CORPUS

Corpus

2 | 2003
La distance intertextuelle

Temps verbaux et linéarité du texte : recherches sur les distances dans un corpus de textes latins lemmatisés

Dominique Longrée et Xuan Luong

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/33
DOI : 10.4000/corpus.33
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 décembre 2003
ISSN : 1638-9808

Référence électronique
Dominique Longrée et Xuan Luong, « Temps verbaux et linéarité du texte : recherches sur les distances dans un corpus de textes latins lemmatisés », *Corpus* [En ligne], 2 | 2003, mis en ligne le 15 décembre 2004, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/33 ; DOI : https://doi.org/10.4000/corpus.33

Ce document a été généré automatiquement le 8 septembre 2020.

© Tous droits réservés

Temps verbaux et linéarité du texte : recherches sur les distances dans un corpus de textes latins lemmatisés

Dominique Longrée et Xuan Luong

1 La recherche de critères statistiques permettant d'évaluer des distances linguistiques ou stylistiques entre des textes n'en est plus à ses balbutiements en matière de littérature latine. Diverses études ont mis en évidence les perspectives qu'une telle recherche pouvait offrir en matière d'attribution à un auteur, d'évolution diachronique ou de distinction entre genres littéraires[1]. Les facteurs étudiés ont été de nature assez variées : lexique, parties du discours et catégories grammaticales, constructions syntaxiques... Souvent, néanmoins, ces travaux statistiques ont montré leurs limites, notamment en ce qui concerne les attributions de textes à un auteur[2].

2 Un autre domaine où la recherche s'avère particulièrement délicate est celui de la répartition générique des œuvres. Sur base de critères linguistiques, il est en effet aisé d'opposer des genres clairement distincts : prose et poésie, rhétorique et histoire, comédie et épopée, etc. En revanche la chose se complique singulièrement lorsque l'on tente de reconnaître des distances entre des textes de genres assez proches : poésie lyrique et poésie didactique, discours politiques et discours judiciaires... Souvent, on hésite, dans ce cas, sur le poids respectifs de chacun des facteurs pouvant contribuer à expliquer les distances relevées entre les textes : évolution diachronique entre des textes d'époques différentes, spécificités d'écriture propres à chaque auteur ou, finalement, nature du texte lui-même. Il en va ainsi pour les textes de natures assez diverses qui composent le corpus historique latin. La recherche que nous menons a pour objectif de découvrir des critères efficaces de reconnaissance et de classification de ces textes par auteurs et, surtout, par sous-genres. Les catégories des temps et modes verbaux ont semblé pouvoir constituer un de ces critères. Nous présenterons ici les outils que nous avons développé pour ce faire : l'originalité de la méthode réside

dans une prise en compte non seulement des fréquences des unités d'analyse, mais aussi de leurs localisations dans la chaîne linéaire que constitue le texte.

1. Spécificités du corpus historique

3 Le corpus des textes historiques latins présente un champ d'investigation tout à fait privilégié par rapport à la question qui nous préoccupe. On dispose en effet d'un ensemble d'œuvres qui, toutes, ont comme but avoué de raconter des événements historiques. Ces œuvres relèvent cependant d'intentions littéraires souvent fort différentes. Sous le terme de « prose historique », les latinistes regroupent en effet des œuvres assez diverses :

- 1. des commentaires militaires, comme la *Guerre des Gaules* ou la *Guerre civile* de César, ou comme les œuvres de ses continuateurs (8e livre de la *Guerre des Gaules*, *Guerre d'Espagne*, *Guerre d'Afrique*, *Guerre d'Alexandrie*),
- 2. des textes s'inscrivant dans la tradition de l'histoire annalistique, comme l'œuvre de Tite-Live ou les *Histoires* et les *Annales* de Tacite,
- 3. des monographies historiques comme la *Conjuration de Catilina* ou la *Guerre de Jugurtha* de Salluste,
- 4. des œuvres biographiques comme celle de Cornelius Nepos ou comme les *Vies de 12 Césars* de Suétone,
- 5. et enfin des œuvres quelque peu hybrides, comme la *Vie d'Agricola* de Tacite – œuvre qui combine éloge funèbre, biographie et récit historique – ou comme l'*Histoire d'Alexandre le Grand* de Quinte-Curce, monographie historique où, de par le sujet même, la composante biographique est bien présente.

4 Par delà cette diversité des œuvres, les latinistes opèrent néanmoins deux regroupements majeurs, en opposant les œuvres à proprement parler « historiques » aux œuvres « biographiques ». Des différences manifestes apparaissent en effet tant sur le plan des conceptions littéraires ou historiographiques que dans la structure même des deux types d'œuvres : une des oppositions les plus nettes se situe dans l'emploi des discours rapportés, fréquents chez les historiens, rarissimes et remplacés par de brèves citations dans les œuvres biographiques. Rares sont toutefois les études tendant à dégager les caractéristiques linguistiques et stylistiques propres aux divers types de textes composant le corpus historique[3]. On peut dès lors se demander dans quelle mesure une bipartition du corpus entre histoire et biographie est ou non confirmée par l'emploi des modes et temps verbaux.

5 La recherche en la matière est certes compliquée par le peu d'œuvres qui nous sont parvenues pour la période classique : ainsi, toutes les œuvres conservées qui relèvent du commentaire militaire tournent autour de César, et, si l'on fait abstraction de fragments, l'histoire de type annalistique n'est plus représentée que par deux auteurs, séparés par plus d'un siècle. Seules les œuvres de deux biographes sont arrivées jusqu'à nous. Il est donc particulièrement délicat de faire la part de ce qui revient à des variations génériques, par delà les différences liées à l'époque ou à l'auteur.

2. Nature et structure du corpus informatisé

6 Malgré les difficultés que nous venons de signaler, la question des distances entre textes historiques latins peut être réexaminée aujourd'hui avec l'aide de nouveaux moyens d'investigation. On a vu en effet s'élargir de manière très sensible, durant ces dernières années, l'important corpus informatisé et lemmatisé que le Laboratoire d'Analyse Statistique des Langues anciennes (L.A.S.L.A.) de l'Université de Liège développe depuis sa fondation en novembre 1961. Ce corpus présente un avantage

capital : chaque forme du texte est non seulement associée à son lemme, -c'est-à-dire à une étiquette identifiant le lexème auquel elle appartient-, mais est également liée à un codage morphosyntaxique[4]. On dispose ainsi, pour bon nombre de textes, de fichiers où, à chaque forme, se trouvent rattachés un lemme, une référence textuelle, un numéro d'ordre dans le texte, une analyse morphologique complète, à laquelle, s'ajoute un codage de la subordination des verbes[5]. Voici un extrait d'un de ces fichiers (César, *Guerre des Gaules*, livre 7, 1, 1: *Quieta Gallia, Caesar, ut constituerat, in Italiam ad conuentus agendos profisciscitur*) :

&QVIETVS	quieta	001000100100121F00-2AE	007 34028
&GALLIA	NGallia	001000100200211F00	007 34029
&CAESAR	NCaesar	001000100300313A00	007 34030
&VT	1ut	001000100400466015	007 34031
&CONSTITVO	constituerat	001000100500553C**15-XD**	007 34032
&IN	in	001000100600670300	007 34033
&ITALIA	NItaliam	001000100700711C00	007 34034
&AD	2ad	001000100800870300	007 34035
&CONVENTVS	conuentus	001000100900914L00	007 34036
&AGO	agendos	00100010100105CL50 4	007 34037
&PROFICISCOR	proficiscitur	00100010110115LC**11 &** 2	007 34038

7 De tels fichiers sont particulièrement précieux lorsqu'il s'agit de repérer dans un textes toutes les occurrences d'un même temps verbal : on notera ici les codes en gras (15-XD et 11 &) qui permettent de repérer automatiquement les mode, temps et fonction syntaxique (prédicat de subordonnée ou de principale) des verbes *constituerat* et *proficiscitur*. Un problème auquel on reste toutefois confronté est que la banque de données textuelles du L.A.S.L.A ne comprend pas encore l'ensemble des textes historiques conservés. Seule une des deux œuvres clairement annalistiques a été encodée, celle de Tacite. L'absence de l'œuvre de Tite-Live ne permet pas pour l'instant de se prononcer d'un point de vue statistique sur les affinités linguistiques et stylistiques que l'on pourrait rencontrer dans ces œuvres appartenant au même sous-genre. Du côté de la biographie, l'œuvre de Suétone n'est encore que partiellement accessible et l'analyse de Cornelius Nepos est en cours. La comparaison entre histoire et biographie ne reposera donc pour l'instant que sur des extraits de Suétone. Cependant les fichiers dont nous disposons semblent bien permettre dès à présent une première évaluation des distances entre œuvres et de leur regroupement par auteur, époque et sous-genre. Voici la liste des œuvres sur lesquelles nous avons pu travailler, classées par ordre chronologique et suivies des abréviations qui seront utilisées dans la suite de l'étude :

- César, *Guerre des Gaules*, livres 1-7 (Gall1-7)
- César, *Guerre civile*, livres 1-3 (Civ1-3)
- *Guerre des Gaules*, livre 8 (Gall8)
- *Guerre d'Alexandrie* (Alex)
- *Guerre d'Afrique* (Afri)

- *Guerre d'Espagne* (Hisp)
- Salluste, *Conjuration de Catilina* (Cat)
- Salluste, *Guerre deJugurtha* (Jug)
- Quinte Curce, *Histoire d'Alexandre*, livres 3-10 (QC3-10)
- Tacite, *Vie d'Agricola* (Agri)
- Tacite, *Germanie* (Germ)
- Tacite, *Dialogue des Orateurs* (Orat)
- Tacite, *Histoires*, livres 1-5 (Hist1-5)
- Tacite, *Annales*, livres 1-6 (Ann1-6)
- Tacite, *Annales*, livres 11-16 (Ann11-16)
- Suétone, *Vie de Jules César* (Iul)
- Suétone, *Vie d'Auguste* (Aug)
- Suétone, *Vie de Tibère* (Tibe)
- Suétone, *Vie de Galba* (Galb)
- Suétone, *Vie d'Othon* (Otho)
- Suétone, *Vie de Vitellius* (Vite)
- Suétone, *Vie de Titus* (Titus)
- Suétone, *Vie de Vespasien* (Vespa)
- Suétone, *Vie de Domitien* (Domi)

8 Dans ce corpus, ont été intégrées deux œuvres, la *Germanie* et le *Dialogue des Orateur*s, qui ne relèvent pas du genre historique : la première est un traité géographique et la seconde, un traité rhétorique. Celles-ci sont toutefois dues à l'auteur des *Histoires* et des *Annales*. Leur prise en compte, du moins dans un premier temps, aidera à mieux évaluer les poids respectifs des facteurs de distance que constituent le genre et l'auteur.

3. Étude des fréquences

9 Une méthode largement éprouvée déjà pour mesurer des distances entre textes consiste à considérer les unités d'analyses, – ici, les formes verbales –, d'une manière globale, sur base de leur fréquence dans le texte. Pour relever les occurrences en question dans les fichiers du L.A.S.L.A., nous avons fait appel à deux logiciels conçus dans le cadre de l'UMR 6039 « Bases, Corpus et Langage » (CNRS - Nice) : le premier, Estela, a été spécialement conçu comme un logiciel de consultation et d'extraction ; le second n'est autre qu'une nouvelle version du programme Hyperbase d'Étienne Brunet, adaptée à l'exploitation des codages grammaticaux propres aux textes latins[6]. Nous avons ainsi obtenus des dénombrements qui permettent de constituer des tableaux de contingence à partir duquel des AFC[7] ont permis de représenter des similarités entre profils textuels. Après avoir mis en évidence des écarts significatifs dans les proportions globales de chaque temps selon l'auteur, l'époque ou le genre, il a paru intéressant d'affiner l'analyse en examinant la distribution des temps entre les principales et les subordonnées[8]. La figure 1 montre les difficultés que l'on peut rencontrer dans l'interprétation des AFC obtenues.

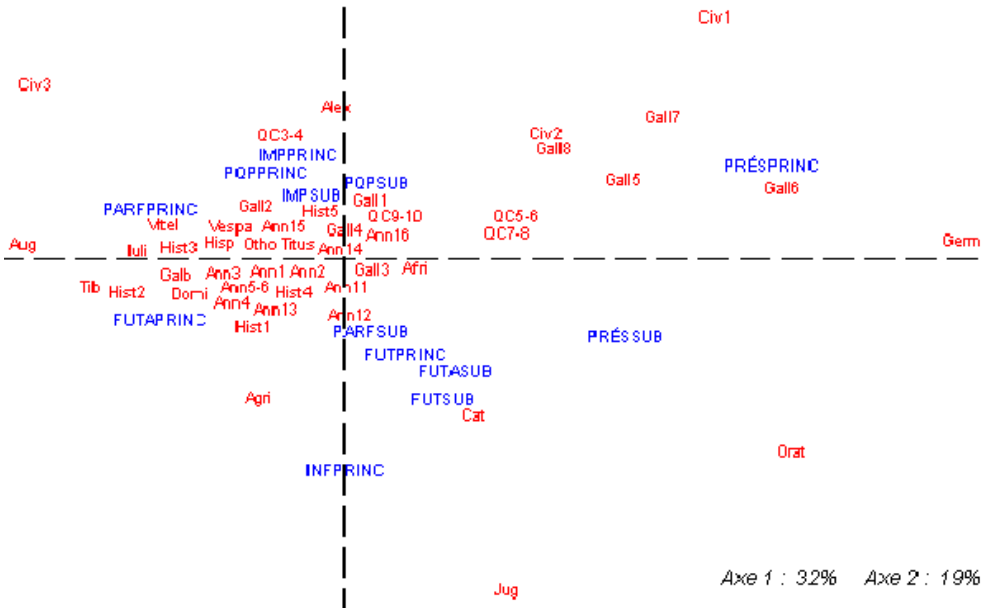

Figure 1

10 On a ici pris en compte l'ensemble des temps de l'indicatif, en distinguant les occurrences en principales et en subordonnées et on a ajouté l'infinitif de narration, une forme qui intervient en concurrence avec l'indicatif en proposition principale.

11 On voit d'emblée que le premier facteur a un poids relativement limité, ce qui montre bien la complexité des phénomènes entrant en ligne de compte. Ce premier facteur semble, de prime abord, lié à une évolution diachronique : une bonne partie des œuvres de la fin de l'époque républicaine (1er s. avant J.C.) se trouve à la droite du tableau, alors que les œuvres de Tacite et Suétone (fin 1er - début 2ème siècle après J.C.), dans leur grande majorité, se rassemblent à gauche. On remarque toutefois des exceptions à cette répartition : Gall.2 et Civ.3 à gauche ; Germ. et Orat. à droite.

12 Ces deux dernières exceptions sont particulièrement intéressantes par rapport à la question du genre : dues à Tacite, elles s'opposent clairement au reste de son œuvre; elles s'écartent aussi très sensiblement de toutes les autres œuvres historiques ; le phénomène s'explique par l'emploi des futurs dans le *Dialogue* et l'abondance des présents dans la *Germanie*; le genre littéraire semble donc bien primer ici sur l'auteur ou l'époque. En revanche, cette première AFC ne permet pas d'attribuer clairement les écarts entre les autres textes soit à l'auteur, soit au sous-genre. Pour mieux percevoir ces écarts, on peut éliminer le *Dialogue* et la *Germanie,* de même que les formes de futur et futur antérieur. On obtient alors la figure 2. Celle-ci montre clairement que certaines œuvres présentent une très grande cohérence dans l'emploi des temps alors que d'autres apparaissent beaucoup plus éclatées : les diverses *Vies* de Suétone (à gauche) apparaissent groupées ; de même pour les livres de la *Vie d'Alexandre* (QC3-10 en haut); si l'on relève par ailleurs une très grande proximité entre la fin de la *Guerre des Gaules* (Gall.5-7) et le début de la *Guerre civile* (Civ.1-2) (à droite), le troisième livre de cette dernière œuvre (Civ.3 à gauche) présente, lui, un profil assez différent; parmi les continuateurs de César, Hirtius (Gall.8) reste proche du modèle, alors que les auteurs de la *Guerre d'Espagne* (Hisp.) et de la *Guerre d'Afrique* (Afri.) s'en écartent très sensiblement.

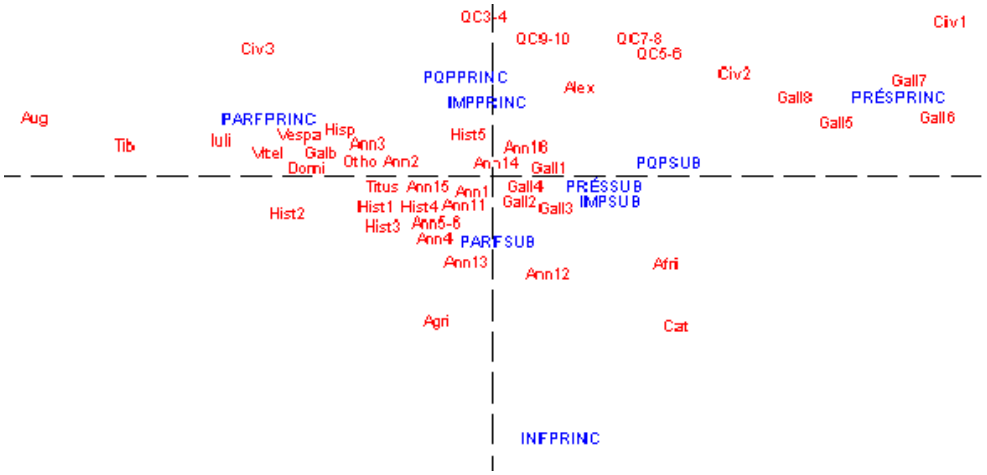

Figure 2

13 Le premier facteur d'analyse est à l'évidence d'ordre chronologique : sauf exception (comme Civ.3), le présent historique est utilisé largement dans les œuvres les plus anciennes, comme celles de César et de Salluste, alors que le parfait prend le dessus à l'époque impériale, spécialement chez Suétone, plus d'un siècle et demi plus tard. Le phénomène est toutefois à relativiser dans la mesure où des œuvres importantes, comme le début de la *Guerre des Gaules* ou les *Annales*, proches de l'origine des axes, semblent fort peu sensibles à celui-ci. Le second facteur d'analyse, – figuré le long de l'axe vertical –, est lié à l'utilisation de l'infinitif de narration, très fréquent chez Salluste (Cat. et Jug.), dans la *Vie d'Agricola* et dans certains livres des *Annales*.

14 Ce poids de l'infinitif de narration se marque clairement sur le deuxième plan d'analyse (figure 3). Sur le 3e axe (vertical), apparaît une opposition dans la manière dont les auteurs rapportent à l'imparfait ou au plus-que-parfait des faits relevant de l'arrière-plan narratif : au début de leurs œuvres, Quinte-Curce et Tacite (en bas) rapportent ces faits plus fréquemment en principale, alors que, chez César ou Suétone (en haut), ils sont plus volontiers décrits dans des subordonnées.

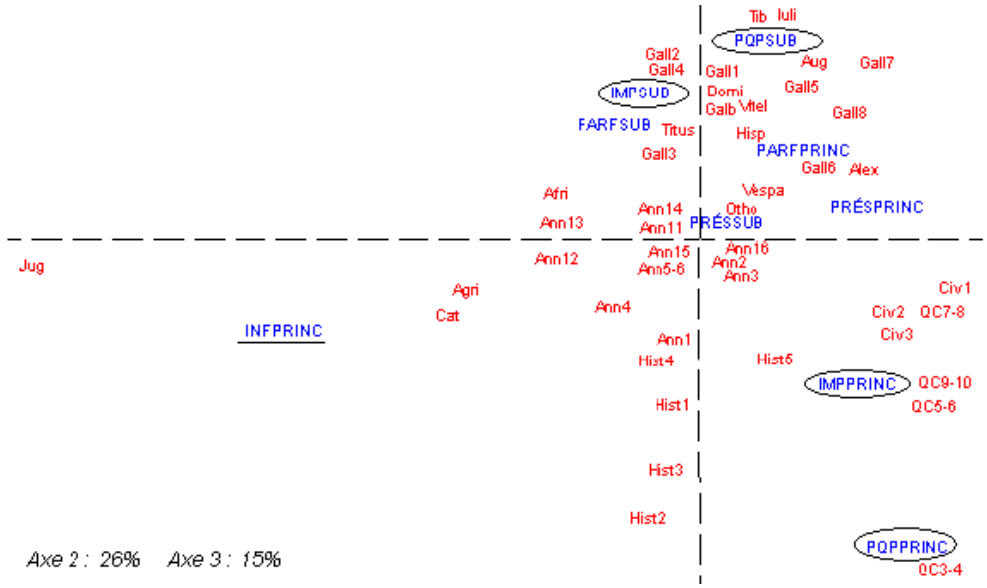

Figure 3

15 Les distances semblent bien relever ici soit du style propre de l'auteur, soit de l'évolution de son style au fil de son œuvre. Qu'en est-il alors du facteur générique ? Son poids semble de prime abord ci fort limité, puisque des œuvres à composante biographique comme l'*Agricola* de Tacite et les *Vies* de Suétone semblent ne pas se rapprocher. Le genre de l'œuvre influe certes sur la distribution temporelle, mais essentiellement lorsqu'il s'agit de genres clairement différents, comme on l'a vu avec la figure 1. L'étude de la distribution des temps entre principales et subordonnées est insuffisante à elle seule pour arriver à caractériser des sous-genres au sein du corpus historique. De simples dénombrements des formes présentent en effet le lourd inconvénient de ne pas pouvoir rendre compte de la répartition des formes dans les textes. Or si l'on considère qu'un texte est avant tout un ensemble ordonné, il paraît utile de rechercher une méthode pour rendre compte de cette structure d'ordre.

4. Étude des séquences

16 La structure des fichiers du L.A.S.LA. facilite une prise en compte de la linéarité du texte : à partir des fichiers sources, il est aisé, avec un simple tableur, de pouvoir extraire les codes qui correspondent aux mode, temps et fonction syntaxique (prédicat de subordonnée ou de principale) des formes verbales (*cf. supra*). On peut ainsi obtenir des listes de nombres représentant l'ensemble des formes verbales d'un texte dans leur ordre linéaire. Ainsi la suite de nombre 115 011 011 132 011 correspond aux premières formes verbales du 7e livre de la *Guerre des Gaules* : le premier chiffre indique le statut syntaxique (1 = prédicat d'une subordonnée; 0 = prédicat d'une principale) ; le deuxième signale le mode (par exemple, 1 = indicatif ; 3 = le subjonctif) ; le troisième correspond au temps (par exemple, 1 = présent ; 2 = imparfait; 5 = plus-que-parfait); le code 115 indiquera donc un prédicat de subordonnée à l'indicatif plus-que-parfait (*constituerat*).

17 A partir de ces listes, deux approches de la répartition linéaire des formes verbales sont possibles : la première consiste à s'intéresser à la répartition globale des formes dans l'ensemble du texte en étudiant leur distribution à travers les différentes zones qui constituent ledit texte[9] ; la seconde porte sur les séquences de formes. Celles-ci intègrent en effet l'axe syntagmatique en ce qu'elles rendent compte des effets de succession, – enchaînements ou ruptures –, dans l'emploi des temps verbaux. Certaines séquences sont linguistiquement conditionnées: ainsi le phénomène de la concordance des temps limite la liberté du choix du temps en subordonnée. En revanche, les enchaînements temporels de proposition principale à proposition principale sont généralement libres. Pour l'étude des séquences, on a donc réduit les textes à la simple liste des codes caractérisant les propositions principales (pour le début du 7ème livre de la *Guerre des Gaules*, la liste 115 011 011 132 011 se réduit à 011 011 011 ou mieux encore à des codes à deux chiffres 11 11 11). A partir de telles listes, il est possible de dénombrer diverses séquences de deux codes identiques successifs, puis de trois, de quatre, de cinq, etc., ou de dénombrer des ruptures (séquences de deux codes différents).

18 Pour ce faire, il a fallu mettre au point une programmation informatique spécifique, en s'appuyant sur une base de programmation modulaire en langage Java mise au point par Xuan Luong. La manière de procéder avec cette base est la suivante : à partir d'une plate-forme centrale, on pilote plusieurs modules indépendants, chacun, -représenté par une fenêtre écran-, étant destiné à effectuer une tâche spécifique. Pour chaque

question posée, un nouveau module de programmation a donc été crée et intégré à l'environnement existant, ce qui n'a nécessité qu'un minimum de codes et de manipulations informatiques.

19 Grâce à cette programmation, nous avons pu commencer par étudier la succession en principale de temps identiques. Deux approches complémentaires de la question s'offraient à nous : soit étudier des séquences « cumulées », soit envisager la longueur absolue des séquences. Les deux exemples qui suivent illustreront les deux méthodes.

4.1. Étude des séquences cumulées

20 La première des deux méthodes consiste à dénombrer des séquences cumulées d'éléments successifs identiques, à commencer par des séquences de deux éléments. Voici quelques exemples illustrant la manière dont le comptage a été effectué :

- dans la liste 14 14 14 14 12 11 12 14 14 14, on relève 5 séquences de deux indicatifs parfaits (= 14 14) successifs sur 7 parfaits ;
- dans la liste 12 12 15 12 12 11 14 14 12 14, on note 2 séquences de 2 indicatifs imparfaits (= 12 12) sur 4 occurrences d'imparfait et 1 séquence de 2 indicatifs parfaits (= 14 14) sur 3 occurrences de parfait ;
- la liste 14 14 14 14 14 14 14 11 14 14 compte 7 séquences de 2 indicatifs parfaits sur 9 occurrences de parfait.

21 Ce même type de dénombrement a été effectué informatiquement pour tous les textes du corpus. Les données obtenues ne pouvaient toutefois pas être soumises telles quelles à une AFC. En effet, puisque ce qui nous intéressait, c'était de comparer la manière dont chaque auteur emploie chaque temps, il semblait utile de pouvoir confronter des données équivalentes : il va de soi que, par exemple, le parfait aura plus de chances dans l'absolu d'apparaître dans des séquences de 2 temps identiques chez un auteur qui emploie fréquemment ce temps que chez un auteur qui ne l'emploie que plus rarement. Mais cela ne veut pas dire automatiquement que l'auteur qui emploie la forme moins souvent a pour autant tendance à la faire apparaître moins fréquemment dans des séries que ne le font les autres auteurs. Pour pouvoir centrer l'étude sur la manière même dont les temps sont employés et non pas sur la fréquence de cet emploi, on a donc rapporté le nombre de séquences cumulées de chaque temps au nombre d'occurrences de ce même temps dans le texte : pour la première liste, cela reviendrait à diviser le nombre des séquences de parfaits, c'est-à-dire 5, par le nombre des occurrences de parfait de la liste, soit 7, ou dans la dernière liste, à diviser 7 par le nombre de parfaits, soit 9. Les rapports ainsi obtenus ont été soumis à une AFC dont voici le résultat.

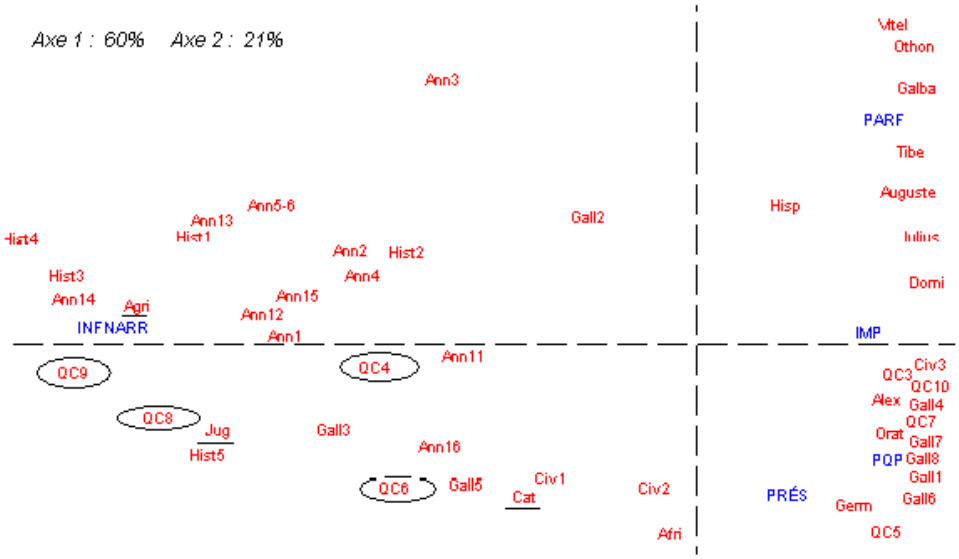

Figure 4

22 Les latinistes s'accordent à reconnaître que l'infinitif de narration présente une tendance très nette à apparaître au sein de séries, ce qui est vrai pour des textes où ce temps est fréquent (soulignés : *Vie d'Agricola*, *Guerre de Jugurtha*...), mais aussi pour des textes où il est beaucoup plus rare (entourés), comme chez Quinte-Curce. Cette particularité de l'infinitif de narration rend le premier plan d'analyse peu lisible. Le facteur 2 apparaîtra plus clairement sur le deuxième plan d'analyse (figure 5). Sur l'axe 2, on remarque que l'imparfait et le plus-que-parfait, quoique rares en principale chez César ou Salluste, apparaissent souvent dans des séries chez ces auteurs (soulignés à gauche), ce qui n'est pas le cas chez Suétone qui, lui, privilégie les séquences de parfaits (soulignés à droite). Le facteur chronologique semble bien conserver ici un rôle non négligeable. Sur l'axe 3, le poids de la *Germanie* (entourée) semble déterminant, les présents se succédant en longues séquences dans cette description géographique. Le genre de l'œuvre a donc bien un poids non négligeable sur la répétition de deux temps identiques successifs, mais le critère ne permet pas à proprement parler de reconnaître des groupements d'œuvres appartenant à un même sous-genre.

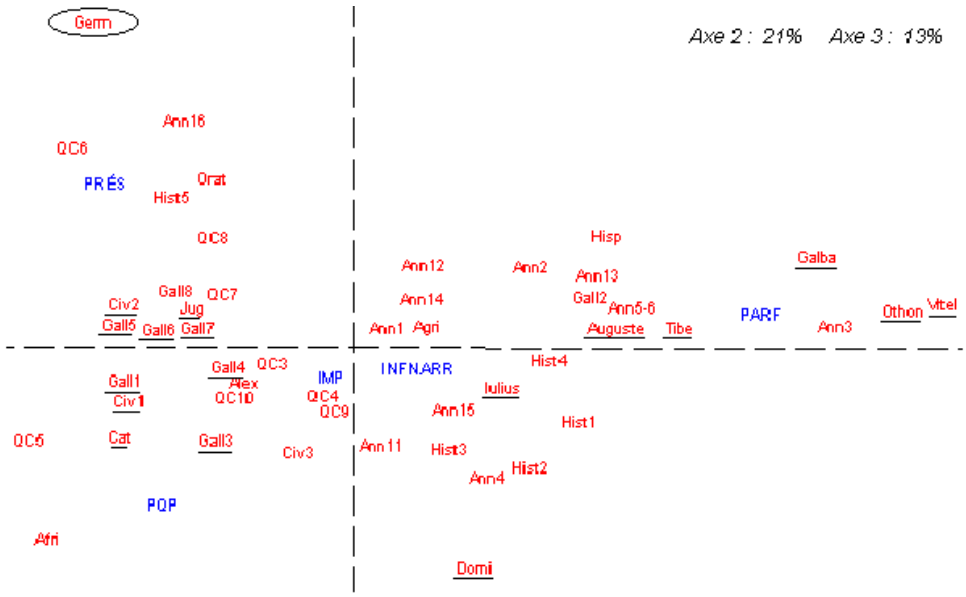

Figure 5

23 Pour préciser l'analyse, on peut limiter l'étude à la confrontation de l'œuvre de deux auteurs, en l'occurrence Tacite et Suétone. Le résultat de cette étude apparaît sur la figure 6.

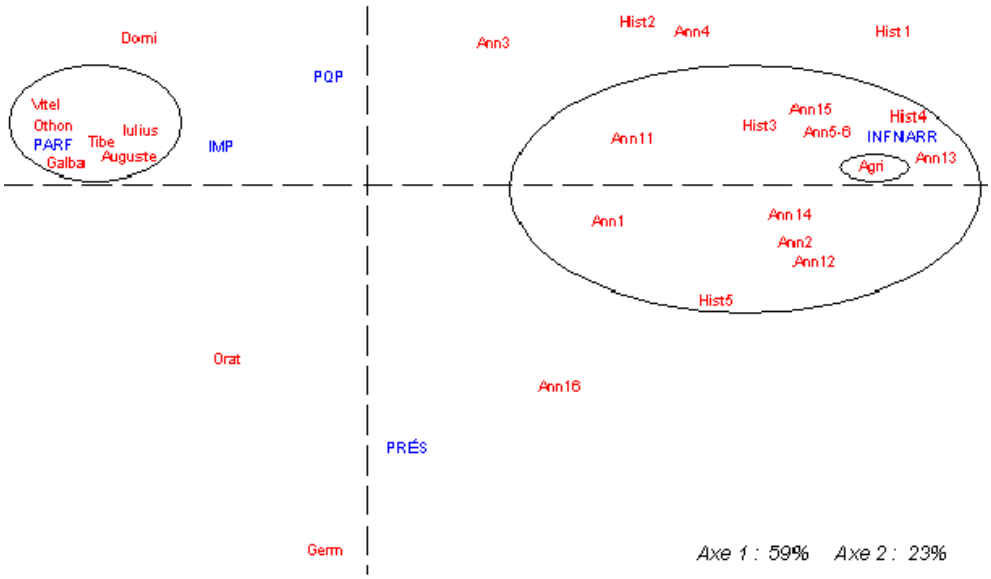

Figure 6

24 Les deux œuvres s'opposent nettement aux deux extrémités de l'axe horizontal, avec une exception toutefois pour la *Germanie* et le *Dialogue des Orateurs*. Dans ses œuvres historiques, Tacite, contrairement à Suétone, préfère, en principale, l'alternance de temps différents, avec la seule exception de l'infinitif de narration. Le regroupement des *Vies* de Suétone, de la *Germanie* et du *Dialogue des orateurs* indique que le genre littéraire des œuvres prime ici sur l'auteur. La présence de la *Vie d'Agricola* parmi les *Histoires* et les *Annales* semble indiquer que la composante biographique de l'œuvre ne suffit pas à l'écarter significativement du reste de l'œuvre. Ici encore le poids de

l'infinitif de narration pourrait fausser la perspective. Pour mieux percevoir les distances entre les œuvres, on peut affiner l'analyse en prenant cette fois en compte la longueur absolue des séquences.

4.2. Étude des longueurs des séquences

25 Les séquences de parfaits successifs semblent avoir un poids important dans la distance existant entre les œuvres de tacite et Suétone. C'est donc par l'examen de ces séquences que nous illustrerons l'étude de la longueur des séquences. Voyons, à partir des listes déjà examinées précédemment, comment les dénombrements ont été effectués :

- dans la liste 14 14 14 14 12 11 12 14 14 14, on note 1 séquence de 4 indicatifs parfaits et 1 séquence de 3 parfaits sur 7 parfaits ;
- dans la liste 12 12 15 12 12 11 14 14 12 14, on compte 1 seule séquence de 2 indicatifs parfaits sur 3 occurrences au total ;
- la liste 14 14 14 14 14 14 14 11 14 14 compte 1 séquence de 7 indicatifs parfaits et 1 séquence de 2 parfaits sur 9 occurrences.

26 Utilisant le logiciel modulaire décrit plus haut, on a procédé au comptage de ces différentes séquences dans l'ensemble du corpus, puis, comme cela a été fait pour les séquences cumulées, on a divisé les chiffres obtenus par le nombre d'occurrences de parfait dans chaque texte et on a ainsi obtenu des rapports auxquels a été appliquée une analyse factorielle.

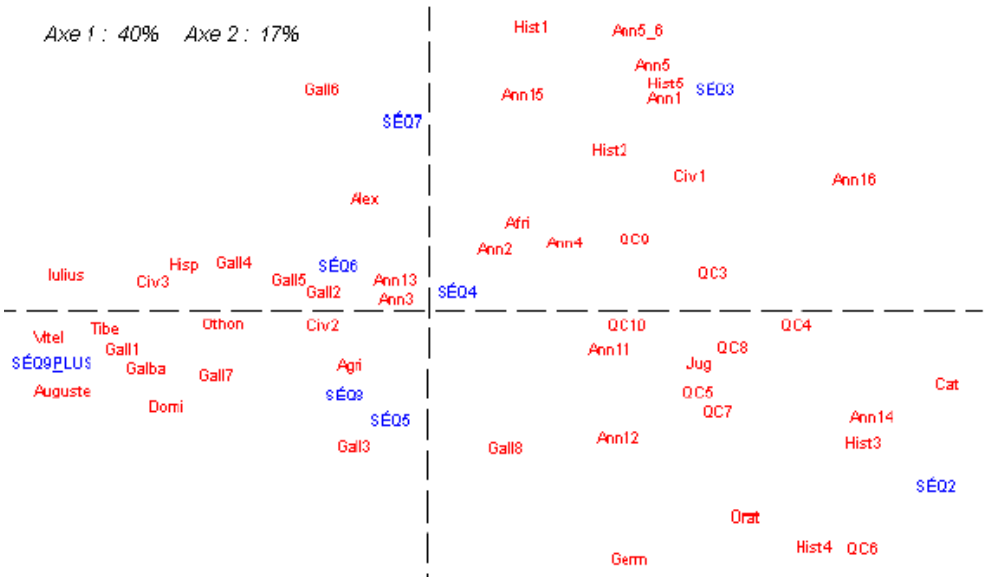

Figure 7

27 Sur la figure 7, on voit clairement que l'axe 1 s'organise essentiellement selon la longueur des séquences : on constate que le parfait apparaît essentiellement en longues séquences chez Suétone qui emploie largement ce temps, mais aussi chez César, chez qui ce temps est nettement moins fréquent. Même si les différences dans la longueur des phrases et le nombre des subordonnées ne permettent pas de parler de caractère césarien pour l'œuvre de Suétone, on ne peut ici que constater un nouveau rapprochement entre les deux auteurs. En revanche Salluste, Quinte-Curce et Tacite utilisent de préférence des séquences plus courtes. L'opposition est d'autant plus claire lorsque l'on limite l'étude aux seules œuvres de Tacite et Suétone : sur la figure 8, les

deux œuvres s'opposent clairement aux deux extrémités du tableau. Il est intéressant de constater ici que la *Vie d'Agricola* semble occuper une position médiane entre le reste de l'œuvre de Tacite où des séquences courtes de 2 ou 3 parfaits prédominent et les *Vies* de Suétone où l'on rencontre essentiellement des séquences longues de 5, 6 ou 7 parfaits consécutifs. Cette proximité pourrait s'expliquer par la composante biographique de l'œuvre. Seul le retour au texte peut bien sûr ici permettre d'étayer cette hypothèse.

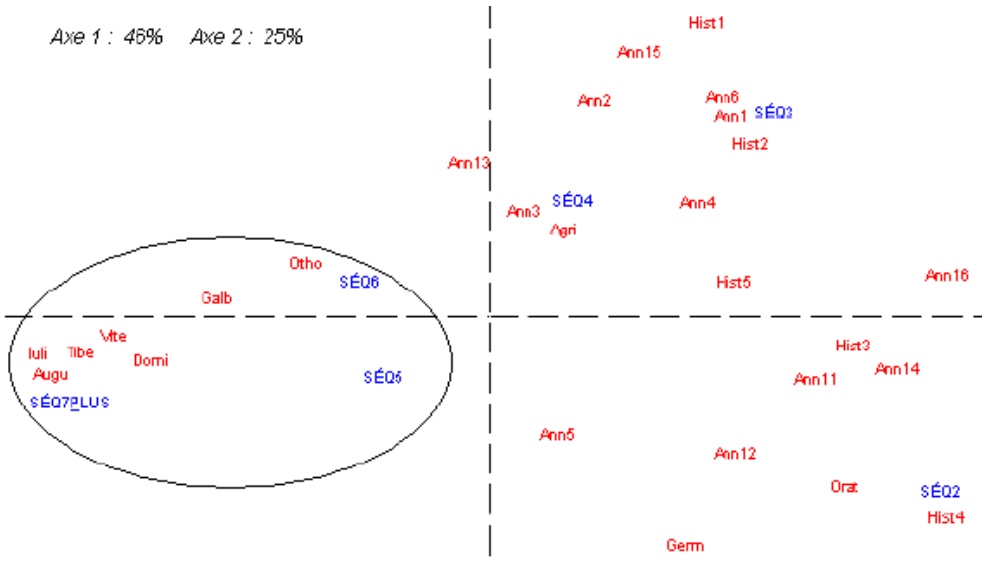

Figure 8

5. Conclusions provisoires

28 En prenant en compte l'aspect dynamique du texte, la méthode décrite ici permet de mettre en évidence des distances et des proximités entre auteurs qu'une simple analyse de fréquences ne permet pas de relever : ainsi, si les œuvres de César et Suétone s'opposent quant aux proportions de présents et de parfaits en principale, les deux auteurs affectionnent l'un comme l'autre les longues séquences de parfaits. La méthode mise au point pourrait bien évidemment être utilisées pour d'autres corpus. En ce qui concerne le latin, on peut espérer, grâce à ce type d'analyse, pouvoir à terme dégager des critères permettant de caractériser plusieurs sous-genres au sein de la prose historique latine. La recherche devra se poursuivre : il faudra notamment intégrer au corpus informatisé les deux œuvres de références importantes qui manquent encore : pour l'histoire annalistique, ce que nous avons conservé de Tite-Live, et pour la biographie, les *Vies* de Cornelius Nepos. L'étude des fréquences ou des longueur de temps identiques pourrait en outre être complétée par un examen de la répartition de ces séquences dans les différentes zones du texte. On pourra ainsi sans doute contribuer à une meilleure connaissance de l'œuvre des historiens latins, mais aussi mettre au point de nouvelles méthodes d'analyse en matière de calcul de distance et de topographie textuelle.

BIBLIOGRAPHIE

Denooz J. (1978). « L'ordinateur et le latin, techniques et méthodes ». *R.E.L.O.*, vol. 4 : 1-36.

Denooz J. (1988a). « Quelques observations sur la fréquence des substantifs et des prépositions dans la littérature latine ». *Revue, Informatique et statistique dans les sciences humaines* 24 (= *Le nombre et le texte, Hommage à Étienne Évrard*) : 151-162.

Denooz J. (1988b). « Application des méthodes d'analyse factorielle à la fréquence des catégories grammaticales en latin », *Les Cahiers de l'Analyse des Données* 1 : 19-40.

Dubrocard M. (1988). « Problèmes d'attribution : le choix des critères », *Revue, Informatique et statistique dans les sciences humaines* 24 (= *Le nombre et le texte, Hommage à Étienne Évrard*) : 163-179.

Évrard É. (1966). « La fréquence des phénomènes grammaticaux est-elle constante ? », in *Actes du premier colloque international de linguistique appliquée*, Nancy, 163-179.

Longrée D. (2003). « Tacite et Suétone : linguistique comparative et genres littéraires », in G. Lachenaud & D. Longrée (éds), *Grecs et Romains aux prises avec l'histoire, Représentations, récits et idéologie,* Vol. 1, Rennes, 315-326.

Longrée D. (à paraître). « Temps verbaux et spécificités stylistiques chez les historiens latins : sur les méthodes d'analyse statistique d'un corpus lemmatisé », in G. Calboli *et alii*, *Actes du 12ème Colloque International de Linguistique latine, Bologne, 9-14 juin 2003.*

Longrée D. & Xuan L. (2003, à paraître). « Spécificités stylistiques et distributions temporelles chez les historiens latins : sur les méthodes d'analyse quantitative d'un corpus lemmatisé », in G. Williams (éd.), *Actes des 2èmes Journées de la Linguistique de Corpus (Lorient, 12 au 14 septembre 2002)*, Rennes : Presses universitaires de Rennes.

Mellet S. (1994). « Logiciels d'exploitation de la banque de données de textes latins du L.A.S.L.A. », *Revue, Informatique et statistique dans les sciences humaines* 30 : 91-108.

Mellet S. (1996). « Les atouts de la lemmatisation », in G. Moracchini (éd.), *Actes du Colloque international "Bases de données linguistiques : conceptions, réalisations, exploitatione"*, Univ. de Corse / Univ. de Nice - Sophia Antipolis, 309-316.

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

Mellet S. & Purnelle G. (2002). « Les atouts multiples de la lemmatisation : l'exemple du latin », in A. Morin et P. Sébillot (éds), *JADT 2002, 6èmes Journées internationales d'Analyse statistique des Données Textuelles, Saint-Malo, 13-15 mars,* Vol. 2, Rennes, 529-538.

NOTES

1.. Entre autres É. Étienne (1966), J. Denooz (1988a, 1988b) ou dans le présent volume X. Luong & S. Mellet.
2.. Comme l'ont mis en évidence, entre autres, M. Dubrocard (1988) et S. Mellet (2002).
3.. Sur cette question, *cf.* D Longrée (2003).
4.. Sur la constitution de la base de données du L.A.S.L.A, *cf.* J. Denooz (1978).
5.. Sur l'intérêt d'un tel encodage, *cf.* S. Mellet (1996) et S. Mellet &G. Purnelle (2002).
6.. Sur ces logiciels, *cf.* S. Mellet (1994) et S. Mellet & G. Purnelle (2002).

7.. Celles-ci ont été réalisées avec le logiciel ANAR (Analyse factorielle et arborée), version bêta - octobre 2000, fournie par Étienne Brunet en complément du logiciel Hyperbase.
8.. Pour les résultats détaillés de ces recherches, *cf.* D. Longrée (à paraître).
9.. Une telle recherche est actuellement en cours au sein du laboratoire « Bases, Corpus et Langage », UMR 6039, UNSA / CNRS (Sylvie Mellet, Xuan Luong & Dominique Longrée) et fera l'objet d'une communication aux JADT 2004, 7èmes Journées internationales d'Analyse statistique des Données Textuelles, Louvain-la-Neuve (Belgique).

RÉSUMÉS

Le calcul de distance entre les textes repose souvent sur l'étude de fréquences ; ainsi, les diverses œuvres des historiens latins peuvent être caractérisées par des dénombrements de catégories grammaticales, comme, par exemple, les temps et modes verbaux ; nous proposons ici une autre méthode prenant en compte la répartition des formes verbales le long de l'axe syntagmatique et rendant compte des effets de succession, – enchaînements et ruptures – , au sein de l'ensemble ordonné que constitue le texte. Des exemples mettront en évidence l'intérêt de la démarche quand il s'agit d'évaluer le poids des divers facteurs à l'origine des écarts entre textes : époque, auteur, genre littéraire...

Verbal tenses and linearity of the text
The intertextual distance is often calculated by studying frequencies. For instance, the Latin historians' works have been characterized by computing frequencies of grammatical categories such as tenses and moods. The method discribed here takes into account not only frequencies of verbal forms, but also their distribution and their order in the texts, by studying the way tenses follow one another (tense sequences or ruptures). The reliability of the method to differentiate texts according to diachronic evolution, author style or literary genre will be evaluated.

INDEX

Mots-clés : distance intertextuelle, analyse factorielle, axe syntagmatique, catégories grammaticales, séquences et ruptures temporelles, latin

AUTEURS

DOMINIQUE LONGRÉE

« Bases, Corpus et Langage », UMR 6039, Univ. d'Angers

XUAN LUONG

« Bases, Corpus et Langage », UMR 6039, UNSA