Temps verbaux, axe syntagmatique, topologie textuelle : analyses d'un corpus lemmatisé

Dominique Longrée[1], Xuan Luong[2], Sylvie Mellet[3]

[1]« Bases, Corpus et Langage » (UMR 6039), Univ. d'Angers – 49000 Angers – France
[2]« Bases, Corpus et Langage » (UMR 6039), UNSA – B.P. 209, 06204 Nice cedex 3 – France
[3]« Bases, Corpus et Langage » (UMR 6039), CNRS – B.P. 209, 06204 Nice cedex 3 – France
longree@fusl.ac.be, mellet@unice.fr

Abstract

The intertextual distance is often calculated by studying frequencies. As part of a research project on texts topology, the methods described here try to take into account the text linearity in this calculation. In fact, we will study the distribution of verbal forms, according to their tenses, in the general framework of the text. The corpus is based on a latin historical lemmatized texts databank. Several methods will be compared : texts segmentation by division or by neighborhood, distances representation by curves or trees.

Résumé

La mesure des distances entre les textes est le plus souvent effectuée à partir du dénombrement des unités d'analyse. Dans le cadre d'une réflexion générale sur la topologie textuelle, nous présentons quelques outils développés pour prendre en compte, dans ce calcul, la linéarité du texte et les localisations des unités au fil de cette chaîne linéaire. En l'occurrence, il s'agit d'examiner la répartition globale des temps verbaux le long de l'axe syntagmatique pour évaluer les distances et similarités entre des textes lemmatisés d'historiens latins. Plusieurs méthodes sont comparées : segmentation des textes par découpage ou par voisinage, représentation des distances par courbes ou par analyses arborées.

Mots-clés : distance intertextuelle, topologie textuelle, mesures de voisinage, courbes de distribution, analyses arborées, temps verbaux, latin.

1. Objectifs

La recherche ici exposée représente une étape méthodologique importante dans le cours de recherches entamées depuis quelques années pour tenter d'améliorer les calculs de distance entre les textes en prenant en compte deux paramètres originaux : travaillant sur des textes latins lemmatisés et étiquetés[1] les auteurs s'intéressent en effet depuis longtemps d'une part au rôle des catégories grammaticales dans la caractérisation d'un texte, d'autre part à la distribution de ces catégories au fil des textes, c'est-à-dire à la topologie textuelle entendue au sens des mathématiciens (cf. Luong et Mellet, 1995 ; Salem, 2002 ; Lamalle et Salem, 2002 ; Longrée et Luong, 2003). En intégrant ces deux paramètres, très généralement délaissés dans les études de statistique linguistique, nous espérons pouvoir mieux évaluer les distances et similarités intertextuelles. Les catégories grammaticales retenues ici seront les temps et modes verbaux, paramètres jugés pertinents pour la comparaison des textes étudiés ; on espère ainsi obtenir un critère efficace (parmi d'autres sans doute) de reconnaissance et de classification

[1] Nos corpus reposent sur les fichiers informatisés et lemmatisés par le L.A.S.L.A. à l'Université de Liège. Pour cette étude, nous avons travaillé sur l'ensemble du corpus césarien, y compris ses continuateurs, sur l'ensemble des œuvres conservées de Quinte-Curce et de Tacite, et sur 9 *Vies* extraites des *Vies des douze Césars* de Suétone. Pour une présentation plus détaillée, cf. Longrée et Luong (2003).

par auteurs et, surtout, par sous-genres. L'objectif ultime est en effet de pouvoir caractériser de manière endogène et aussi objective que possible, au sein de l'histoire latine, le sous-genre du commentaire, celui des annales, celui des biographies, etc. Le propos sera, plus précisément, de présenter quelques outils que nous avons développés pour prendre en compte les localisations des unités d'analyse dans la chaîne linéaire des textes : segmentation des textes par découpage et voisinage, représentation des distances par courbes de distribution et analyses arborées.

2. L'emploi des temps verbaux

Le terme *emploi* recouvre au moins trois réalités, trois objets d'étude statistique différents.

2.1. ***Les fréquences*** : les unités d'analyses sont considérées globalement et en vrac ; elles donnent lieu à des dénombrements qui permettent de constituer de classiques tableaux de contingence à partir desquels des AFC, des analyses arborées ou des analyses en composantes principales permettront de représenter les similarités des profils textuels (cf. Longrée, à paraître 2004). Il est à noter cependant que les matrices issues du dénombrement des catégories grammaticales présentent des spécificités par rapport aux matrices lexicales qui nécessitent un aménagement des analyses traditionnelles (obligation de travailler sur les fréquences et non pas en termes de présence / absence ; prise en compte de quelques colonnes creuses et néanmoins significatives (cf. Luong et Mellet, 2003)).

2.2. ***Les séquences*** : elles intègrent 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, en latin, est-il rarissime qu'un infinitif de narration apparaisse isolément ; on a plutôt affaire à des séries d'au moins trois occurrences et souvent plus. La séquence des temps en propositions subordonnées est le plus souvent contrainte par le temps du verbe principal (phénomène dit de « concordance des temps »). Mais la plupart des enchaînements de proposition principale à proposition principale sont libres. C'est pourquoi nous avons réduit nos textes à la simple succession des codes caractérisant les formes verbales des propositions principales et indépendantes afin de pouvoir dénombrer les séquences de deux codes identiques successifs, puis de trois, de quatre, de cinq, etc. codes identiques successifs (soit de manière absolue, soit de manière cumulative, une séquence de quatre codes identiques successifs étant alors également comptabilisée comme 2 séquences de trois codes et 3 séquences de deux codes ; complémentairement, cette réduction permet aussi de mettre en évidence et d'analyser les types de ruptures dans la succession des temps principaux. Les résultats, publiés dans Longrée et Luong (2003) et Longrée (à paraître, 2003) sont prometteurs.

2.3. ***La répartition globale sur l'ensemble d'un texte*** : il s'agit d'une troisième réalité qui associe la saisie de l'axe syntagmatique à courte portée déjà présente dans les séquences et une saisie à longue portée qui tient compte de la répartition des unités étudiées dans les différentes zones du texte. C'est là qu'on retrouve véritablement la notion de topologie textuelle.

3. Les méthodes : les représentations graphiques

Il s'agit donc de synthétiser les informations contenues dans les deux graphes ci-dessous et de leur donner une forme exploitable pour un calcul de distance.

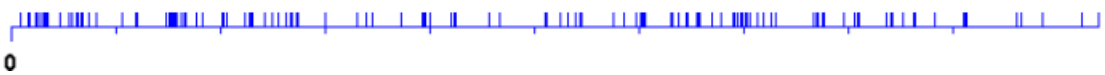


Figure 1. Répartition linéaire des imparfaits dans la Guerre d'Alexandrie

Ce graphe visualise la répartition des formes d'imparfaits dans la chaîne linéaire du texte : il montre assez clairement les zones de densité importante dans les premier et troisième quarts du texte, et les zones de faible présence dans les deux autres quarts, tout particulièrement à la fin du texte. Les données numériques associées à ce graphe sont les suivantes :

Nombre de mots : 10666

Nombre d'imparfaits : 160

Distance moyenne théorique entre deux imparfaits : 66,66

Distance maximale observée : 482

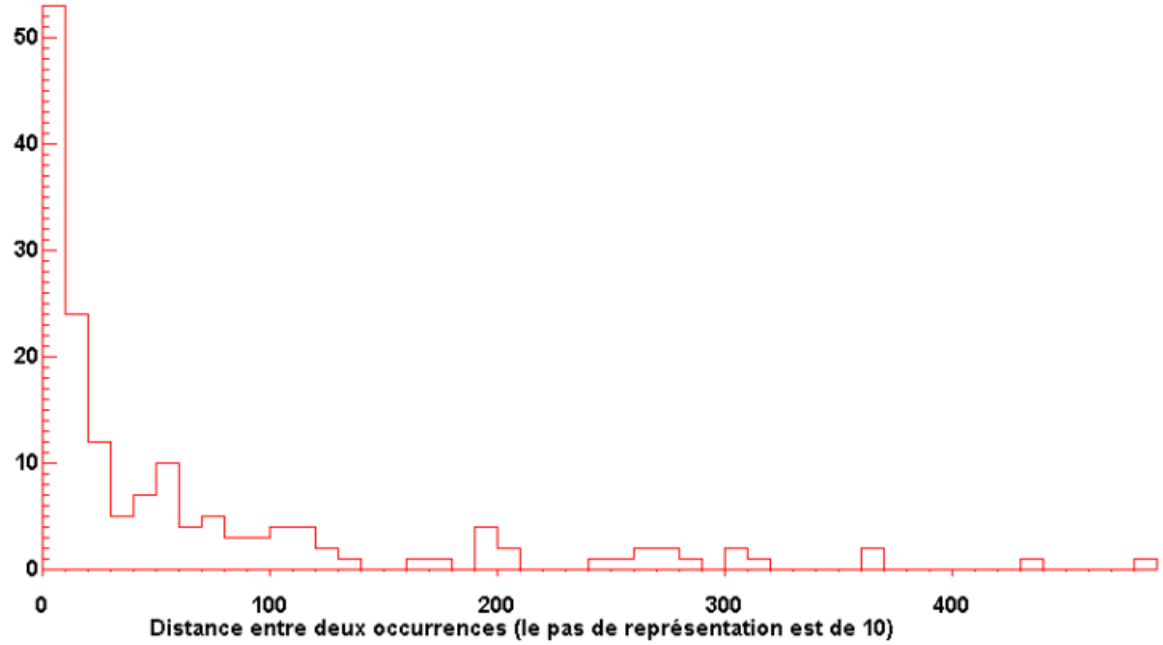


Figure 2. Graphe de répartition des imparfaits dans la Guerre d'Alexandrie

La figure 2 offre un autre mode de représentation de la même réalité, donnant à voir le nombre d'occurrences d'imparfait séparées par un intervalle inférieur ou égal à 10 formes graphiques (ici 53), le nombre de celles qui sont séparées par un intervalle compris entre 10 et 20 formes (24), etc. Les espacements réduits sont majoritaires, mais on relève aussi quelques espacements très importants (2 de plus de 370 formes, 1 de plus de 430 et 1 de plus de 480).

3.1. ***Le découpage en zones*** : la première idée est de découper chacun des textes du corpus en « zones » ou « tranches » successives, de longueur arbitraire, et de dénombrer les occurrences des différents temps verbaux des propositions principales dans chacune de ces tranches pour pouvoir ensuite traiter la chaîne numérique ainsi obtenue. Concrètement, un texte T divisé en N tranches sera caractérisé par une série de profils tels que :

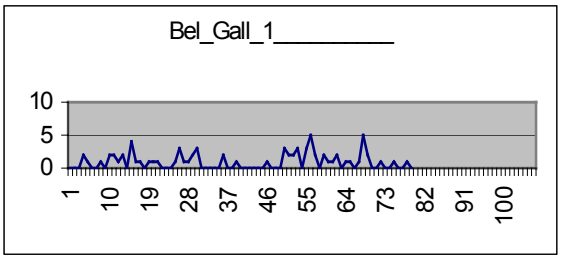
$$I_T = f_{i\,1} \quad f_{i\,2} \quad f_{i\,3} \quad \ldots \quad f_{i\,n}$$

$$P_T = f_{p\,1} \quad f_{p\,2} \quad f_{p\,3} \quad \ldots \quad f_{p\,n}$$

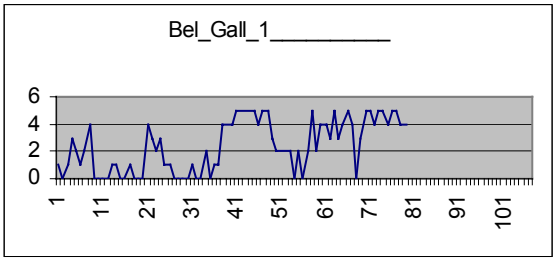
$$PR_T = f_{pr\,1} \quad f_{pr\,2} \quad f_{pr\,3} \quad \ldots \quad f_{pr\,n}$$

où $f_{i\,n}$ note la fréquence de l'imparfait dans les tranches 1, 2, 3 jusqu'à N du texte T, où $f_{p\,n}$ note la fréquence du parfait dans les tranches 1 à N du même texte et où $f_{pr\,n}$ note les fréquences du présent.

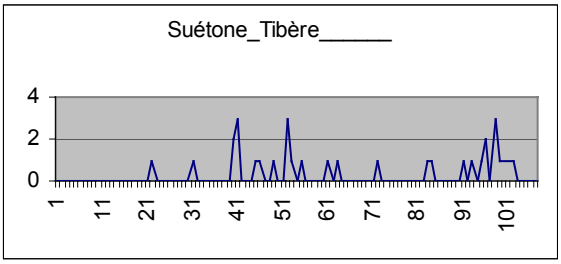
Une première représentation des données peut être réalisée sous formes de courbes grâce au simple outil Excel. Les graphes ci-dessous en donnent quelques exemples : deux textes, le livre 1 de la *Guerre des Gaules* et la *Vie de Tibère* de Suétone ont été réduits à la succession des codes caractérisant le temps et le mode de leurs verbes principaux, puis découpés en tranches de 5 codes chacune ; le premier est constitué de 79 tranches de 5 codes, le second de 107 tranches. On a décompté d'abord les occurrences du code d'imparfait, ensuite celles du code de parfait (dans chaque tranche leur nombre peut évidemment varier entre 0 et 5) :



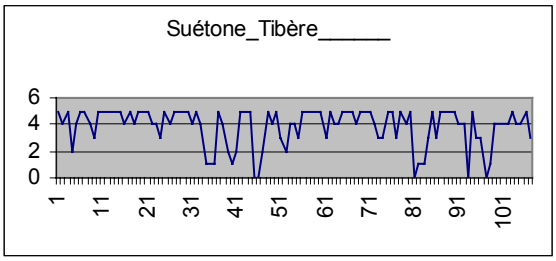
Imparfaits dans GALL_1



Parfaits dans GALL_1



Imparfaits dans TIBERE



Parfaits dans TIBERE

Cette représentation a le mérite d'être plus parlante qu'un tableau numérique. On y constate la relative complémentarité de la distribution de l'imparfait et du parfait, surtout dans le livre 1 de la *Guerre des Gaules* où la plupart des pics de l'un des temps correspondent aux creux de l'autre. On y voit que ni dans l'un, ni dans l'autre texte l'imparfait n'abonde dans les premiers chapitres, contrairement à ce que l'on pourrait attendre d'un temps descriptif d'arrière-plan destiné à poser le cadre des actions à venir. On y voit surtout la différence d'écriture entre les deux textes, avec un emploi beaucoup plus monotone du parfait dans la *Vie de Tibère* par Suétone. Ces analyses restent cependant très approximatives et de portée limitée.

On peut en outre se demander si cette méthode des tranches est tout à fait satisfaisante ; elle implique en effet une conception discontinue de la chaîne linéaire du texte. C'est pourquoi on estime pouvoir sensiblement améliorer la représentation en recourant à une mesure de densité des occurrences qui prend appui sur une véritable topologie discrète des textes.

3.2. ***La mesure de voisinage*** : il s'agit d'affecter à chaque unité constitutive du texte (c'est-à-dire, rappelons-le, à chaque code de temps verbal) une mesure de son voisinage calculée en fonction d'une propriété de ce voisinage jugée pertinente. Là encore, comme précédemment pour la longueur des tranches, la taille du voisinage peut être parfaitement arbitraire. Nous avons choisi une taille de 11 unités. Au sein de ce voisinage on a dans un premier temps retenu la propriété de présence du code 14 affecté au parfait de l'indicatif, temps fondamental de la narration. Ainsi, pour chaque unité du texte, un programme examine les cinq codes précédents et les cinq codes suivants et décompte le nombre d'occurrences du code 14 : nous donnons à cette mesure le nom de ***densité***. Le résultat aboutit à une représentation topologique discrète du texte qui rend mieux compte que la méthode précédente de sa continuité

puisque la même mesure est affectée à chaque unité successivement. Voici les résultats obtenus sur le début du livre 2 de la *Guerre Civile* :

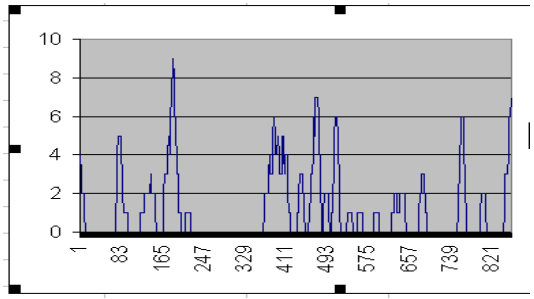
Chaîne linéaire des codes de temps et de modes affectés à chacun des verbes principaux :

11 12 11 11 11 11 11 11 12 12 12 12 12 12 12 12 12 12 11 11 11 11 11 15 15 12 15 15 14 11 11 15 11 11 11 11 11 11 11 15 15 15 11 12 12 15 14 12 12 12 12 12 12 14 14 14 14 11 11 14 14 12 14 14 14 14 14 14 14 14 14 14 12 12 12 12 14 14 14 14 14 14 14 14 14 14 14 14 15 14 14 12 12 12 12 14 14 etc.

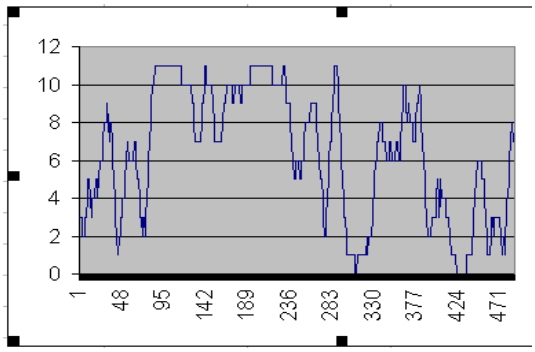
Topologie établie à partir des mesures de voisinage de chaque unité (taille du voisinage = 11, propriété du voisinage = nombre d'occurrences du code 14) :

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 2 3 4 5 4 4 5 6 6 7 8 8 8 8 8 9 10 10 10 10 9 8 7 7 7 7 7 7 7 8 9 10 11 11 10 10 10 9 8 7 6 6 6 6 etc.

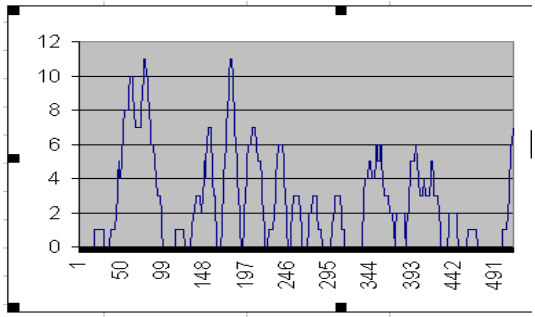
Cette nouvelle suite numérique peut de nouveau donner lieu à une représentation graphique sous forme de courbe. Voici celles de la *Guerre Civile* livres 1 et 2, la *Guerre des Gaules* livre 5, la *Guerre d'Espagne*, des livres 3 et 5 des *Histoires* de Quinte-Curce, des *Vies* de Iulius et de Tibère par Suétone et des livres 12, 13, 14 et 15 des *Annales* de Tacite.



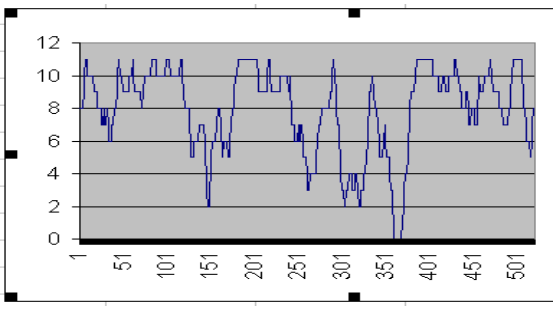
Guerre Civile 1 (César)



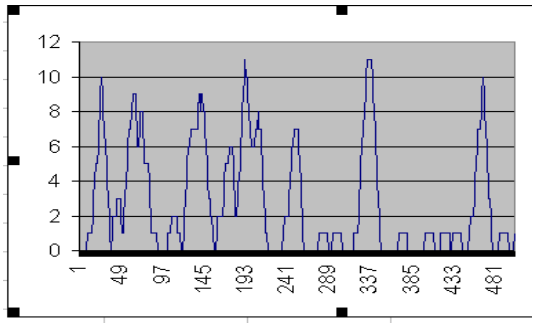
Guerre d'Espagne (imitateur de César)



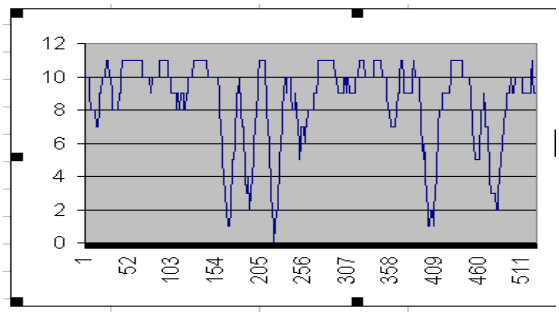
Guerre Civile 2 (César)



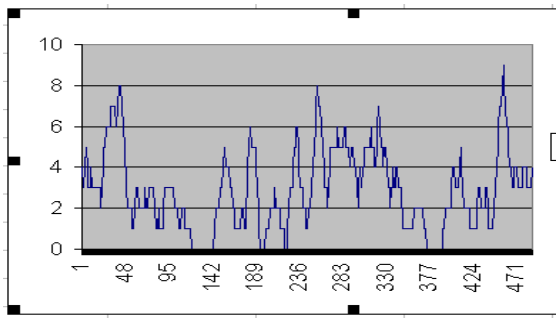
Vie de Iulius (Suétone)



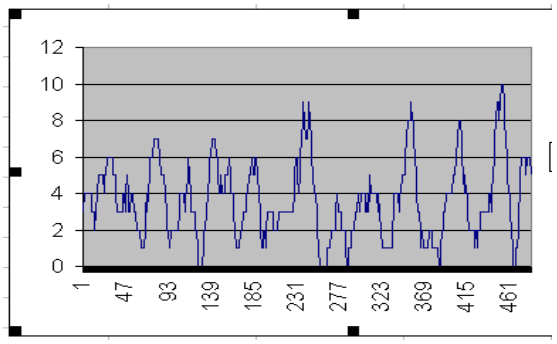
Guerre des Gaules 5 (César)



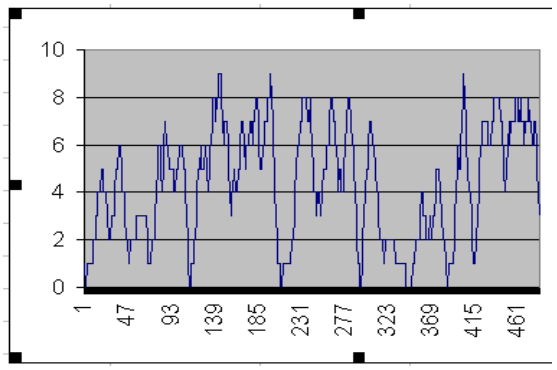
Vie de Tibère (Suétone)



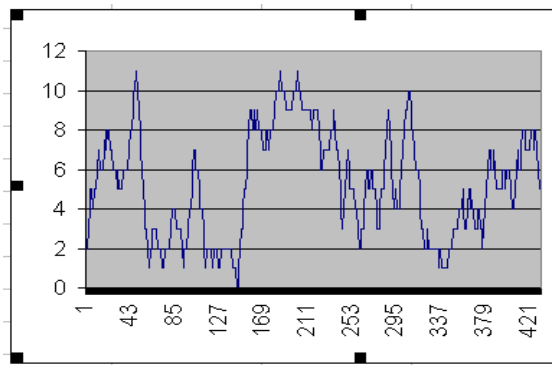
Quinte-Curce livre 3



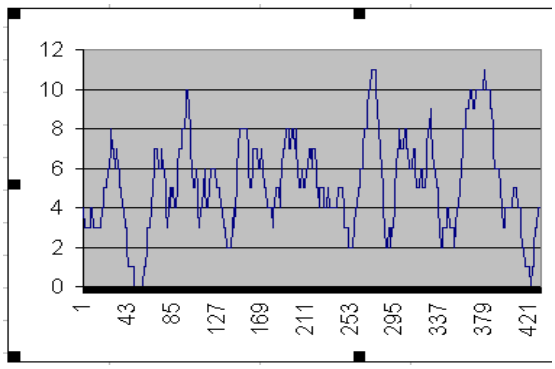
Quinte Curce livre 10



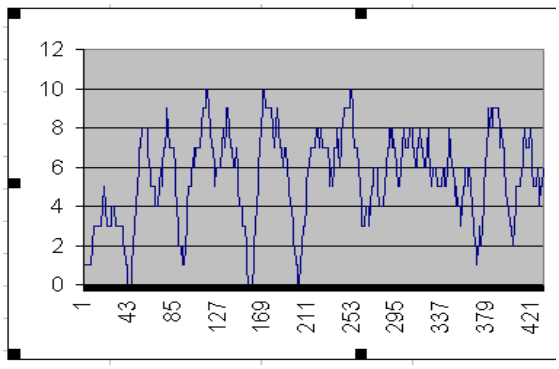
Annales 12 (Tacite)



Annales 13 (Tacite)



Annales 14 (Tacite)



Annales 15 (Tacite)

Là encore les courbes sont assez parlantes ; sans entrer dans une analyse approfondie, remarquons simplement la différence très nette entre les courbes des deux livres de César et celle de son imitateur, auteur de la *Guerre d'Espagne*. Remarquons aussi le caractère très particulier des courbes des *Vies* de Suétone. Notons que les livres de César commencent tous sur un mode qui tend à exclure le parfait alors que ceux de Suétone et, dans une moindre mesure, ceux de Quinte-Curce et de Tacite acceptent plus volontiers ce temps verbal dès les premiers paragraphes.

Cependant ces courbes restent peu exploitables en l'absence de méthode pour comparer deux courbes et surtout des courbes de longueur différentes. C'est pourquoi il nous a paru indispensable de revenir à des méthodes d'analyse moins qualitative des tableaux de données numériques.

4. Les calculs de distances et les analyses arborées

4.1. ***Le problème de la longueur*** se pose aussi, à dire vrai, pour les tableaux numériques. On touche là l'une des deux difficultés majeures auxquelles nous avons été confrontés et qui, toutes deux, concernent le découpage du texte :

– d'une part ce découpage ne peut pas être fait selon un empan fixe puisque le nombre de colonnes constitutives des différents profils de textes serait alors variable en fonction de la longueur des textes (un texte contenant 300 verbes principaux et un texte en contenant 350, découpés l'un et l'autre en tranches de 10, auraient, pour l'un, un profil de 30 tranches, pour l'autre, un profil de 35 tranches) ;

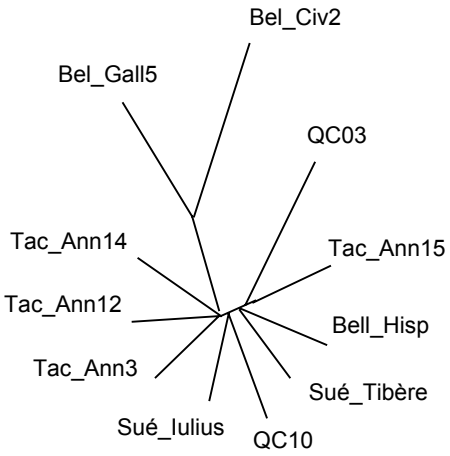
– d'autre part, dans le cadre d'un découpage devant aboutir à un nombre fixe de tranches (et donc de colonnes dans la matrice), quel est le meilleur critère de division ? On peut songer à un découpage « naturel » (du type : introduction – narration – péripétie – dénouement – conclusion). Cela pourrait être pertinent pour certains types de textes : contes traditionnels, notices explicatives ou modes d'emploi, exposés scientifiques, et même, dans le domaine latin, prose oratoire. Mais ce choix semble totalement inapplicable à textes qui sont, précisément, de structures variées, voire peu structurés. Par ailleurs, même dans l'hypothèse où un tel découpage paraîtrait praticable, on peut se demander s'il n'introduirait pas *de facto* dans les données ce qu'on y cherche, viciant ainsi dès le début toute la procédure d'analyse.

Il faut donc revenir à des tranches arbitraires et discontinues. La notion de paragraphe n'a aucun sens en latin, langue dont les scribes pratiquaient la *scriptio continua*. Celle de phrase est tout autant une projection des éditeurs modernes sur un texte en général non ponctué, mais au moins s'appuie-t-elle le plus souvent sur des critères syntaxiques et sémantiques assez fiables. Elle pourrait donc être adoptée. Cependant, étant donné la réduction que nous avons opérée sur les textes, qui les condense en une série de codes symbolisant les verbes principaux, il nous a semblé plus facile et plus pertinent de retenir cette dernière unité d'analyse comme unité de découpage (elle coïncide partiellement avec la notion de phrase, mais pas entièrement dans la mesure où une seule phrase peut être formée à partir de plusieurs propositions principales coordonnées ou juxtaposées). Nous avons également restreint le corpus à des textes de taille sensiblement égale (la longueur varie d'un facteur 1,2[2]).

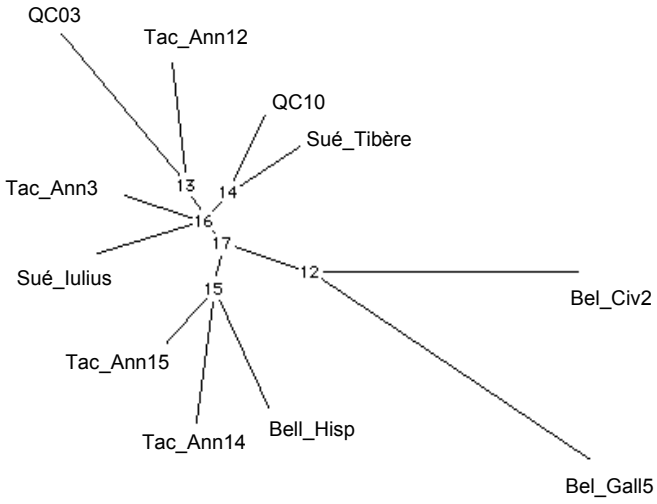
Nous avons ensuite choisi d'établir le fractionnement des textes en un nombre de tranches arbitraire qui varie de manière décroissante : les textes ont d'abord été découpés en 20 tranches, puis en 16, en 12, en 8, en 6 et enfin en 5 tranches. La réitération des calculs en faisant régulièrement et progressivement varier le nombre de tranches s'appuie sur l'espoir d'une part de trouver empiriquement la division la mieux adaptée à l'ensemble du corpus, d'autre part — et surtout, vu l'objectif à long terme de notre étude — de détecter les divisions les plus idoines pour chacun des sous-ensembles du corpus : en effet, il n'est pas exclu *a priori* que tel type de texte ou d'écriture révèle mieux son rythme narratif dans un profilage aux séquences très brèves et que tel autre s'insère mieux dans un profilage aux séquences plus longues. Le nombre d'occurrences du code 14 dans chaque tranche de chaque texte est comptabilisé, fournissant ainsi un profil du texte. Chacune des matrices de profils textuels ainsi obtenue a été soumise à un calcul de distance par le chi² et représentée graphiquement par une analyse arborée.

[2] Bel_Civ_2 : 520 codes verbaux ; Bel_Gal_5 : 512 ; Bel_Hisp : 496 ; QC03 : 558 ; QC10 : 492 ; Tac_Ann3 : 541; Tac_Ann12 : 519 ; Tac_Ann14 : 531 ; Tac_Ann15 : 575 ; Sué_Tibère : 536 ; Sué_Iulius : 529.

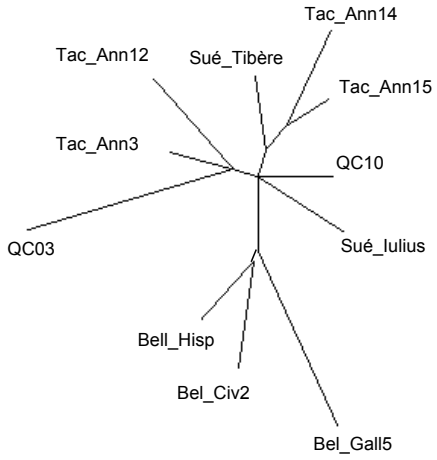
4.2. Premiers résultats et perspectives



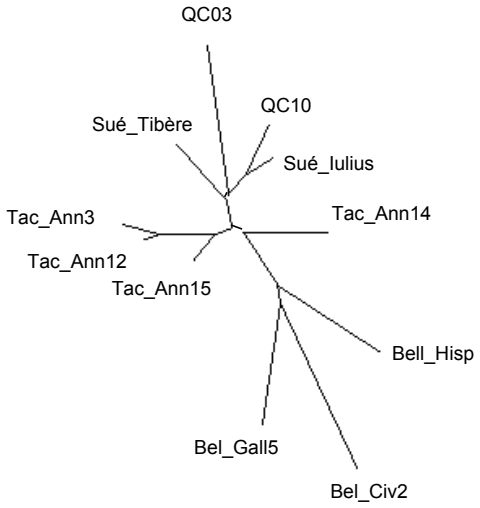
Découpage 20, code 14



Découpage 16, code 14

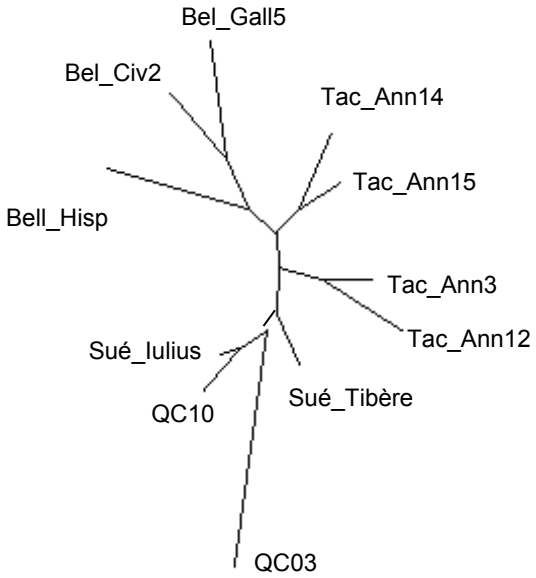


Découpage 8 , code 14



Découpage 6, Code 14

Comme on le voit sur les graphes ci-dessus et sur le graphe ci-dessous, le « découpage 20 » ne donne lieu qu'à un seul regroupement pertinent, celui des deux livres de César (qui reste stable à travers tous les découpages). Le reste prend la forme d'une étoile, dénotant l'absence de structure et de tout autre regroupement pertinent. Le « découpage 16 » ébauche une forme un peu mieux structurée, notamment autour du nœud 15, sans être toutefois très claire. Il faut en fait attendre les « découpages 6 et 5 » pour obtenir des arbres parfaitement structurés et interprétables, dans lesquels on observe un très net regroupement des œuvres par auteur.



Découpage 5, Code14

C'est donc un découpage en un nombre restreint de tranches qui permet de mettre en évidence cette structuration ; cela veut dire que le caractère dominant qui détermine la similarité ou la dissemblance entre les textes est leur structuration en larges zones correspondant à l'organisation du récit. Ce caractère semble devoir être attribué au sous-genre narratif (commentaire, biographie, etc.) plutôt qu'au style d'un auteur ; un bon indice en est la proximité récurrente des textes de Quinte-Curce et de Suétone, qui comportent les uns et les autres une forte composante biographique.

5. Conclusion

Les deux méthodes utilisées, — découpage du texte en un certain nombre de zones et mesure de voisinage —, mettent en évidence des aspects qui, tantôt rapprochent les textes, tantôt les séparent. La mesure de voisinage est un outil souple permettant de réduire la part d'arbitraire du descripteur. Cette méthode a l'avantage d'allier approche quantitative et approche qualitative dans la mesure où la qualité définitoire du voisinage peut être très variée : comme propriété pertinente, nous avons retenu ici la simple présence d'occurrences d'un temps donné au sein du voisinage, mais on peut également dénombrer, au sein du voisinage, des séquences diverses, comprenant ou excluant la présence de telle ou telle occurrence ; en outre des seuils de fréquences peuvent être fixés. La propriété du voisinage peut donc être associée à un objet de nature complexe. Toutefois, quand il s'agit de comparer des textes de longueurs différentes, la méthode présente un inconvénient majeur : le manque d'instruments de comparaison des courbes de distribution obtenues (à cet égard, toute suggestion serait la bienvenue). Le découpage du texte en un nombre fixe de tranches est une méthode moins subtile, mais cette segmentation présente l'avantage d'aboutir à un nombre fixe de colonnes dans la matrice et de permettre, pour la mesure des distances intertextuelles, d'utiliser l'outil éprouvé que constitue l'analyse arborée. En outre, cette méthode tend à mettre en évidence la structure même des textes, plutôt que les techniques d'écriture propres à chaque auteur. Une piste pour associer les avantages de chacune des deux méthodes serait de confier à un programme informatique le soin de découper la série numérique qui rend compte de la topologie du texte en N tranches respectant au mieux les variations de densité observées. Il s'agirait en quelque sorte de pro-

duire une structure en N tranches approximant au mieux la structure réelle de la topologie. Une recherche en ce sens est actuellement en cours.

Références

Évrard Ét. et Mellet S. (1998). Les méthodes quantitatives en langues anciennes. LALIES, vol. (18) : 111-155.

Lamalle C. et Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In *Actes des JADT 2002* : 403-411.

Longrée D. et Luong X. (à paraître, 2003). Spécificités stylistiques et distributions temporelles chez les historiens latins. In Williams G. (Ed.), *Actes des 2èmes Journées de la Linguistique de Corpus (Lorient, 12-14 sept. 2002)*. A paraître aux Presses universitaires de Rennes.

Longrée D. (à paraître, 2004). Temps verbaux et spécificités stylistiques chez les historiens latins. In Calboli G. *et al.* (Eds), *12ème Colloque intern. de Linguistique latine* (Bologne, 9-14 juin 2003).

Longrée D. et Luong X. (2003). Temps verbaux et linéarité du texte : recherches sur les distances dans un corpus de textes latins lemmatisés. *Corpus*, vol. (2) : 119-140.

Luong X. et Mellet S. (1995). Les calculs multidimensionnels au service de l'analyse syntaxique diachronique. In Bolasco S., Lebart L. et Salem A. (Eds), *Analisi statistica dei Dati testuali*. CISU : 281-288.

Luong X. et Mellet S. (2003). Mesures de distance grammaticale entre les textes. *Corpus,* vol. (2) : 141-166.

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

Pincemin B. (2002). Similarités texte – textes. Expérience d'une application de diffusion ciblée et propositions. In *Actas del segundo Seminario de la Escuela interlatina de altos Estudios en Lingüística aplicada, Matemáticas y Tratamiento de Corpus* (s.n) : 35-52.

Salem A. (2002). Topographie textuelle dans l'analyse quantitative des textes. In *Actas del segundo Seminario de la Escuela interlatina de altos Estudios en Lingüística aplicada, Matemáticas y Tratamiento de Corpus* (s.n.) : 53-59.