Distance intertextuelle et classement des textes d'après leur structure : méthodes de découpage et analyses arborées

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

[1]« Bases, Corpus et Langage » (UMR 6039), Univ. d'Angers – 49000 Angers – France

[2]« Bases, Corpus et Langage » (UMR 6039), CNRS – B.P. 3209, 06204 Nice cedex – France

[3]« Bases, Corpus et Langage » (UMR 6039), UNSA– B.P. 3209, 06204 Nice cedex – France

Abstract

The intertextual distance is often calculated by studying frequencies. As part of a research project on texts topology, we will evaluate here the efficiency of one of the methods we described in 2004. For the calculation of the intertextual distance, this method tries to take into account the text linearity by a text segmentation into a succession of *n* contiguous parts. By combining this method with tree analyses, we will study the distribution of a linguistic phenomenon (in this case, a tense) in the general framework of texts. Our long term aims are to evaluate the variability of linguistic phenomena correlation according to the parts of the text (introduction, text body, conclusion, etc) and to calculate distances between texts according to their structure. By studying a corpus based on a latin historical lemmatized texts databank, we will try here to establish the limits of the size heterogeneousness of the analysed texts and to evaluate the efficiency of the method according to the selected parameters.

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 évaluons ici la performance d'un des outils présentés aux JADT04 et développés pour prendre en compte, dans ce calcul, les localisations des unités au fil de la chaîne linéaire du texte. Par un découpage des textes en un même nombre de *n* de tranches contiguës et par le recours à l'analyse arborée, il s'agit d'examiner la répartition d'un même fait linguistique (en l'occurrence, un temps verbal) dans les différentes parties de chaque texte. Les objectifs à long terme sont de déterminer la variabilité des corrélations de traits selon les parties d'un même texte (introduction, développement, etc.) et de préciser les distances ou similarités entre les textes en fonction de leur structure. À partir de l'examen d'un corpus de textes historiques latins lemmatisés, on tentera ici de déterminer les limites à respecter dans l'hétérogénéité des tailles de textes et d'évaluer l'efficience de la méthode en fonction des paramètres d'analyse choisis.

Mots-clés : distance intertextuelle, classement générique, topologie textuelle, structures textuelles, découpage, analyses arborées, temps verbaux, latin.

1. Rappels

Lors des JADT 2004, à partir d'un corpus d'historiens latins, nous avions montré que la classification automatique endogène des textes gagnait à prendre en compte non seulement la fréquence globale du ou des paramètres d'analyse dans chacun des textes comparés, mais aussi leur distribution au fil de chaque texte – rejoignant par là d'autres observations, faites, par exemple, à propos de la variabilité des corrélations de traits selon les parties d'un même

texte (introduction, développement, etc.)[1]. La méthode que nous proposions consistait à découper chaque texte en un même nombre *n* de tranches successives et contiguës, et à dénombrer les occurrences d'un paramètre dans chacune des tranches, – paramètre dont on présumait que sa distribution pouvait être liée d'une manière ou d'une autre à la structuration du texte : il s'agissait alors des formes de parfait de l'indicatif, temps de la narration par excellence. Chaque texte se voyait ainsi affecter un profil caractéristique, formé par la succession ordonnée des *n* valeurs correspondant à la fréquence du parfait dans chacune de ses *n* tranches constitutives. Le programme de calcul de distance développé par Xuan Luong était alors appliqué à la matrice des profils, aboutissant à une représentation arborée radiale, donnant à voir les apparentements et les oppositions entre les textes.

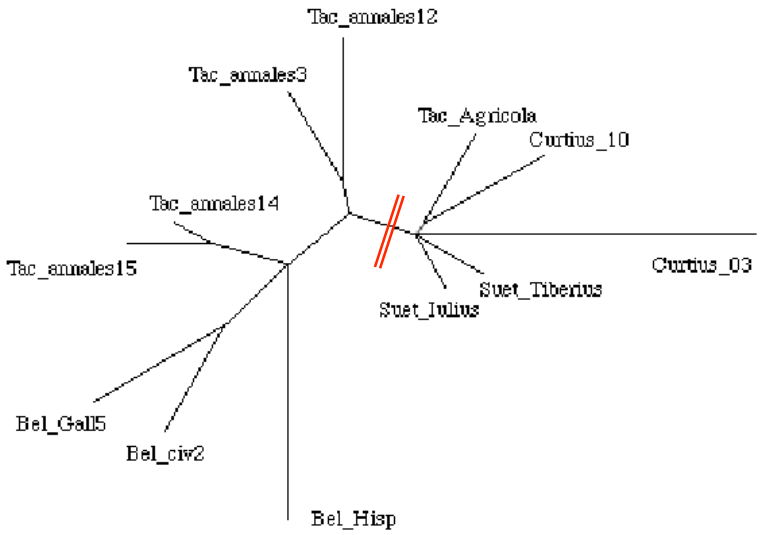
Un point important est à souligner : dans cette méthode, la détermination du nombre optimal de tranches pour découper les textes de la manière la plus pertinente possible est obtenue de façon empirique, par tests successifs ; elle ne repose sur aucun a priori linguistique. Pour notre corpus initial, nous avions retenu la division en 5 tranches, qui fournissait l'analyse arborée la plus fortement structurée[2]. Or le corpus était constitué de textes dont la taille (calculée ici en nombre de prédicats principaux – ou verbes de propositions principales) variait dans les limites restreintes d'un facteur 1,2 ($493 \leq N \leq 575$)[3]. Une question vient alors immédiatement à l'esprit : le même découpage et la même méthode d'analyse vaudraient-ils encore pour traiter un corpus plus hétérogène, où les tailles des textes varieraient plus sensiblement ? C'est le point que nous allons approfondir ici. En d'autres termes, quelle est la fiabilité de la méthode et quelles sont les caractéristiques textuelles auxquelles elle est sensible et qu'elle est susceptible de mettre au jour ? Par ailleurs, nous voudrions également tester la stabilité des résultats lorsqu'on modifie légèrement le paramètre utilisé pour représenter le texte et donner une image de sa structure. L'exposé se veut donc avant tout méthodologique.

2. Validité de la méthode pour des corpus comprenant des textes de petite taille

2.1. Limites de l'hétérogénéité acceptable : extension du corpus initial en direction de textes de petite taille

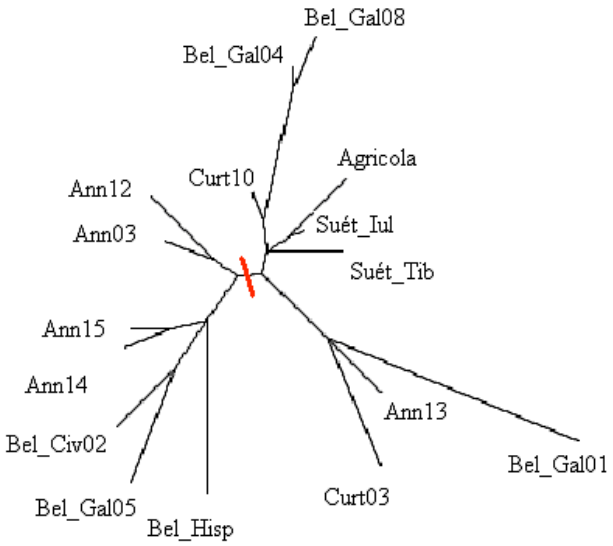
Les résultats obtenus jusqu'ici[4] sur un corpus d'œuvres de tailles homogènes offrent une classification remarquable, avec des regroupements sous-génériques très clairs qui associent sur un même nœud l'ensemble des œuvres d'inspiration biographique opposées aux œuvres plus proprement historiques (*Annales* de Tacite, regroupées deux par deux, et commentaires césariens) ; on note que l'impact du sous-genre est suffisamment fort pour détacher la *Vie d'Agricola* des autres œuvres de Tacite, et le rattacher notamment aux *Vies* de Suétone.

[1] Voir notamment Sueur, 1982 ; Biber & Finegan, 1994.
[2] On parle bien ici de la structuration de l'arbre, c'est-à-dire de la force et de la stabilité de ses regroupements, et non pas de l'adéquation qualitative de la classification aux attentes du philologue, laquelle doit être jugée *in fine* seulement, au moment de l'évaluation de la méthode.
[3] Bel_Civ_2 : 520 ; Bel_Gal_5 : 512 ; Bel_Hisp : 496 ; QC03 : 558 ; QC10 : 492 ; Tac_Ann3 : 541 ; Tac_Ann12 : 519 ; Tac_Ann14 : 531 ; Tac_Ann15 : 575 ; Tac_Agricola : 493 ; Sué_Tibère : 536 ; Sué_Iulius : 529. La recherche s'appuie sur le corpus de textes lemmatisés et étiquetés du Laboratoire d'Analyse Statistique des Langues Anciennes (LASLA) de l'Université de Liège.
[4] Cf. Longrée *et al.*, 2004 ; Longrée et Mellet, 2007.



Graphe 1

Qu'en est-il si on accroît le corpus et, par là-même, si on adjoint aux textes initiaux des textes de tailles plus variées ? Commençons par leur associer des textes sensiblement plus petits ($284 \leq N \leq 575$, variation d'un facteur à peine supérieur à 2), à savoir les livres 1, 4, 8 de la *Guerre des Gaules* ainsi que le livre 13 des *Annales* de Tacite. Chacun de ces livres est, comme les précédents, réduit à une série de codes représentant la succession de ses prédicats principaux, puis découpé en 5 tranches égales au sein desquelles on dénombre les occurrences du code de parfait de l'indicatif. Les profils ainsi obtenus sont intégrés au tableau rectangulaire qui sert de matrice pour le calcul des distances.



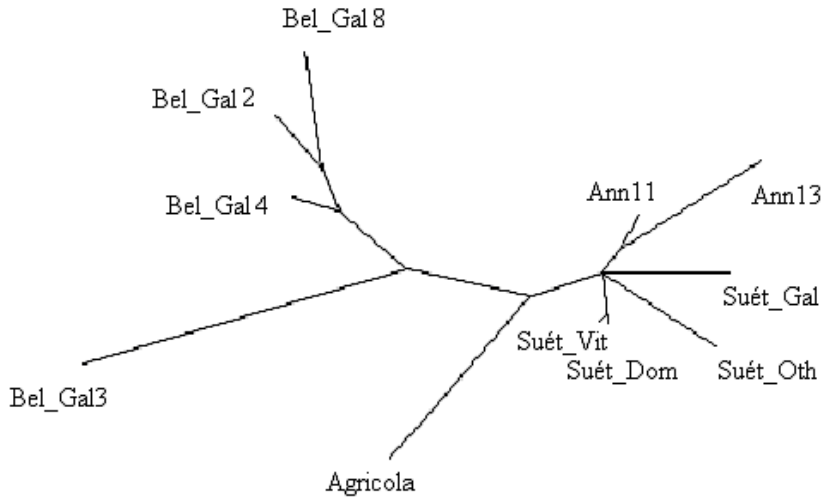
Graphe 2

L'attention se porte d'emblée sur le positionnement dans l'arbre des petits textes qui viennent d'être ajoutés au corpus : ceux-ci manifestent une tendance assez nette au regroupement, deux par deux : le livre 4 et le livre 8 de la *Guerre des Gaules* d'un côté et, moins proches, le livre 1 de la *Guerre des Gaules* avec le livre 13 des *Annales*. Par ailleurs ces deux regroupements se situent d'un même côté du nœud principal de l'arbre. Mais, à l'évidence, ce deuxième arbre est moins bien structuré que le précédent et ne permet plus une analyse en termes d'oppositions sous-génériques aussi claire. On observe en particulier deux phénomènes sur lesquels on sera amené à revenir : d'une part l'originalité de positionnement du livre 13 des *Annales* par rapport aux autres livres des *Annales* et l'éloignement important du livre 1 de la *Guerre des Gaules* manifestant un très mauvais rattachement à l'arbre, d'autre part le déplacement du livre 3 de Curtius dans un des groupes de petits textes. Nous laissons pour l'instant ces remarques en suspens[5].

Pour revenir au positionnement des petits textes intégrés au corpus, deux hypothèses explicatives viennent à l'esprit : ou bien c'est la seule hétérogénéité du corpus – en termes de taille des textes – qui est cause de cette relative marginalisation des petits textes, ou bien c'est la méthode elle-même – et notamment le découpage en 5 tranches – qui n'est pas adaptée au traitement de textes dont la longueur est inférieure à un certain seuil. Il convient donc d'appliquer la méthode proposée à un corpus composé ***exclusivement*** de petits textes pour pouvoir évaluer sa pertinence et son efficience dans un tel cadre.

2.2. Efficience de la méthode pour des textes de petite et très petite taille

Nous avons donc sélectionné un ensemble de textes[6] dont le nombre de prédicats principaux varie entre 95 et 493 et nous les avons traités de la même manière que les précédents. Voici l'arbre issu de cette analyse :



Graphe 3

[5] Voir §3.1.

[6] Il s'agit des livres 2, 3,4 et 8 de la *Guerre des Gaules*, des livres 11 et 13 des *Annales*, de la *Vie d'Agricola* par Tacite, des *Vies de Galba, Othon, Domitien* et *Vitellius* par Suétone.

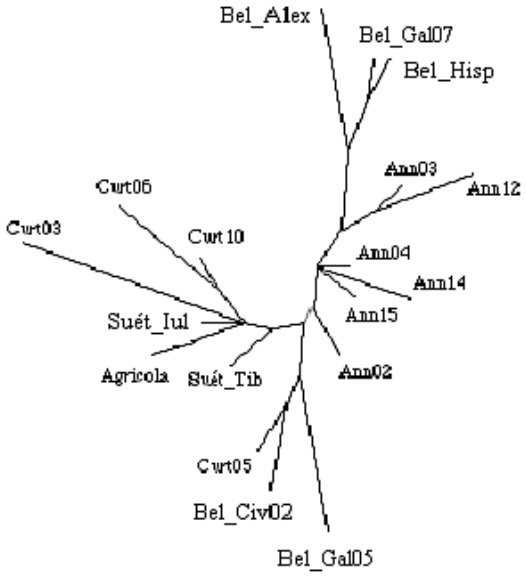
En dépit d'un taux de variation important dans les tailles des textes retenus (facteur 5 entre le texte le plus court et le texte le plus long), l'arbre offre une bonne structuration, avec des regroupements interprétables. On note simplement que la classification repose ici davantage sur des oppositions entre auteurs que sur des oppositions sous-génériques. Mais il est vrai que l'impératif de petite taille nous a conduit à intégrer au corpus deux textes qui mériteraient d'être traités avec prudence : les livres 11 et 13 des *Annales*, le premier parce qu'il s'ouvre sur une lacune de la tradition manuscrite, le second parce qu'il ouvre une héxade de l'œuvre (point sur lequel nous reviendrons). En outre ce même impératif limite sensiblement le nombre de textes du corpus, si bien que les oppositions sous-génériques en viennent à se superposer avec les oppositions d'auteurs et avec l'évolution chronologique. Il n'en reste pas moins que la méthode paraît pouvoir mettre en lumière ce qui devait l'être et que le découpage des textes en 5 tranches, en dépit de leur taille réduite, semble rester pertinent.

Nous allons passer maintenant à l'extension du corpus vers des textes de plus grande taille.

3. Validité de la méthode pour des corpus comprenant des textes de grande taille

3.1. Limites de l'hétérogénéité acceptable : extension du corpus initial en direction de textes de grande taille

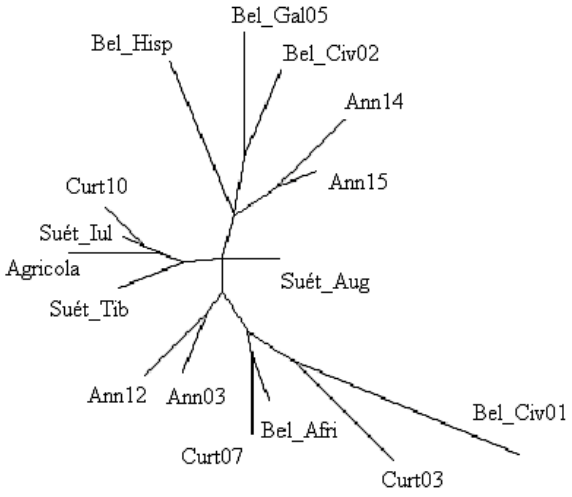
Nous repartons donc du corpus initial auquel nous ajoutons d'abord les textes suivants : livre 7 de la *Guerre des Gaules*, *Guerre d'Alexandrie*, livres 5 et 6 de Curtius, livres 2 et 4 des *Annales*. La longueur des textes (estimée, rappelons-le, en nombre de prédicats principaux) varie entre 493 et 797, soit d'un facteur 1,62 (graphe 4) ; puis, dans un deuxième test, nous poussons encore l'extension en revenant au corpus initial et en lui ajoutant des textes encore plus longs (N ≥ 800)[7], ce qui induit un facteur de variation supérieur à 1,8 (graphe 5).



Graphe 4

[7] A savoir le livre 1 de la *Guerre civile* (N = 869), la *Guerre d'Afrique* (839), le livre 7 de Curtius (889) et la *Vie d'Auguste* par Suétone (839).

À l'évidence, il existe un seuil d'hétérogénéité qu'on ne peut pas dépasser sous peine d'altérer la qualité des résultats. Dans le graphe 4 les regroupements restent majoritairement pertinents, alors que ce n'est plus le cas dans le graphe 5. Dans le graphe 4 en effet, nous avons un regroupement global des livres des *Annales*, une stabilité du positionnement de l'*Agricola* au sein d'un groupe solide de biographies[8] ; le corpus césarien, lui, se répartit sur deux branches opposées.



Graphe 5

Dans le graphe 5, les très grands textes manifestent une tendance au regroupement quel que soit leur auteur ou leur sous-genre, comme l'avaient déjà fait de leur côté les petits textes dans le graphe 2. On en conclut que les différences de taille importantes deviennent, avec cette méthode d'analyse, un critère de classification, au même titre que les différences d'auteurs ou de genres.

3.2. Limites de l'hétérogénéité acceptable : de quelques autres hétérogénéités spécifiques

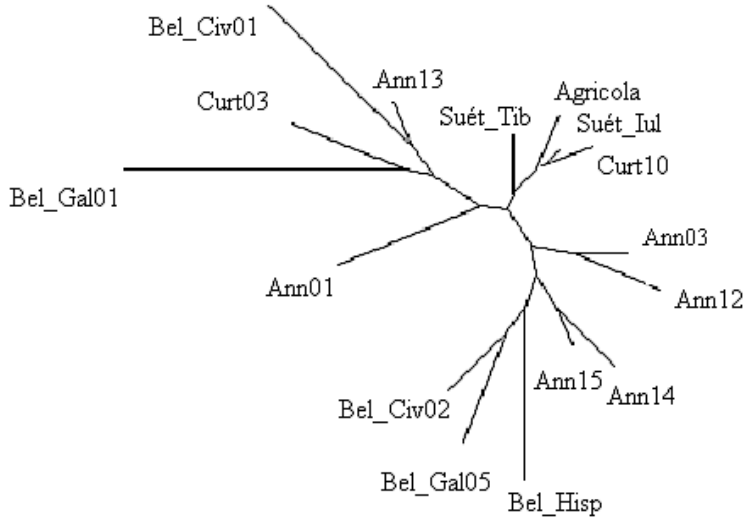
Une fois posé ce constat, en vérité assez prévisible, sur les problèmes liés à la taille des textes, voyons plus en détail un certain nombre de phénomènes.

La première observation est que le texte le plus éloigné de tous les autres et donc le moins bien intégré à l'analyse arborée du graphe 5 est le ***premier*** livre de la *Guerre civile*. Ceci nous renvoie au même phénomène déjà observé dans le graphe 2, cette fois pour le positionnement du ***premier*** livre de la *Guerre des Gaules*. Ces deux observations suggèrent que les premiers livres d'une œuvre pourraient avoir une structure particulière les différenciant assez sensiblement des autres livres de la même œuvre. Cette hypothèse est corroborée par le rattachement, dans le graphe 2, du livre 13 des *Annales* à un nœud portant aussi le livre 1 de la *Guerre des Gaules*. Or les *Annales* sont constituées de trois héxades (ou cycles narratifs constitués de six livres chacun que l'on peut considérer comme des entités relativement indépendantes, suscitant à chaque fois une reprise introductive, avec un récapitulatif des faits narrés dans la précédente héxade)[9]. Il semble donc que les premiers livres d'une œuvre ou

[8] Échappe à ce groupe le livre 5 de Curtius, qui devrait pourtant en faire partie, mais qui comporte une lacune pouvant expliquer son mauvais positionnement.

[9] Le livre 7 des *Annales* qui introduisait la deuxième héxade et qui aurait donc pu être considéré, lui aussi, comme un premier livre de cycle est perdu.

d'un cycle narratif échappent à la classification générique et même à la classification d'auteur. Il convient donc d'en vérifier de manière plus systématique le comportement. Nous proposons donc d'ajouter au corpus initial 4 premiers livres : le livre 1 de la *Guerre des Gaules* (Cés_Gal01), le livre 1 de la *Guerre civile* (Cés_civ01), le livre 1 des *Annales* (Ann01) et le livre 13 des mêmes *Annales* (Ann13) ; voici le résultat de la nouvelle analyse arborée :



Graphe 6

Il apparaît très clairement que, lorsqu'on intègre au corpus toute une série de livres qui démarrent une œuvre ou un cycle narratif suffisamment autonome, ceux-ci créent un nœud de classification spécifique sur lequel ils se regroupent.

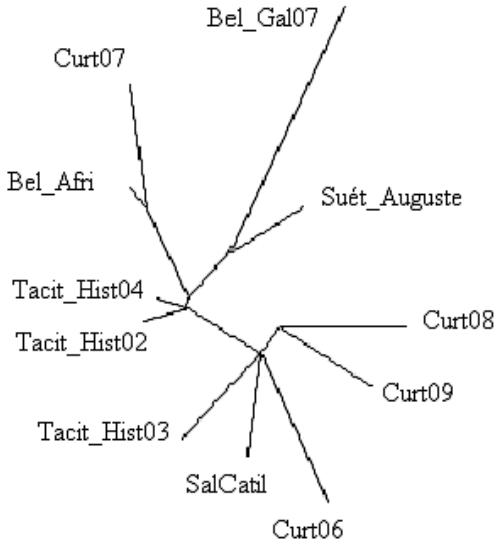
Nous observons par ailleurs le positionnement de Curtius03 dans ce groupe spécifique auquel il n'a apparemment aucune raison d'appartenir – pas plus qu'il n'avait de raison, précédemment, de se rattacher au groupe des très petits textes (graphe 2) ou à celui des très grands textes (graphe 5). On peut penser à une incohérence, qui, loin de remettre en cause la méthode de classification, suggère que celle-ci est apte à dévoiler la particularité d'un texte à travers ses comportements erratiques pourvu qu'on prenne la peine de multiplier les tests de classification et d'interpréter en profondeur les analyses obtenues. C'est en effet ici l'occasion de rappeler que, si une lecture rapide de l'arbre en fonction des distances entre ses feuilles et des structurations dessinant ses branches suffit généralement pour avoir une assez bonne évaluation des regroupements textuels pertinents, c'est en réalité la dynamique de construction de l'arbre qui apporte les enseignements les plus riches et les plus fiables. L'algorithme qui gère cette construction dynamique commence par opérer les regroupements les plus stables, et remplace aussitôt chaque ensemble d'éléments ainsi regroupés par son nœud de regroupement (et non par une valeur moyenne comme on le pratique le plus souvent dans les arbres plantés). Les regroupements successifs peuvent donc être amenés à traiter des nœuds intermédiaires avant même d'avoir épuisé l'ensemble des éléments initiaux. Par ailleurs, le graphe représentant non pas seulement les distances, mais aussi les structures de classification, il a la propriété d'accepter toutes les bipartitions possibles : où qu'on le coupe, la bipartition est pertinente.

Dans le cas qui nous occupe (graphe 6), le script de construction de l'arbre montre que le premier regroupement effectué est précisément celui qui rassemble le livre 3 de Curtius et le livre 1 de la *Guerre des Gaules* (et ce, avec un taux d'agrégation de 0,967 ce qui est un très bon taux pour un premier regroupement). Viennent ensuite quatre autres regroupements d'éléments initiaux (dont celui du livre 1 de la *Guerre civile* avec le livre 13 des *Annales*), puis laissant de côté les textes restants – encore au nombre de 6 – l'algorithme établit le nœud secondaire auquel se rattachent la branche portant Curt03 et Gall01 et celle portant Civ01 et Ann13. C'est dire la force d'attraction qui rapproche les livres dont le point commun est d'ouvrir une œuvre. La présence du livre 3 de Curtius dans ce groupe doit donc recevoir une interprétation. C'est alors au philologue ou au stylisticien de prendre le relais pour comprendre les raisons de ce comportement particulier. On pourrait simplement expliquer ce regroupement étonnant par la perte des quelques premiers chapitres de ce livre, perte qui rapprocherait son profil de ceux de livres ouvrant une oeuvre, mais on est également légitimé, nous semble-t-il, à s'interroger sur la tradition philologique elle-même qui, sur des bases manuscrites très succinctes, affirme, sans jamais remettre l'assertion en cause, qu'il s'agit là du troisième livre d'une œuvre dont les deux premiers ont été perdus. Or peu d'éléments garantissent cette interprétation de la lacune initiale de l'œuvre de Quinte-Curce. On pourrait aussi bien envisager que seuls quelques chapitres ont été perdus et donc que ce soi-disant livre 3 soit en fait un livre 1 étêté[10] ; auquel cas, son positionnement dans les différents arbres redeviendrait compréhensible puisque, que ce soit dans le graphe comprenant les petits textes ou dans celui comprenant les très grands textes, il s'est toujours rattaché rapidement à d'autres premiers livres.

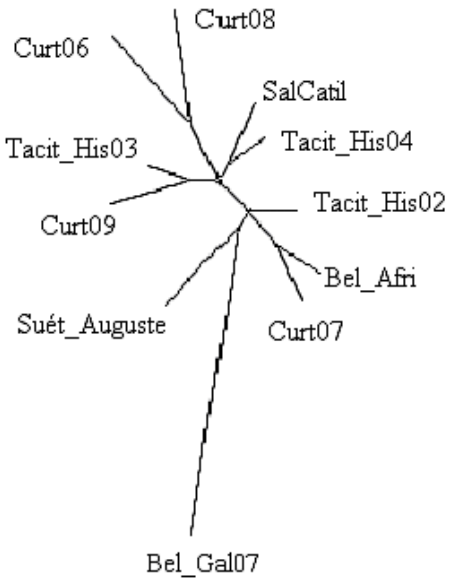
3.3. Efficacité de la méthode pour des textes de très grande taille

Comme pour les petits textes, la faible structuration du graphe 5 pose la question de savoir si seule l'hétérogénéité du corpus est en cause ou si, plus profondément, la méthode de découpage des textes en 5 parties est pertinente pour ceux dont la taille excède un certain seuil. Comme précédemment par conséquent, nous allons tester notre analyse sur un corpus composé exclusivement de grands textes ($732 \leq N \leq 976$) et en excluant, bien sûr, les premiers de chaque œuvre considérés désormais comme éléments perturbateurs.

[10] Il y aurait lieu de s'interroger plus globalement sur le découpage de l'oeuvre en livres de tailles très inégales et sur la taille des lacunes existant entre les livres 5 et 6 et dans le livre 10 ; l'œuvre pourrait fort bien ne pas avoir commencé avec le couronnement d'Alexandre en 336, mais seulement avec le début de l'expédition vers la Perse à la fin de l'année 334 ; il s'agirait alors de déterminer quel pourrait avoir été le nombre exact de livres composant l'œuvre (8, 10 ou 2 x 6).



Graphe 7



Graphe 8

La structuration est, on le voit (graphe 7), moins bonne que pour des textes de taille moyenne et l'interprétation des oppositions n'est pas aussi aisée. On peut donc légitimement penser que le découpage en 5 parties est peut-être moins adapté à cette taille de textes dont la structure, se déployant plus largement, peut davantage se complexifier et se diversifier. Divers tests portant sur des découpages en 6 ou 7 parties (graphe 8) n'ont pas donné de résultats réellement meilleurs, peut-être parce que le vaste espace narratif qui s'ouvre à l'historien lui laisse plus de marges de manœuvre, plus de liberté pour mener son récit ; le cadre narratif étant ainsi moins contraignant, les structures utilisées sont moins récurrentes et ne fournissent plus la base suffisante à la détection de parentés structurelles entre les textes et à leur classification

sur la base du paramètre d'analyse retenu. Ce qui nous amène à terminer notre exposé en évoquant rapidement la question du choix du paramètre pertinent.

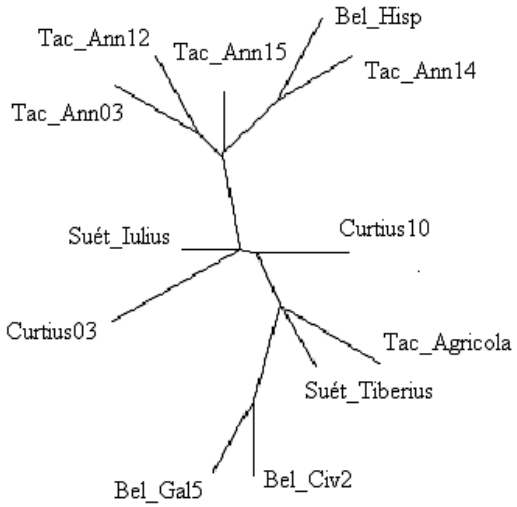
4. Le choix du paramètre d'analyse

4.1. Prédicats principaux vs ensemble de tous les prédicats (principaux et subordonnés)

Jusqu'ici le paramètre utilisé pour représenter les textes était, nous l'avons dit, le temps des prédicats principaux. Que se passe-t-il si l'on change de paramètre, ne serait-ce que de manière apparemment infime, c'est-à-dire en réintroduisant dans l'analyse l'ensemble des prédicats, principaux et subordonnés ?

Bien évidemment, et en premier lieu, la taille des réductions textuelles sur lesquelles nous travaillons (séries de codes verbaux) est modifiée : les chaînes de codes verbaux représentant la structure du texte sont multipliées par un facteur variant entre 2 et 3. Les rapports de taille entre les textes peuvent ainsi changer[11] : un auteur à l'écriture complexe, aux phrases longues, verra la taille de son texte augmenter plus sensiblement que celui dont l'écriture est plus simple et utilise moins volontiers les propositions subordonnées. Par exemple, dans l'*Agricola* réduit à ses seuls prédicats principaux, la première partie d'un découpage en 5 tranches égales s'achève au premier paragraphe du chapitre 10 alors que, dans la réduction à l'ensemble des prédicats, la première partie s'achève au premier paragraphe du chapitre 11. Il y a donc un chapitre de décalage.

Parallèlement, la distribution des parfaits de l'indicatif au sein de la nouvelle partition des textes en 5 tranches est aussi modifiée. Ceci est lié à l'existence de zones de parataxe plus ou moins intense ou, au contraire, au cumul de phrases complexes. Ainsi, pour l'*Agricola*, le profil de distribution des parfaits dans chacune de ces partitions varie de la façon suivante : on obtient la série 55-47-36-26-47 pour le premier cas de figure et 62-41-34-28-46 pour le second. Le reparamétrage donne lieu à des profils textuels assez différents, entraînant de nouveaux regroupements de textes, comme on peut le voir sur le graphe 9.



Graphe 9

[11] Seulement trois textes ne changent pas de rang, dont le plus petit (*Vie d'Agricola*) et le plus grand (*Annales* 15). Curtius_03 manifeste encore son originalité en passant du rang 11 au rang 6, les autres modifications étant moins importantes.

4.2. Les corrélations de traits

Par ailleurs, nous sommes tout à fait conscients qu'un paramètre n'entre pas seul en jeu dans la structuration des différentes parties d'un récit. A priori, l'usage du temps narratif par excellence, le parfait de l'indicatif, devait être révélateur, et c'est bien ce que nous ont confirmé nos différents tests. Mais il serait intéressant de pouvoir prendre en compte d'autres traits linguistiques corrélés à l'usage des temps verbaux. Nous avons tenté une première extension du nombre de paramètres, sans résultats significatifs pour l'instant : il s'avère en effet que l'usage des temps d'arrière-plan tels que l'imparfait et le plus-que-parfait est beaucoup moins structurant que celui du parfait[12] ; par ailleurs l'emploi de certains types de subordonnées, pourtant globalement caractéristique de l'écriture d'un auteur, ne détermine pas non plus la structuration d'un récit en différents épisodes – en tout cas pas de manière récurrente et pas en lien avec un style d'auteur ou un sous-genre littéraire : il est vrai en effet que l'arrivée massive dans le récit de subordonnées au subjonctif ou de propositions infinitives signale un passage contenant des discours rapportés en style indirect, mais de tels passages ne semblent pas avoir de place prédéterminée dans la structure du récit.

Pour approfondir cette hypothèse des corrélations de traits sur un corpus comme le nôtre, c'est-à-dire dont les textes ne sont quand même pas aussi structurés qu'un article scientifique contemporain ou un conte traditionnel, il faudrait pouvoir relever et dénombrer des traits linguistiques aussi subtils que ceux qui portent l'effacement énonciatif du locuteur ou, à l'inverse, manifestent son engagement assertif et sa posture de commentateur. On se heurte alors aux limites bien connues de l'encodage des textes : ces traits sont souvent trop fugaces ou, du moins, s'inscrivent dans des formes lexicales ou morpho-syntaxiques trop variées pour pouvoir faire l'objet d'un traitement automatique.

5. Conclusion

La méthode proposée ici demande à être testée plus largement : sur d'autres corpus et en mesurant d'autres paramètres. Par ailleurs, il faudra tenter d'évaluer de manière plus formelle en quoi le facteur de taille contribue à créer les regroupements entre textes, ce qui permettrait d'affiner l'algorithme de calcul de distance. Dans son état actuel, l'exploration méthodologique que nous venons de présenter nous semble toutefois aboutir à quelques conclusions intéressantes. D'une part, la méthode du découpage des textes en 5 tranches, – nombre apparemment arbitraire d'un point de vue linguistique, mais retenu au terme d'une série de tests procédant par approximations successives –, s'avère fiable et efficace : elle fournit des profils de structure textuelle qui, soumis à l'analyse arborée, donnent lieu à des rapprochements ou des oppositions en général relativement stables et toujours interprétables. D'autre part, on a pu constater que cette méthode d'analyse et de visualisation permettait de rendre compte successivement des nombreuses variables qui peuvent affecter la plus ou moins grande proximité entre les textes d'un corpus : les regroupements par auteurs sont bien sûr mis en valeur ; ils sont souvent transcendés par des faits de structures spécifiques, tels que ceux qui affectent les premiers livres d'une œuvre et, plus largement, par les apparentements génériques ou sous-génériques ; mais ils laissent aussi la place au critère de taille, qui devient prédominant lorsque celle-ci introduit une hétérogénéité trop grande dans le corpus. L'analyse

[12] Cf. Juillard *et al*., soumis.

arborée est bien une analyse multi-dimensionnelle, particulièrement performante, et nous attribuons cette qualité au fait que son algorithme n'est pas celui d'une classification hiérarchique traditionnelle, mais qu'il repose entièrement sur une approche topologique.

Références

Biber D. & Finegan E. (1994). Intra-textual variation within medical research articles. In Oostdijk N. & de Haan P. (eds), *Corpus-based Research into Language*. Amsterdam : Rodopi : 201-222.

Bronckart J.-P. (1996). Genres de textes, types de discours et opérations discursives. *Enjeux*, 37 / 38 : 31-47.

Juillard M. & Luong X. (2001). On Consensus between Tree-Representations of Linguistic Data. *Literary and Linguistic Computing*, 16, 1 : 59-76.

Juillard M., Longrée D., Luong X. & Mellet S. (soumis). Methods in Linguistic Topology. *Literary and Linguistic Computing*.

Lamalle C. & Salem A. (2002). Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels. In Morin A. & Sébillot P. (éds.), *JADT 2002*, (actes des 6èmes Journées internationales d'Analyse statistique des Données Textuelles), Saint-Malo, IRISA et INRIA, vol. 2 : 403-411.

Longrée D., Luong X. & Mellet S. (2004). Temps verbaux, axe syntagmatique, topologie textuelle. analyse d'un corpus lemmatisé. In Purnelle G., Fairon C. & Dister A. (éds), *Le poids des mots* (actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles – JADT 2004). Louvain, P.U. de l'U.C.L. : 743-752.

Longrée D. & Luong X. (2005). Spécificités stylistiques et distributions temporelles chez les historiens latins : sur les méthodes d'analyse quantitatives d'un corpus lemmatisé. In Williams G. (éd.), *La linguistique de corpus*. Rennes, P.U. de Rennes : 141-152.

Longrée D. & Mellet S (2007, à paraître). Temps verbaux et prose historique latine : à la recherche de nouvelles méthodes d'analyse statistique. In Denooz J. et Purnelle G. (éds), *Ordre des mots et cohérence, Communications du 13e Colloque international de Linguistique latine ICLL- Bruxelles - 4-9 avril 2005,* Liège, P.U. de Liège.

Sueur J.-P. (1982). Pour une grammaire du discours : élaboration d'une méthode ; exemples d'application. *Mots*, 5 : 145-185.