Légitimité d'une unité textométrique : le motif

Sylvie Mellet[1], Dominique Longrée[2]

[1] BCL, CNRS, Université Nice – Sophia Antipolis – mellet@unice.fr
[2] LASLA, Université de Liège et SeSLa, FUSL (Bruxelles) – dominique.Longrée@ulg.ac.be

Abstract

At the JADT 2008, we introduce a complex textometric unit, -called "motif"-, defined by formal criteria and demonstrating at once a characterization function. According to its multidimensional nature, a "motif" includes expressions which present various lexical, morphological and syntactical variations: expressions without any common point in their surface form can therefore be included in the same "motif". This paper will investigate the theoretical and practical legitimacy of such abstractions (which remind us of the old debate between lemma and graphic forms). By studying some examples based on a corpus of Latin texts, we will also evaluate the characterization power of the "motives", both in their multiple forms and in their only abstract schemas.

Résumé

Lors des JADT 2008, nous avons proposé une unité textométrique complexe, – le motif –, définie de manière formelle et manifestant d'emblée une fonction caractérisante. En vertu de ses caractéristiques multi-dimensionnelles un même motif peut regrouper des expressions présentant de multiples variations lexicales, morphologiques et syntaxiques, au point qu'elles peuvent en surface ne plus avoir aucun point commun. Nous poserons la question de la légitimité théorique et pratique de telles abstractions (rappelant le vieux débat entre lemme et formes). A partir d'exemples concrets portant sur un corpus de textes latins, nous évaluerons également les propriétés caractérisantes des motifs tant dans la variété de leurs formes que dans l'unicité de leur schème abstrait.

Mots-clés : motif, formes de surface, schèmes abstraits, unité textométrique, distances intertextuelles.

Il y a 4 ans, lors des JADT 2008, nous avons proposé d'introduire dans l'ensemble des unités textométriques une unité complexe à laquelle nous avons donné le nom de ***motif***. Initialement conçu dans le cadre d'une analyse topologique des textes (Barthélemy & Mellet, 2007 ; Barthélemy, Longrée, Luong & Mellet, 2009), le motif était défini de manière formelle comme « un sous-ensemble ordonné de (E) [(E) désignant l'ensemble textuel], formé par l'association récurrente de n éléments de l'ensemble (E) muni de sa structure linéaire. » (Longrée, Luong & Mellet, 2008). Conjointement lui était d'emblée associée une fonction caractérisante, l'emploi d'un motif, par définition récurrent, étant susceptible de spécifier un type de textes ou de séquences textuelles (*ibid*.). Cependant, ni les implications concrètes de la définition formelle pour l'analyse quantitative des textes, ni le caractère nodal de la définition fonctionnelle ne nous étaient encore apparus clairement. Depuis, comme nous l'avions alors annoncé, nous

avons multiplié les études particulières pour faire fonctionner la notion de motif et pour mieux en appréhender les propriétés[1]. Aujourd'hui nous souhaitons présenter ici un bilan de ces recherches qui attestent de l'intérêt du motif comme nouvelle unité textométrique. Nous voulons aussi introduire une réflexion à la fois théorique et empirique sur le processus d'abstraction qui nous conduit à postuler pour chaque motif l'existence d'un schème unique sous-jacent à des réalisations assez variées – processus qui n'est pas sans rappeler celui de la lemmatisation, et qui donc pose les mêmes questions que celles qui ont dominé le débat entre tenants des lemmes et tenants des formes graphiques.

Nous commencerons par rappeler la genèse du concept de motif et la façon dont sa définition s'est précisée au fur et à mesure que notre recherche progressait. Nous illustrerons ensuite la fonction caractérisante de certains motifs en montrant, à partir d'un corpus de prosateurs latins emprunté à la base de données du LASLA, dans quelle mesure certains motifs sont spécifiques de certains genres littéraires et de certains auteurs. Nous montrerons aussi que les différentes variantes d'un même motif se distribuent elles aussi de manière significative selon les œuvres et les auteurs du corpus et contribuent ainsi à caractériser de manière pertinente et à différencier des modes d'énonciation et des styles différents. Ce qui nous conduira pour finir à comparer le pouvoir discriminant de différents motifs appréhendés et dénombrés sous leurs diverses réalisations concrètes dans les textes *vs* le pouvoir discriminant du même ensemble de motifs appréhendés sous leur forme schématique sous-jacente (donc en regroupant au sein d'une même unité textométrique l'ensemble de ses variantes).

1. Genèse et définition du motif

Nous avons introduit la notion de motif dans le cadre d'une réflexion, menée en collaboration avec Xuan Luong et Jean-Pierre Barthélemy, sur la structure et les propriétés topologiques des textes. Le passage de la lexicométrie à la textométrie obligeait en effet à une analyse approfondie des structures textuelles : il ne suffisait pas de dire que les textes ne sont pas de simples « sacs de mots » ; encore fallait-il définir de manière précise et exploitable les propriétés de leur structure, depuis longtemps pressenties, jamais entièrement formalisées. La première de ces propriétés est évidemment la linéarité ; c'est celle qui a donné lieu aux développements formels les plus avancés : à bien y réfléchir elle sous-tend aussi bien la loi de Zipf que les chaînes de Markov, et, du côté de la linguistique, les principes de la sémantique harrissienne. L'autre propriété, plus récemment mise en évidence, notamment par les travaux de Jean-Marie Viprey, est celle de réticularité ; en réalité cette notion, sans être ainsi dénommée, était déjà sous-jacente dans les premiers travaux de Rastier consacrés aux isotopies. Elle a ensuite été reprise par J.-M. Viprey qui considère notamment que les collocations sont un « aspect décisif de la texture » textuelle (2006 : 73) et, dans le même temps, D. Legallois (2006 : 70) mettait en évidence sa « parfaite congruence avec l'étymologie du mot *texte* ».

Pour notre part, nous avons tenté de modéliser simultanément ces deux propriétés grâce à la notion mathématique de voisinage, allant jusqu'à suggérer qu'un texte répondait aux critères formels d'une structure en treillis. Cet excursus mathématique un peu hasardeux pour

1 Cette recherche a été financée par un programme d'échanges entre le CNRS et le FRS-FNRS, ainsi que par un programme de coopération « PHC - Tournesol » conjoint entre le Ministère français des Affaires Etrangères et Européennes et le WBI (Wallonie-Bruxelles International).

nous, linguistes latinistes, nous a permis de revenir mieux armés vers l'analyse des données textuelles, terrain plus familier, et a mis en lumière l'impérieuse nécessité de concevoir **une nouvelle unité textuelle**, qui intégrant formes graphiques, lemmes, catégories grammaticales, patrons syntaxiques et, éventuellement, schèmes métriques ou prosodiques, permette de traiter à la fois l'imbrication hiérarchique de ces différents niveaux linguistiques, leur association syntagmatique dans la chaîne linéaire des énoncés et leur récurrence structurante au niveau macro-textuel.

Ceci nous a conduit à proposer une première définition formelle du motif qui est la suivante : de manière strictement formelle, un motif se définit par l'association récurrente de n éléments du texte muni de sa structure linéaire (Legallois 2006), laquelle donne une pertinence aux relations de successivité et de contiguïté (Longrée, Luong & Mellet 2008 ; Mellet & Longrée 2009). 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. En effet, comme je viens de le dire, la notion de motif est conçue comme un moyen d'intégrer la multidimensionalité (ou le caractère multi-niveau) de certaines formes récurrentes. Ainsi le motif césarien *dum haec Romae geruntur* se définit par la récurrence d'un schème dont les paramètres sont à la fois syntaxiques, morphologiques et lexicaux : proposition temporelle introduite par la conjonction *dum*, avec verbe à l'indicatif présent, 3ème personne du pluriel, voix passive, forme verbale lexicalement contrainte (généralement *geruntur*, exceptionnellement *parantur*), instanciation de l'actant sujet par un anaphorique neutre pluriel (*haec* ou *ea*), présence d'un circonstant locatif.

Cette multi-dimensionnalité du motif n'est pas nécessairement instanciée : on peut envisager des motifs purement métriques, par exemple. Dans un premier temps, en raison des limites logicielles de repérage des motifs, nous avons travaillé sur des motifs narratifs qui consistaient en de simples séquences de temps verbaux. Mais cette multi-dimensionnalité est néanmoins définitoire et c'est elle qui donne à la notion de motif sa puissance heuristique.

La suite de nos travaux nous a conduits à progresser en intégrant aussi à cette première définition un **aspect fonctionnel**, lié bien sûr à la sémantique du motif, mais aussi à sa propriété de récurrence. Ainsi un motif a une fonction textuelle et discursive, à courte et à longue portée : à courte portée un motif aura, par exemple, une fonction cohésive (*dum haec Romae geruntur*, *quibus rebus cognitis*), une fonction résomptive (*ut supra memoraui*), une fonction conclusive (*quae cum ita sint*), etc. A longue portée, sa récurrence permettra de faire progresser le récit en différents épisodes et de structurer temporellement et spatialement les faits narrés ; ou de souligner les différents moments d'une plaidoirie ou de toute autre argumentation ; sa répétition pourra fournir les jalons mémoriels nécessaires à la bonne réception du discours (en dehors du domaine latin, on pense par exemple ici aux formules caractéristiques de la littérature orale : épopée, chanson de geste, etc.).

En raison de cette articulation forte entre fonctionnalité textuelle et définition formelle, un motif admet la présence de variables en son sein. Au niveau lexical, un des items peut être réalisé par divers lexèmes formant paradigme :

geruntur / parantur

Quibus rebus cognitis / quibus rebus nuntiatis / his rebus nuntiatis

ut supra memoraui / ut antea dixi

Les variations peuvent aussi venir de la permutation de deux éléments :

Haec dum Romae geruntur

et jouer sur la présence *vs* absence d'un élément (opérations de suppression ou d'ajout) :

Haec dum Ø geruntur

His rebus gestis / his rebus ***ita*** *gestis*

Enfin, on peut observer aussi des variations sur les catégories grammaticales :

Quibus rebus nuntiatis / qua re nuntiata

Plusieurs de ces variations peuvent être combinées ; même un motif apparemment très stable peut présenter l'ensemble de ces variantes :

Quae cum ita sint « puisqu'il en est ainsi »,

motif apparemment quasi figé, connaît pourtant quelques variantes, dont celle-ci, qui combine expansion et variation morphologique sur le sujet (*quae*, neutre pluriel, laisse la place à *quae res*, féminin singulier, « cette chose, cette situation »), commutation lexicale et variation morphologique sur le prédicat (*sint*, subjonctif présent du verbe « être » laisse la place au subjonctif imparfait de la locution *se habere* « se présenter ») :

quae cum ita se res haberet « puisque la situation se présentait ainsi »

La variation peut donc être formalisée en quelques règles simples : la principale limite qui lui est imposée est celle de la stabilité du sens et de la fonction textuelle d'autre part. L'imbrication des niveaux, la variation et la fonctionnalité distinguent le motif du simple n-gramme ou du segment répété (Salem 1987). On voit alors que ce qui compte dans le motif, ce qui justifie d'en faire une unité textuelle spécifique, c'est qu'il s'agit d'une cellule organisée et fonctionnelle. Finalement, le motif est donc un « cadre collocationnel » ou une « construction lexico-grammaticale » (Gledhill & Frath 2007) associé à un stock restreint d'éléments fixes et de variables, et auquel on peut attribuer une fonction de marqueur discursif structurant. Désormais nous mettrons entre crochets droits la forme standard du motif, celle que nous choisissons pour le représenter dans toutes ses variantes. Mais cette abstraction n'est pas sans poser question. Est-il légitime de regrouper sous un même schème des formes de surface qui n'ont absolument aucun élément en commun, telles que *ut supra demonstrauimus* et *quos antea dixi* ? Certes, elles apparaissent comme les deux extrêmes d'un continuum de variantes ; mais comment valider l'existence de ce continuum ? C'est de nouveau en prenant appui sur l'une des propriétés fondamentales des motifs en tant qu'unité textométrique – à savoir leur propriété caractérisante – que nous allons tenter de répondre à la question.

2. L'abstraction du motif : perte d'information ou synthétisation structurante ?

C'est sur le pouvoir caractérisant de la distribution des motifs dans le corpus que nous allons évaluer si l'abstraction que fait subir la notion de motif à tout un ensemble de formules correspond à une perte d'information ou, au contraire, à un resserrement structurant.

2.1. Le motif en tant que schème abstrait

Que certains motifs soient spécifiques à certains genres littéraires, à certaines modalités énonciatives ou au style de certains auteurs est un fait déjà relevé dans les études classiques. L'exemple le plus connu en est sans doute celui des « clichés de liaison » mis en évidence par Chausserie-Laprée (1969) chez les historiens latins et auxquels appartiennent certains des exemples déjà donnés ci-dessus : [*quibus rebus cognitis*], [*hac re nuntiata*], [*his rebus ita gestis*], [*dum haec Romae geruntur*]. Inutile, par exemple, de recourir à un lourd attirail statistique pour apprécier la distribution de [*quibus rebus cognitis*] (toutes variantes confondues) :

Cic. (discours et traités)	Cés. (histoire)	Sall. (histoire)	Curt (histoire)	Tac. (histoire)	Sén. (traités)	Pétr. (roman)	Total
14	58	4	8	13	2	3	102

Figure 1 : distribution de [quibus rebus cognitis] (toutes variantes confondues)

Ce motif est propre à l'énonciation narrative historique (83 occurrences sur 102), et au sein de celle-ci, il est la marque plus particulière de l'écriture césarienne (58 occurrences). A l'inverse, le motif [*quae cum ita sint*], lui, ne se rencontre, dans toute la base textuelle gérée par HYPERBASE-Latin, que chez Cicéron, traités et discours.

Mais, par ailleurs, un examen attentif de la distribution des variantes de chaque motif peut être tout aussi intéressant. Ainsi, pour revenir au motif [*quibus rebus cognitis*], dans notre corpus il ne se rencontre sous cette forme précise que chez César : 5 occurrences dans la *Guerre des Gaule*, 5 dans la *Guerre Civile*. Chez tous les autres auteurs, ce sont des variantes, plus ou moins complexes, qui apparaissent. Un tel choix parmi les variantes d'un même motif est-il lui aussi significatif et caractérisant ?

2.2. Les variantes d'un motif : distribution et distances intertextuelles

Pour tester cette hypothèse de la pertinence des variantes de motif en matière de distance intertextuelle, nous commençons par traiter le motif [*ut supra memoraui*]. Nous l'avons retenu parce qu'il présente deux avantages méthodologiques : d'une part il est présent aussi bien en discours qu'en histoire : les éventuelles spécificités de sa distribution ne pourront donc pas se résumer à une simple opposition présence *vs* absence selon le plan d'énonciation et/ou le genre littéraire des œuvres du corpus ; d'autre part, il présente un bon nombre de variantes, mais celles-ci se laissent synthétiser au moyen de quelques règles simples de commutation (au sein d'un paradigme lexical ou au sein du paradigme morphologique de la flexion verbale, voire entre deux structures syntaxiques) et ne nous entraînent pas trop loin sur le terrain plus

mouvant des expansions et des suppressions, dont il n'est pas toujours facile de fixer les limites acceptables et qui, de fil en aiguille, peuvent conduire à construire un ensemble quelque peu hétérogène. Ici seul un adverbe peut ou non s'insérer dans le schéma collocationnel de base qui est donc le suivant :

[pronom relatif ou subordonnant comparatif + adverbe d'antériorité intradiégétique *ante(a)* ou *supra* + verbe déclaratif au passé]

On observe donc d'abord l'alternance entre le relatif (*Rel*) et le subordonnant comparatif (*Comp*), celui-ci pouvant prendre la forme *ut* ou *sicut*, parfois aussi *quemadmodum*. Une autre variation porte sur la présence ou l'absence de l'adverbe anaphorique intradiégétique, ainsi que sur sa forme : on relève deux adverbes en concurrence : *ante(a)* « auparavant » et *supra* « ci-dessus », précédés ou non de *paulo* « peu », parfois de *iam* « déjà » (= *x*). Le lexème verbal est choisi au sein d'un paradigme sémantique : *dico* « dire », *demonstro* « démontrer, argumenter, présenter », *memoro* « rappeler » ; plus rarement : *commemoro* « rappeler », *doceo* ou *expono* « exposer », *refero* « rapporter », *loquor* « parler », *nomino* « mentionner », *ostendo* « montrer ». Le temps verbal est très majoritairement le parfait ; mais on a quelques occurrences d'imparfait (*ut paulo ante dicebam*, Tacite, *de Oratoribus*) et de plus-que-parfait (*quos paulo ante commemorare coeperam*, Cicéron, *Discours*). Enfin, la personne verbale connaît aussi des variations intéressantes : à côté de la première personne du singulier (*1s*), on trouve aussi la première du pluriel (*1p*) et la troisième personne du passif impersonnel (*P*) (« comme il a été montré plus haut »).

Comment se distribuent ces différentes variantes chez les prosateurs latins ? Et sont-elles à même de caractériser suffisamment le style de nos auteurs pour permettre de calculer une distance intertextuelle ?

Pour répondre à ces questions, nous avons dénombré dans le corpus que constitue la base textuelle d'HYPERBASE-Latin toutes les occurrences de chaque variante du motif étudié et commencé par éliminer les textes qui ne présentaient pas un nombre d'occurrences suffisant pour autoriser les calculs statistiques. Cela nous a permis de conserver les *Discours* de Cicéron, les *Traités* de Cicéron et les œuvres des quatre historiens : la *Guerre des Gaules* et la *Guerre Civile* de César, l'œuvre complète de Salluste, celle de Quinte-Curce, les *Annales* et les *Histoires* de Tacite. Du côté des variantes, nous avons aussi supprimé les plus rares et opéré quelques regroupements qui nous ont paru *a priori* acceptables (*ante* avec *antea*, *ut* avec *sicut*) : ce qui nous a conduit à conserver la liste suivante :

Comp_x_ante_1s_dico
Comp_x_ante_P_dico
Comp_supra_1s_dico
Comp_supra_P_dico
Comp_supra_1p_demonstro
Comp_supra_P_demonstro
Rel_x_ante_1s_dico
Rel_x_ante_P_dico
Rel_x_ante_1s_memoro
Rel_supra_1p_dico
Rel_supra_1p_demonstro

Nous avons ensuite construit la matrice dans laquelle chaque œuvre est caractérisée par la fréquence d'emploi de chacune des formes du motif, et dans laquelle, simultanément, chaque

forme du motif est caractérisée par sa distribution à travers les œuvres du corpus. Le calcul de distances a été fait grâce au logiciel ANAR associé à Hyperbase, qui transforme le tableau de contingences en un tableau d'écarts réduits, et la représentation choisie est celle de l'AFC. Par ailleurs, le nombre de paramètres de variation au sein du motif étant assez important, nous avons avancé progressivement dans l'analyse et nous présentons donc les résultats par étapes.

Nous avons d'abord neutralisé la variation de personne, pour nous concentrer sur les alternances lexicales conjointes du verbe et de l'adverbe. On obtient l'AFC suivante :

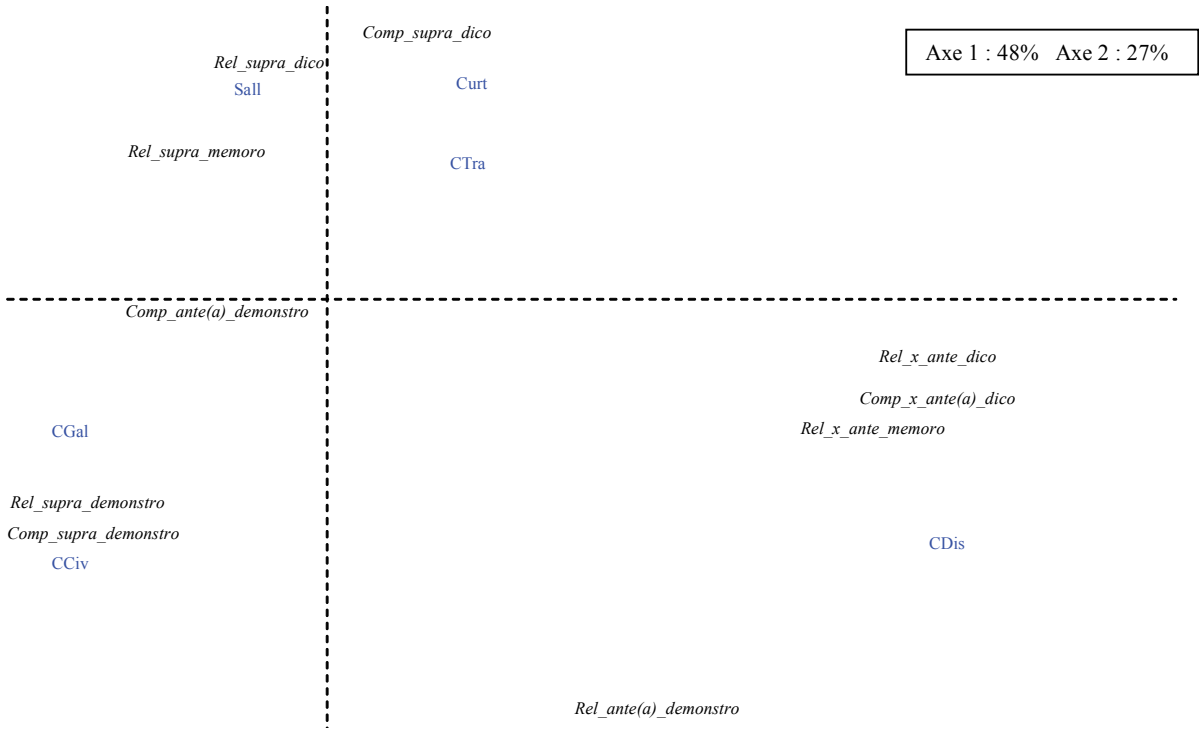

Figure 2 : AFC sur les variantes du motif [ut supra memoraui], avec neutralisation de la personne verbale. Axes 1 et 2.

On voit ici que l'axe horizontal du graphique (axe 1 qui exprime 48% de l'information de la matrice) oppose les formes du motif contenant *ante(a) dico*, *ante(a) memoro* dans les *Discours* de Cicéron[2] aux formes contenant *supra demonstro* dans les deux *Commentaires* de César. Cette distribution était fortement prévisible pour l'adverbe : *ante* (« auparavant ») réfère au temps de la parole vivante, énoncée, tandis que *supra* (« ci-dessus ») réfère à l'agencement spatial d'un texte rédigé et matériellement inscrit sur un support. En ce qui concerne le verbe, l'emploi préférentiel de « dire » en discours est aussi très banal ; on pourrait en revanche s'étonner de voir un historien prétendre « démontrer » quelque chose au lieu de le raconter. Mais on reconnaît bien là le but à peine déguisé de César, qui n'est pas de produire une histoire objective de la guerre des Gaules ou de la guerre civile, mais bien d'argumenter un commentaire à visée apologétique. C'est d'ailleurs ce qui distingue profondément César des autres historiens, comme le montre l'axe 2 de l'AFC (opposition le long de l'axe vertical) qui regroupe les motifs bâtis

2 Le symbole x dans les schèmes représentant les diverses variantes du motif correspond à la présence possible de l'adverbe *paulo* « peu », éventuellement à celle de *iam* « déjà ».

autour de *demonstro* dans le quadrant inférieur gauche, alors que la partie supérieure accueille les lexèmes *dico* et *memoro*. L'emploi de l'adverbe *supra* est, lui, commun à tous les historiens : c'est donc bien, au-delà des lexèmes isolés, l'unité phraséologique qui caractérise l'écriture de chacun des auteurs. Mais elle caractérise aussi des genres, car on notera, dans cette AFC, l'assez grand éloignement entre les *Discours* et les *Traités* de Cicéron.

Si l'on neutralise au contraire ces variations lexicales pour se concentrer sur la variation morphologique de la personne verbale, on obtient l'AFC suivante (Nous avons représenté ici les axes 1 et 3 car la distribution sur l'axe 2 est écrasée par le seul poids de la variante [comparatif + *supra* + verbe à la 1pers du sg.] qui caractérise les *Traités* de Cicéron) :

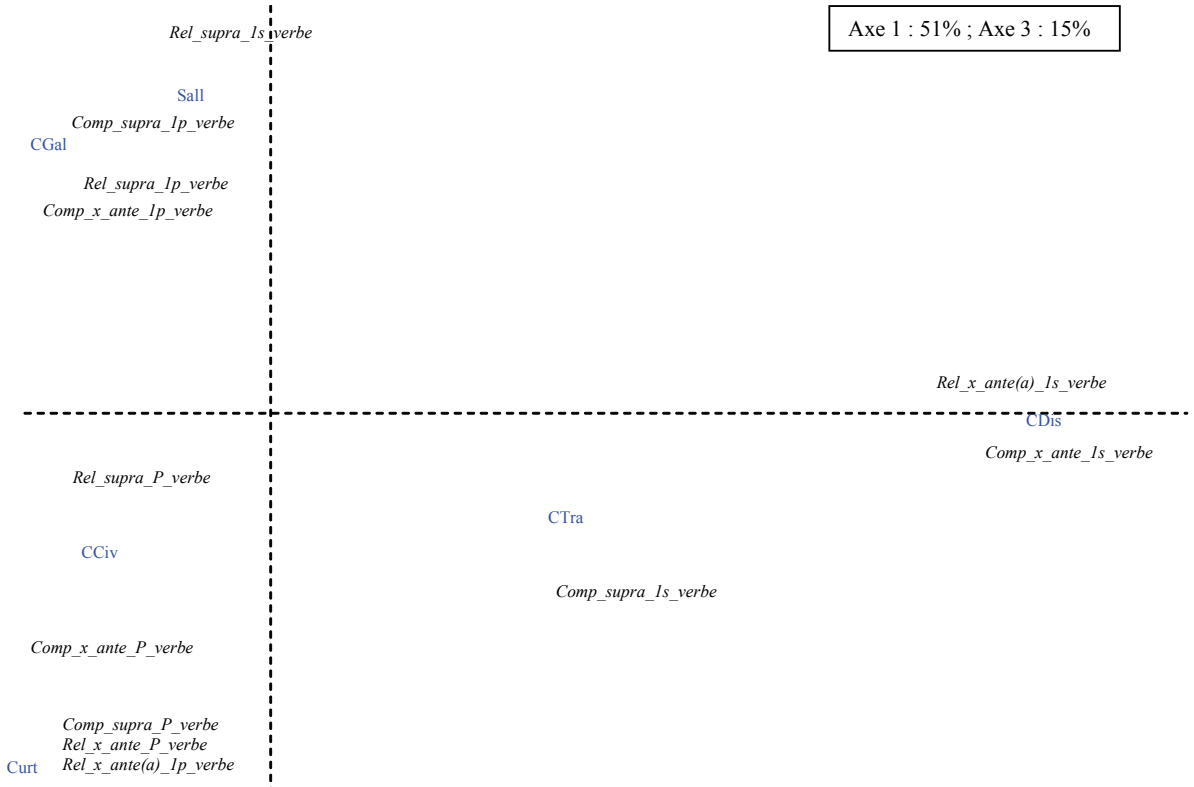

Figure 3 : AFC sur les variantes du motif [ut supra memoraui], avec neutralisation du lexème verbal. Axes 1 et 3.

L'opposition entre les plans d'énonciation histoire *vs* discours est flagrante : les œuvres de Cicéron occupent la partie droite du graphe, distribuées exclusivement autour de formes verbales à la première personne du singulier ; sans être très proches ces deux œuvres d'un même auteur manifestent donc leur parenté quant aux plans d'énonciation benvenistiens. La partie gauche de l'axe 1 est plus hétéroclite mais le troisième facteur d'analyse (qui représente encore 15% de l'information) fait apparaître un paramètre de distribution intéressant : le positionnement dans le quadrant inférieur gauche de toutes les formes de passif impersonnel, en grande partie regroupées à proximité de l'œuvre de Quinte-Curce. A noter en revanche qu'avec ce paramétrage, les deux œuvres de César s'éloignent sensiblement l'une de l'autre.

Reste à synthétiser tous ces paramètres en un seul et même graphe de distances intertextuelles. C'est ce que fait la figure 4 (voir également Longrée et Mellet 2012, où l'on retrouve des regroupements attendus :

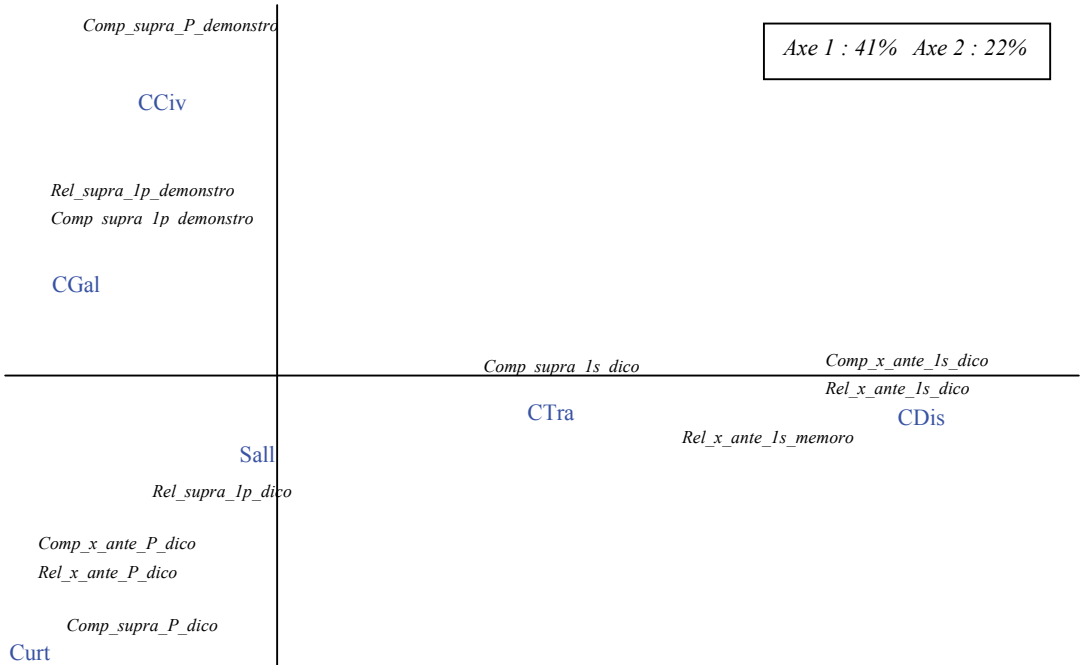

Figure 4 : AFC sur les variantes du motif [ut supra memoraui], dont la fréquence est supérieure à 5 ; axes 1 et 2

L'opposition entre les plans d'énonciation histoire *vs* discours déjà pointée à plusieurs reprises reste flagrante : les œuvres de Cicéron occupent toujours la partie droite du graphe, distribuées exclusivement autour de formes du motif contenant la première personne du singulier du verbe *dico* « dire » et se distanciant nettement des œuvres historiques qui utilisent des formes organisées autour soit de la première personne du pluriel, soit du passif impersonnel. Cette opposition énonciative coïncide donc logiquement avec une opposition générique. Toujours le long de l'axe 1 (qui exprime 41% de l'information de la matrice) se dessine ensuite la différence plus subtile entre *Discours* et *Traités*, les premiers donnant la préférence à l'adverbe *ante*, les seconds à l'adverbe *supra*. Comme on l'a dit, cette distribution est, elle aussi, interprétable puisqu'elle correspond logiquement aux deux deixis énonciatives, temporelle et spatiale, étroitement associées au genre discursif de chacune des œuvres. Dans la partie gauche du graphique sont regroupés tous les historiens, qui se différencient les uns des autres par d'autres paramètres : les deux œuvres de César sont de nouveau regroupées dans le quadrant supérieur gauche, affichant leur proximité caractéristique avec les variantes du motif bâties sur le verbe *demonstro*. Ce paramètre est donc plus lourd que la variante sur les personnes verbales qui avait séparé *Guerre Civile* et *Guerre des Gaules* dans l'AFC n°2. Le quadrant inférieur gauche du graphe regroupe avant tout les variantes qui comportent la forme de passif impersonnel du verbe *dico* (*ut supra dictum est*) pour lesquelles Quinte-Curce semble avoir un goût marqué ; Salluste est en position plus centrale et peu caractérisée. La représentation des axes 2 et 3 permet, elle, d'afficher de manière claire la distance entre les *Discours* et les Traités de Cicéron, pointée dans les deux premières AFC et un peu neutralisée dans la figure précédente.

La preuve semble donc faite de la propriété caractérisante des motifs, aussi bien lorsqu'ils sont appréhendés globalement dans l'ensemble de leurs réalisations (schème abstrait, du type [*quibus rebus cognitis*]) que lorsqu'ils sont au contraire saisis dans leur variation (variantes du motif [*ut supra memoraui*]). La conclusion est-elle extensible à tous les types de motifs ? L'hypothèse que nous formons et à laquelle nous allons apporter pour terminer un premier support empirique, est que les variantes devraient être plus discriminantes lorsque le motif répond à toutes les propriétés formelles du motif, notamment sa forte multi-dimensionnalité, alliée à un sémantisme stable et une fonctionnalité discursive et textuelle marquée, et moins discriminantes lorsque le motif est moins riche, plus lexicalisé, plus proche du segment phraséologique.

2.3. Schème abstrait et variantes dans le cas des segments de type phraséologiques

Nous comparons ici le pouvoir distinctif et classificatoire de la distribution de divers syntagmes circonstanciels de temps, récurrents dans le corpus. Il s'agit des segments suivants :

A = Interpositis X diebus – des jours ayant passé

B = eodem_X_tempore – au même moment

C = isdem_X_temporibus – aux mêmes moments, en ces mêmes temps

D = hoc_X_tempore – à ce moment

E = his_X_temporibus – en ces temps

F = illo_X_tempore - à ce moment-là, à cette époque

G = eo_X_tempore – à ce moment

H = eis _temporibus – en ces temps

I = per_idem_tempus – au même moment

J = sub_idem_tempus – au même moment

K = paucis_multis_ante_diebus_mensibus – peu/beaucoup de jours/mois avant

L = paucis_post_diebus_mensibus – peu de jours/mois après

M = ex_Démonstratif_die – depuis ce jour

N = in_Démonstratif_die - dans ce jour

O = in_Démonstratif_diem - en ce jour

P = in_Dém_diebus - dans ces jours

Q = in_Dém_dies - en ces jours

R = ante_Dém_diem – avant ce jour

S = post_Dém_diem – après ce jour

T = ad_Dém_diem – jusqu'à ce jour

U = per_Dém_dies – à travers ce jour

V = per_dies_Numéral – pendant X jours

Plusieurs schémas formels communs permettent certes de regrouper ces syntagmes : [participe ou adjectif + adverbe *post* ou *ante* + ablatif des noms *dies* ou *mensis*]; [préposition + démonstratif + forme fléchie à l'ablatif ou à l'accusatif des noms *dies* ou *tempus*]; [démonstratif ou *idem* à l'ablatif + ablatif des noms *dies* ou *tempus*], le principal de ces schémas étant le très englobant [préposition + déterminant anaphorique ou déictique + substantif à signifié temporel]. En dépit de ce schème morpho-syntaxique commun, on peut toutefois se demander si l'on a bien affaire là à un motif au sens strict, vu la diversité de sens des variations que l'on peut rencontrer pour chaque schème ; le choix entre les différents syntagmes semble devoir être motivé par le sens propre à chacun d'eux (« en ce jour » *vs* « depuis ce jour » *vs* « pendant tous ces jours », par exemple) et non par un libre choix stylistique de l'auteur. Dès lors ces variations ne relèveraient plus du libre jeu d'écriture à partir du feuilleté complexe d'un motif.

Pour répondre à cette question, nous comparerons deux AFC : l'une portant sur la distribution dans le corpus de tous les syntagmes circonstanciels de temps considérés individuellement, l'autre opérant sur leur regroupement en sept sous-classes effectué sur la base des schèmes sous-jacents les plus cohérents. Dans les deux AFC, le poids de Plaute sur l'axe 2, lié à son emploi abondant et spécifique du syntagme *ante*+Dém+*diem*, nous amène à comparer uniquement ici les représentations des axes 1 et 3.

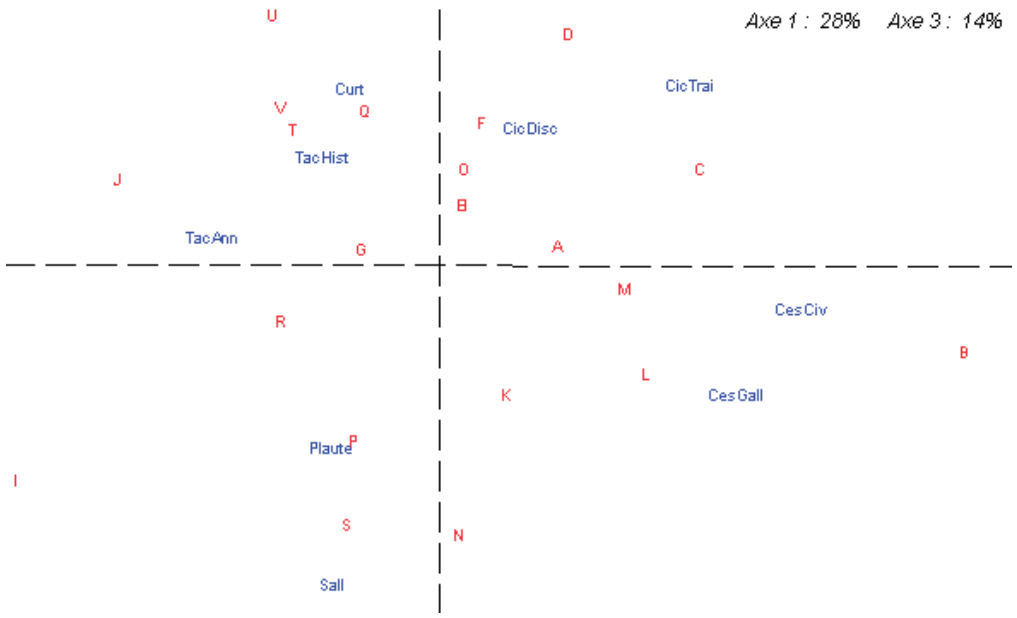

Figure 5 : AFC prenant en compte l'ensemble des syntagmes temporels listés, axes 1 et 3

Dans cette première AFC prenant en compte chacun des syntagmes précédemment listés, l'axe 1 semble opposer la langue très classique de Cicéron et César à celle de tous les autres textes ; l'axe 3, quant à lui, permet de séparer plus clairement César et Cicéron.

Qu'en est-il quand on opère des regroupements entre les variantes de ces syntagmes sous quelques schèmes abstraits plus englobants ? La liste de ces schèmes est la suivante :

a = interpositis+ paucis-multis-ante-post-diebus-mensibus

b = eodem-isdem-illo-illis-eo-eis-X-tempore-temporibus

c = hoc-his-X-tempore-temporibus

d = per-sub-idem tempus

e = ad-ex-in-DEM-die/diem)

f = in-per-DEM-dies, per-dies-NUM

g = ante-post-DEM-NUM-diem

L'AFC portant sur ces schèmes met en évidence les mêmes oppositions : d'une part, sur l'axe 1, entre les œuvres de Cicéron et César et celles de tous les autres auteurs, d'autre part sur l'axe 3, entre César et Cicéron. Et l'on retrouve aussi la même proximité étonnante entre Plaute et Salluste qu'il conviendrait d'analyser de plus près.

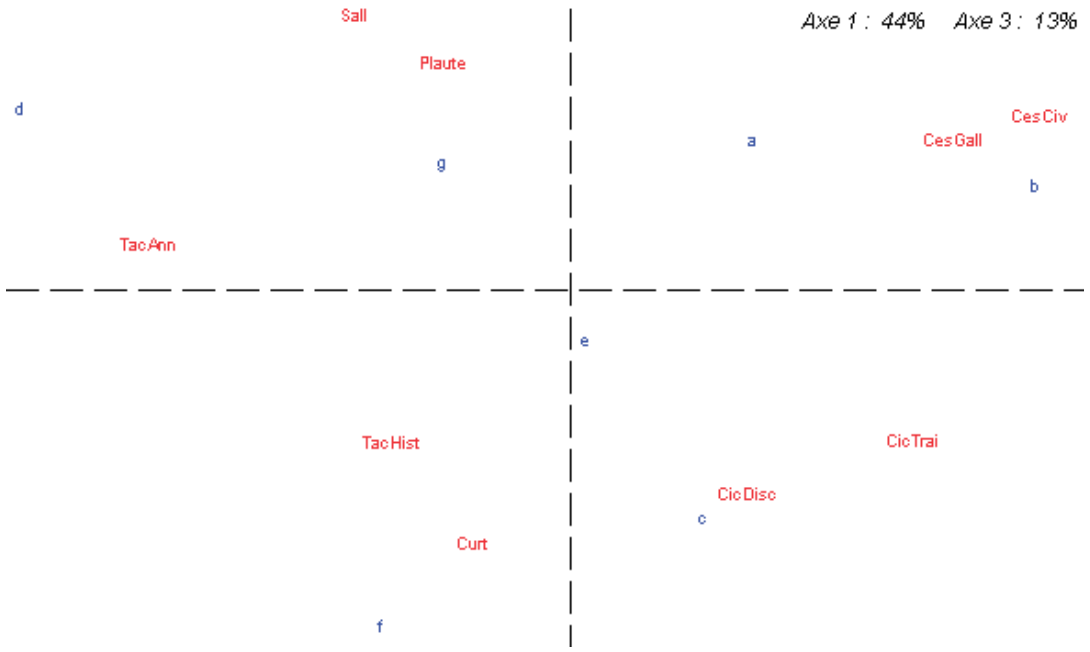

Figure 6 : AFC prenant en compte un regroupement des syntagmes temporels listés en sept sous-classes, axes 1 et 3

On constate ici qu'un calcul de distance fondé sur l'ensemble des formes attestées ne se revèle guère plus instructive que celui portant sur des regroupements effectués sur la base d'une abstraction schématique. Le phénomène s'explique sans doute largement par le fait qu'il n'y a pas ici de véritable unité textométrique sous-jacente à l'ensemble des formes observées : les différents syntagmes sont trop porteurs d'un sens propre, ce ne sont pas des variantes de motifs *stricto sensu*.

3. Conclusion

Cette première étude comparée sur l'impact du regroupement des variantes au sein d'une même unité textométrique appelée motif nous semble confirmer un point fondamental : pour qu'il y ait motif, il faut qu'il y ait une véritable structure multi-dimensionnelle dont la stabilité est assurée par l'unité de sens et de fonctionnalité. La structure est complexe, elle engage plusieurs niveaux linguistiques, mais elle forme une véritable unité. Ceci explique qu'elle puisse à la fois, en tant que telle, être un paramètre discriminant de la distinction entre des textes de genres différents (il s'agit d'une véritable unité textométrique), et à la fois accueillir des variations régulées qui offre à chaque écrivain un espace de liberté avec lequel il peut jouer pour affirmer sa personnalité stylistique. Ces variations peuvent être extrêmement importantes en surface sans rompre avec l'identité intrinsèque du schéma de base. L'abstraction qui consiste à travailler sur le schème du motif fait donc perdre la perception des différences subtiles d'écriture entre auteurs ; en revanche elle reste légitime lorsqu'il s'agit d'appréhender des caractéristiques transversales plus larges telles que les oppositions génériques souvent liées à des différences de positionnement énonciatif.

En revanche, avec le second type de syntagmes étudiés ici, qui rassemblent divers circonstanciels de temps qui ont pour seul point commun cette fonction syntaxique et sémantique large, mais qui ne constituent pas à proprement parler des variantes d'un schème motival, ce va-et-vient entre prise en compte de toutes les variantes et regroupement en quelques sous-classes trouve très vite ses limites. Cela s'explique aisément : comme il n'y a pas véritablement de schème abstrait sous-jacent constituant une unité textométrique indéniable, d'une part le regroupement des différents syntagmes risque d'être peu pertinent et peut aboutir à des résultats peu interprétables, d'autre part les différents syntagmes relevés ne sont pas à proprement parler des variantes ; ce sont des expressions de sens différent (comme « en ces jours-là » *vs* « jusqu'à ce jour ») auxquelles chaque auteur peut avoir besoin de recourir : elles n'ont donc aucun pouvoir caractérisant. On retrouve ici les difficultés que l'on peut avoir lorsqu'il s'agit de lemmatiser : rapporter *boni* « les gens de biens » ou *bona* « les biens (matériels) » au lemme *bonus,a,um* « bon » pose le même type de problème. Plus largement, cette différence de comportement pose la question du statut cognitif du motif et des modalités de sa mémorisation : un travail est en cours en collaboration avec des psycologues.

La possibilité d'opérer une abstraction schématique (de type lemmatisation) avec des résultats pertinents pour la caractérisation et la classification des textes paraît donc être une pierre de touche intéressante pour la reconnaissance des motifs comme unités textométriques complexes.

Références

Barthelemy J.-P., Longrée D., Luong X. & Mellet S. (2009). Représentation du texte pour la classification arborée et l'analyse automatique de corpus : application à un corpus d'historiens latins, *Mathematics and Social Sciences* 187 (3) : 107-121.

Barthélemy J.-P et Mellet S. et. (2007). La topologie textuelle : légitimation d'une notion émergente, *Lexicometrica*. Consultable en ligne à l'adresse : http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/mellet.pdf

Chausserie_Laprée J.-P. (1969). *L'expression narrative chez les historiens latins, Histoire d'un style*, Paris : E. de Boccard.

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

Legallois D. (2006). Des phrases entre elles à l'unité réticulaire de textes, *Langages* 163 : 56-70.

Longrée D., Luong X. et Mellet S. (2008). Les motifs : un outil pour la caractérisation topologique des textes.In S. Heiden et B. Pincemin (éds), *JADT 2008, Actes des 9èmes Journées internationales d'Analyse statistique des Données Textuelles*, vol. 2, Lyon : Presses universitaires de Lyon : 733-744. Consultable en ligne à l'adresse : http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/Longrée-luong-mellet.pdf

Longrée D.et Mellet S. (2012). Le motif : une unité phraséologique englobante ? Etendre le champ de la phraséologie de la langue au discours, *Langages*, à paraître

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

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*, Paris : Klincksiek.

Viprey J.M. (2006). Structure non-séquentielle des textes, *Langages* 163 : 71-85.