre 4 de la même œuvre, associé à la *Guerre d'Espagne* de l'autre, qui appartiennent tous les trois au corpus césarien. Le choix d'un autre motif ou d'une famille de motif ne donne pas de meilleurs résultats. Si donc les motifs participent à la structuration globale du texte, ce n'est pas dans le cadre d'un découpage grossier en cinq parties successives, mais dans une dynamique plus subtile qui fait certainement la part belle aux zones de transition et qui rythme le texte en passages ou mouvements plus courts et plus subtils.

4.3.2. Comparaison des textes du corpus

En revanche le simple dénombrement global des occurrences des motifs dans les textes du corpus permet d'aboutir à une classification intéressante, que l'on travaille sur la liste des 18 motifs initiaux ou sur celle des 13 motifs comprenant quelques regroupements en « supra-motifs ».

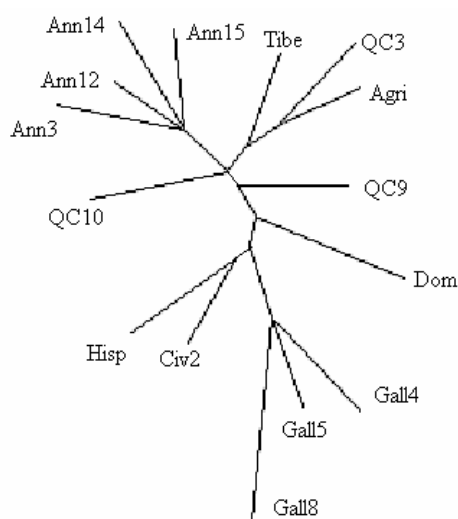


Figure 6 : classification par analyse arborée de 15 textes historiques latins en fonction du nombre d'occurrences global des 18 motifs syntaxiques

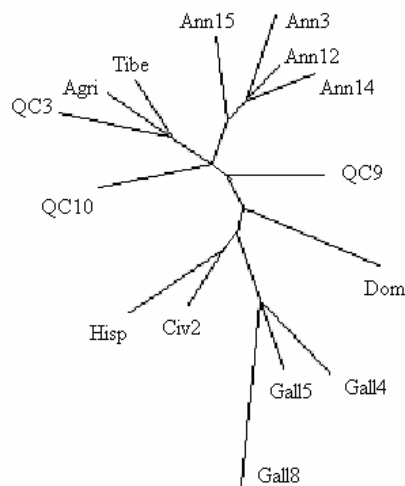


Figure 7 : classification par analyse arborée de 15 textes historiques latins en fonction du nombre d'occurrences global des 13 motifs ou « supra-motifs » syntaxiques

Dans les deux cas on observe une bonne bipartition de l'arbre, opposant aux deux extrêmes le corpus césarien (en bas) et le corpus taciteen (en haut) avec les livres de Quinte-Curce et de Suétone en zone d'intermédiation ; cette classification par auteur est doublée d'une classification chronologique, mais elle est surtout dominée par une classification générique qui non seulement oppose le style des *commentaires* de César et celui des *réécits analytiques* de Tacite, mais qui, de surcroît, positionne au centre de l'arbre toutes les *œuvres d'inspiration biographique* sans exception, y compris celle due à Tacite, qui se trouve ainsi détachée du reste de l'œuvre de cet auteur (*Vie d'Alexandre* par Quinte-Curce, *Vies des Césars* par Suétone, *Vie d'Agricola* par Tacite).

Entre la figure 6 et la figure 7 on constate comme seules différences le renforcement de la proximité de la *Vie de Tibère* (Suétone) avec le livre 3 de Quinte-Curce et la *Vie d'Agricola* (Tacite) et un regroupement moindre du livre 15 des *Annales* de Tacite avec les autres livres de la même œuvre. Mais seules les structures diffèrent très légèrement ; les distances restent identiques. La pertinence classificatoire de ces deux listes de motifs nous paraît donc équivalente, et excellente dans les deux cas.

5. Conclusion

La notion de motif semble donc prometteuse : fondée épistémologiquement dans une conception topologique des textes, elle permettra de conférer aux textes de nouvelles structures modélisables et fournira des bases de classification stables complétant le décompte de unités simples ; grâce à la plasticité de l'outil qui peut accueillir tous les types d'unités textuelles et dont le paramétrage reste à l'entière liberté du linguiste, il est possible d'envisager des analyses multidimensionnelles à toutes les étapes de la méthode ; enfin, la notion permet de réintégrer dans l'analyse le caractère ordonné des micro-structures textuelles.

Reste cependant à multiplier les applications pour d'une part pouvoir répondre aux questions théoriques en suspens évoquées au début de cet exposé et pour d'autre part trouver quels éléments, quels motifs pourraient faire passer du niveau micro-structurel à l'appréhension de la structure globale des textes. Il faudra sans doute, pour ce faire, travailler sur les emboîtements de motifs et sur les zones transitionnelles qui font passer d'un motif à l'autre.

Références

- Barthélemy J.-P. et Luong X. (1998). Représenter les données textuelles par des arbres. In Brunet É. et Mellet S. (éds), *JADT 1998, Actes des 4èmes Journées Internationales d'analyse de données textuelles*, Univ. de Nice : UMR 6039, 49-70.
- Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In Morin A. et Sébillot P. (éds), *JADT 2002, 6es Journées internationales d'Analyse statistique des Données Textuelles*. Saint-Malo : IRISA / INRIA, Vol. 1 : 403-411.
- Longrée D., Luong X. et Mellet S. (2004). Temps verbaux, axe syntagmatique, topologie textuelle : analyse d'un corpus lemmatisé. In Purnelle G., Fairon C. et Dister A. (éds), *Le poids des mots, Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-la-neuve, Presses universitaires de Louvain, vol. 2 : 743-752.
- Longrée D., Luong X. et Mellet S. (2006). Distance intertextuelle et classement des textes d'après leur structure : méthodes de découpage et analyses arborées. In Viprey J.-M. (éd.), *Actes des 8es Journées internationales d'Analyse statistique des Données Textuelles*, Besançon, Presses universitaires de Franche-Comté, vol. 2 : 643-654.
- Longrée D., Luong X. et Mellet S. (2007). Classification et analyse textuelle : l'approche topologique, communication à la Société Francophone de Classification (Paris, septembre 2007), à paraître dans *Mathématiques et Sciences Humaines*.
- Longrée D. et Mellet S. (2008). Temps verbaux et prose historique latine : à la recherche de nouvelles méthodes d'analyse statistique. In Purnelle G. et Denoos J. (éds), *Ordre des mots et cohérence, Communications présentées au 13e Colloque international de Linguistique latine, 4-9 avril 2005, F.U.S.L (Bruxelles)*, (Publication de la Faculté de Philosophie et Lettres de l'Université de Liège), Liège, Faculté de Philosophie et Lettres, à paraître.
- Luong X., Juillard M., Mellet S. et Longrée D. (2007). Trees and after: The Concept of Text Topology. Some applications to Verb-Form Distribution in Language Corpora, *Literary and Linguistic Computing*, 22 (2): 167-186.
- Mellet S. et Barthélemy J.-P. (2007). La topologie textuelle : légitimation d'une notion émergente, *Lexicometrica 7, Topographie et topologie textuelles* (cf. <http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf>).
- Mayaffre D. (2007). L'analyse des données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle, *Lexicometrica 7, Topographie et topologie textuelles* (cf. <http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf>).