Problèmes d'attribution : application de quelques tests statistiques à différents historiens latins, analyse arborée

M. DUBROCARD et X. LUONG
Université de Nice-Sophia Antipolis et
CNRS *Bases, Corpus et Langage* UPRESA 6039

Introduction : Problèmes d'attribution et tracé d'arbres

Voilà tout juste quarante ans, un des pères fondateurs de la statistique linguistique, Charles Guiraud, souhaitait voir se constituer une "caractérologie du langage" susceptible de mettre en évidence des traits caractéristiques, propres à un auteur et même à une œuvre[1].

Depuis cette époque déjà lointaine, les efforts des chercheurs, secondés par des moyens techniques que l'on n'aurait même pas pu, alors, imaginer, ont permis d'enrichir considérablement notre connaissance des caractères statistiques des textes.

Il reste cependant que bien des questions que posait Charles Guiraud n'ont pas encore reçu de réponse vraiment satisfaisante. C'est le cas, en particulier, des problèmes épineux liés à l'identification d'un auteur, ou au choix entre plusieurs auteurs possibles.

À qui attribuer tel ou tel ouvrage ? L'*Iliade* et l'*Odyssée* sont-elles l'œuvre d'un auteur unique ? Combien de rédacteurs différents ont-ils contribué au *corpus Hippocraticum* ?

Avant même d'aborder de tels problèmes, il convient de se demander s'il est possible de reconnaître, chez un auteur déterminé et parfaitement identifié, des traits caractéristiques qui permettraient de le distinguer à coup sûr d'autres écrivains proches de lui par leur langue et leur inspiration.

Pour répondre à cette question, nous disposions d'un instrument mathématique bien adapté, lié aux travaux que nous avons récemment consacrés à la topologie discrète sur un arbre. Nous avons tenté de l'appliquer à quelques données empruntées à des historiens latins. Il en résulte, semble-t-il, un éclairage nouveau sur les notions de filiation, de catégorie, de typicalité qui émergent de l'architecture des représentations arborées.

[1] Ch. Guiraud, *Problèmes et méthodes de la statistique linguistique*, Dordrecht, 1959, p. 25 sqq.

I. Le choix des données

Nous avons donc imaginé l'expérience suivante : nous avons réuni quelques historiens latins, Salluste, César, Tite-Live, Quinte-Curce, Tacite et Suétone.

Nous avons prélevé, dans chaque auteur, un ou plusieurs fragments d'une longueur de 4 000 mots (quatre fragments pour César, Tite-Live et Tacite, deux pour Salluste, Quinte-Curce et Suétone, un pour chacune des *Guerres*), soit, au total, 18 textes de même longueur. Le nombre de fragments retenus est proportionnel à la longueur des œuvres, exception faite pour César, que nous voulions pouvoir opposer ultérieurement à ses continuateurs de la *Guerre d'Alexandrie*, de la *Guerre d'Afrique* et de la *Guerre d'Espagne*, en lui accordant une représentation suffisante.

Les fragments ont été déterminés en fixant leur début de façon aléatoire. Cependant nous avons veillé à ce que les grandes divisions de l'œuvre soient respectées. Ainsi *Catilina* et *Jugurtha* ont-ils fourni un texte chacun, les *Commentaires sur la Guerre des Gaules* et *sur la Guerre civile*, chacun deux textes, de même que les *Histoires* et les *Annales*. Il en a été de même pour les autres auteurs. Ainsi les quatre fragments empruntés à Tite-Live figurent-ils respectivement dans le premier, le second, le troisième et le quatrième quart de l'*Histoire*.

Ces textes ont été empruntés au CD Rom édité par *the Packard Humanities Institute*[2]

La seule modification qui leur a été apportée a consisté à unifier leur présentation typographique, en confondant, pour tous les textes, les lettres *u* et *v*.

Les textes n'étant pas lemmatisés, toutes nos analyses ont porté sur les formes.

Nous avons établi, pour chacun des 18 textes ainsi que pour l'ensemble qu'ils constituent, un index des

[2] *Thesaurus Linguae Latinae "PHI"*, édité par the Packard Humanities Institute

formes, ainsi que des classements tenant compte de l'effectif de chaque forme, dans l'ordre alphabétique et dans l'ordre de fréquence décroissante.
Nous avons retenu cinq tests statistiques :
- longueur des formes
- emploi des graphèmes à l'initiale de la forme
- emploi des graphèmes dans l'ensemble de la forme
- fréquence d'emploi des formes
- répartition des classes de fréquence.
À chacun de ces indices correspond un tableau de 18 lignes, correspondant aux 18 textes. Le nombre des colonnes dépend de la variable observée.
Par ailleurs nous avons effectué un calcul de khi-2 sur chaque colonne, afin d'apprécier la dispersion des résultats obtenus. Nous avons pu ainsi opérer une sorte de filtrage, et ne retenir que les caractères qui faisaient le mieux apparaître l'hétérogénéité du corpus. Ces tableaux, dits "filtrés", écartent les données dont les variations, d'un texte à l'autre, pouvaient avoir une origine aléatoire. Le seuil de probabilité choisi a été de 1/1 000.
Ont été ainsi "filtrés" les tableaux I, II, IV et V. Le tableau III n'a pas fait l'objet d'un filtrage, tous les khi-2 étant largement supérieurs au seuil de signification choisi.
Enfin, nous avons réuni dans un tableau global (tableau VI) les éléments qui présentaient les khi-2 les plus élevés, à quelque caractère statistique qu'ils correspondent.
Après cette première expérience, dont les résultats seront analysés plus loin, nous avons tenté d'aller un peu plus avant dans notre recherche en ajoutant à notre corpus trois textes écrits par des continuateurs de César, les *Commentaires sur la Guerre d'Alexandrie*, *sur la Guerre d'Afrique* et *sur la Guerre d'Espagne*. Ces textes sont très proches de la *Guerre des Gaules* et de la *Guerre civile*, même si César n'en est pas l'auteur. Il sera intéressant de voir si nos différents tests permettent de les isoler du reste du corpus césarien. Nous reprendrons donc les six tableaux déjà utilisés, en leur ajoutant chaque fois trois lignes supplémentaires correspondant aux trois nouveaux textes. Les graphes ainsi obtenus seront analysés à leur tour.

II Représentations arborées

1. Choix d'une méthode d'analyse multivariée.

La figure 1 montre ce qu'on obtient en calculant une représentation approchée du tableau I (18 textes, critères : longueur des mots)**,** dans un espace euclidien de dimension 2 (autrement dit à l'aide du "*multidimensional scaling*"). Nous utilisons ici la distance du khi-2, bien adaptée sur des données provenant des descripteurs très divers.

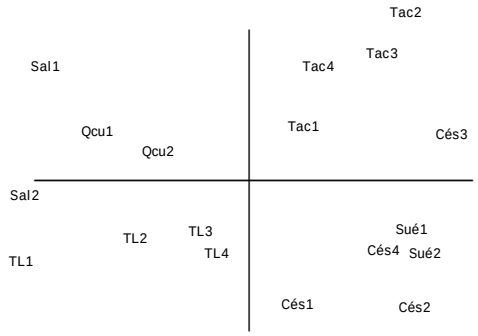


Figure 1 : longueur des formes. *Multidimensional scaling.*

Cependant, le thème de notre travail suggère une représentation des textes qui ferait plutôt appel à des idées de filiation qu'à des emplacements qu'ils occuperaient dans un éventuel espace subjectif.
La figure 2 est une classification ascendante hiérarchique du même ensemble des données. C'est un modèle répandu en taxinomie. Un sommet appelé racine est relié aux feuilles qui sont toutes équidistantes de cette racine. Tous les sommets d'une

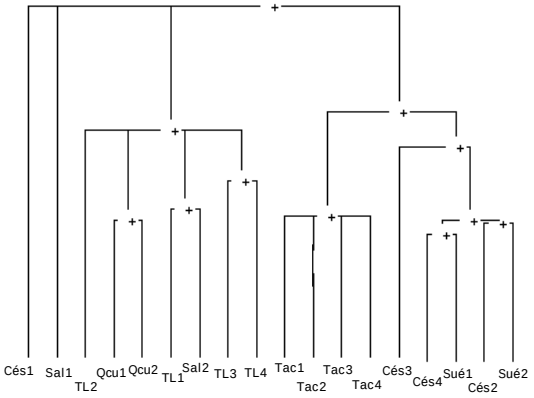


Figure 2 : longueur des formes. Classification ascendante hiérarchique.

même classe, c'est-à-dire dominés par un même nœud, sont équidistants entre eux. Ce modèle ultramétrique présente au moins deux avantages :
- la notion de classe y est définie sans ambiguïté : on passe de classe en classe, de la plus grossière aux plus fines en parcourant l'arbre de la racine aux feuilles.
- la distance d'une feuille à un nœud fournit un indice du niveau de formation d'une classe.

C'est justement cette notion rigide de classe qui nous fait préférer au modèle ultramétrique un modèle plus propre à rendre compte de la ressemblance et de l'air de famille[3]. C'est dans cet esprit qu'a été calculé l'arbre de la figure 3. L'arbre est non planté : aucun de ses éléments n'est privilégié comme la racine de l'arbre. Aux feuilles de cet arbre sont accrochés les 18 textes. Il s'agit là de ce que nous appelons une *représentation arborée* de nos données. L'algorithme utilisé pour cette représentation est de nature topologique. Il a été créé par un des co-auteurs de cet article.

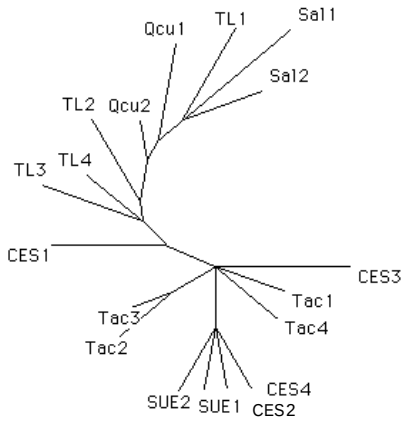


Figure 3 - Arbre I, longueur des formes.
Représentation arborée.

2. Topologie discrète sur un arbre revisité.

Nous énonçons quelques propriétés topologiques des arbres. Elles permettent la compréhension de ce modèle de représentation et sont à la base de nos constructions algorithmiques.

Soit un ensemble d'éléments X muni d'une distance d.

Condition *des 4 points*.
La condition nécessaire et suffisante pour qu'une distance d sur un ensemble X soit une distance arborée est :
pour tout x,y,z,t de X
d(x,y)+d(z,t) <= Max {d(x,t)+d(y,z), d(x,z)+d(y,t) }

Découverte dans les années soixante (Zaretskii, 1965 ; Simoes-Pereira, 1969 ; Buneman 1971 ; etc…, cf Barthélemy & al 1991) elle caractérise un arbre dont les feuilles représentent X (on dit un *X-arbre*) et permet le développement des méthodes de construction arborée.

Propriété 1. Quatre sommets sur un arbre sont toujours dans une configuration en H de la figure 4.

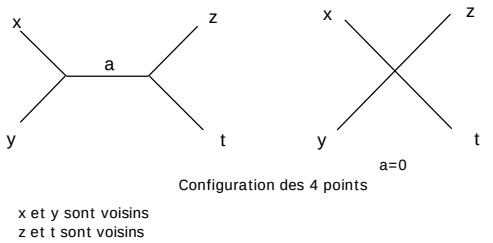


Figure 4 -

Elle exprime de manière imagée la condition précédente : on vérifie facilement que les distances lues sur une configuration en H vérifient cette condition. En cas d'égalité dans la condition des 4 points, le H à gauche de la figure 4 est réduit en une étoile, à droite.

Notre construction d'arbre est basée sur une nouvelle notion de voisinage.

Définition 1. k feuilles d'un X-arbre sont voisins si elles sont toutes reliées à un même nœud intérieur de degré k+1. On note cette relation de voisinage par V.

On remarque que V est une relation d'équivalence dans X.

Définition 2. Un ensemble de voisins est appelé un groupement.

Propriété 2. Deux éléments d'un groupement sont toujours voisins dans n'importe quelle configuration en H où ils se trouvent.

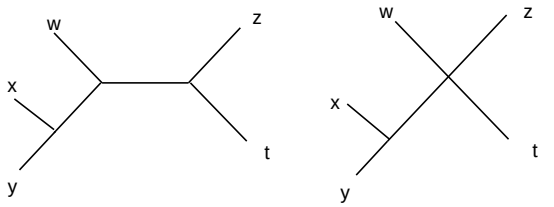
La recherche des éléments voisins revient à examiner toutes les configurations en H. On note dans chaque H les 2 paires de voisins et on dénombre *les scores*, c'est-à-dire le nombre de fois que deux feuilles sont des voisins dans les configurations en H. Dans un groupement, le score d'une paire de ses éléments est maximum et égal à (n-1)(n-2)/2, n étant le nombre de feuilles de l'arbre.

Propriété 3 . Soient x et y deux voisins et w,z,t trois éléments quelconques. On considère les configurations en H {x, w,z,t } et {y,w,z,t}. x et y se comportent, au point de vue de la topologie induite par

[3] WITTGENSTEIN, L (1953) *Philosophical investigations*, New York: Macmillan.

cette notion de voisinage, toujours de la même manière par rapport à w,z et t .

Les démonstrations des propriétés 2 et 3, à partir de la condition des 4 points, sont assez techniques. Ces propriétés se lisent plus facilement à l'aide de la figure 5.



configuration de deux voisins x et y avec 3 éléments quelconques w,z,t

Figure 5 -

Algorithmes.

Soit un X-arbre A muni d'une distance d et soit V la relation de voisinage définie précédemment. L'algorithme de reconstruction d'arbre s'énonce ainsi :

Algorithme
Étape 1 : Examiner toutes les configurations en H de l'arbre A, dénombrer les scores. Déterminer les groupements.
Étape 2 : Représenter chaque groupement par un seul élément. Cela revient à faire une relation d'équivalence A/V . Ces groupements sont des feuilles d'un nouvel arbre. Le noter par A :=A/V . ***Si*** A est réduit à un point, ***fin*** de l'algorithme ***sinon*** aller à *Étape 1*.

Lorsque l'on a un tableau de mesures de dissimilarité, on utilise un algorithme de représentation arborée s'inspirant directement de la construction précédente. On prendra
- pour groupement les éléments de scores maximum,
- pour la relation d'équivalence : x et y sont équivalents s'ils sont dans un groupement.

Remarques : Sous la forme générique que nous avons choisie, nos algorithmes rendent compte (modulo des choix de V et de la relation A/V) de pratiquement n'importe quel algorithme de classification ascendante hiérarchique. Mais contrairement à la classification qui réévalue les distances à chaque fusion, nos algorithmes progressent à chaque étape sans transformation notable des données, car d'après la propriété 3, lorsque l'on représente les éléments d'un groupement par un seul élément, on ne modifie pas les configurations en H de l'arbre A/V. L'algorithme s'appuie à chaque étape sur la topologie induite par la relation de voisinage. Construire un arbre revient à dégager les voisinages successifs de ses éléments.
Cela explique aussi les bonnes mesures statistiques de nos représentations arborées, par exemple les coefficients de corrélation des arbres présentés ici sont égaux en moyenne à 0.92, alors que ceux de la classification ascendante hiérarchique sont à 0,73.

Groupements.

C'est parce qu'elle s'oppose, via les configurations en H à (presque) toutes les autres paires d'objet qu'une paire x,y va fusionner. Les deux exigences, souvent contradictoire, de la classification automatique : homogénéité intra-classes et séparation inter-classes se trouvent ainsi réconciliées.
Par ailleurs la variabilité des longueurs d'arêtes au sein d'un même groupement permet de rendre compte des phénomènes de typicalité ou de représentativité[4].

III. Interprétation des résultats

· **Analyse comparée de six historiens latins**

Les résultats de nos analyses apparaissent sur les arbres numérotés de I à X.

1. Longueur des formes

L'arbre I (figure 3) présente le tracé obtenu à partir du premier tableau de données, c'est-à-dire la mesure de la longueur des formes. Le regroupement des quatre fragments de Tacite et de César, des deux fragments de Quinte-Curce, de Suétone et de Salluste est satisfaisant. En revanche, on observe un certain éparpillement des fragments de Tite-Live (seuls les fragments 3 et 4, empruntés à la deuxième moitié de l'œuvre, sont réunis), et de César.
Le filtrage, effectué suivant la méthode indiquée plus haut, améliore un peu ces résultats. Trois des fragments de Tite-Live sont maintenant regroupés, de même que les fragments de César.

2. Emploi des graphèmes à l'initiale

L'arbre II (figure 6-a) repose sur l'emploi des graphèmes à l'initiale de la forme. En l'absence de filtrage, le résultat est plutôt décevant, puisque seuls

[4] BARTHELEMY, J.P. (1993) "Similitude, arbres et typicalité", in *Sématique et Cognition* (D.Dubois, ed), Paris: Edition du CNRS, Coll. Sciences du Langage **31,** 205-224

quelques textes sont regroupés (Tite-Live 3 et Tite-Live 4, d'ailleurs séparés de Tite-Live 1 et 2, les deux textes de Salluste, deux des textes de Tacite). Les affinités et les disparités relevées entre les autres textes ne peuvent recevoir aucune explication satisfaisante.
Après filtrage (figure 6-b) le résultat est un peu meilleur. Tite-Live, Quinte-Curce et Salluste retrouvent leur unité. Les quatre fragments de Tacite sont regroupés, mais mêlés aux extraits de Suétone. Trois seulement des textes de César sont réunis.
On doit donc admettre que le critère retenu n'est pas d'une grande efficacité pour distinguer entre eux les auteurs de notre corpus.

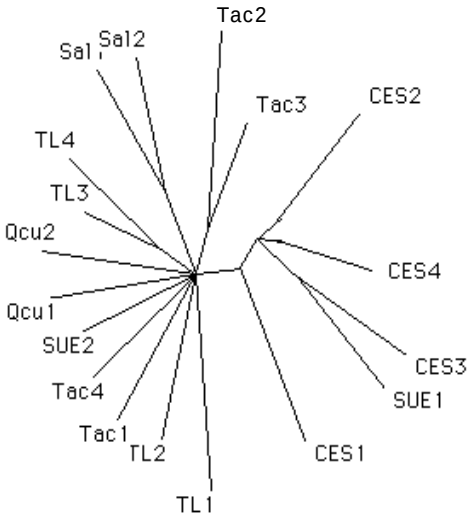


Figure 6-a : Arbre II, emploi des graphèmes à l'initiale de la forme

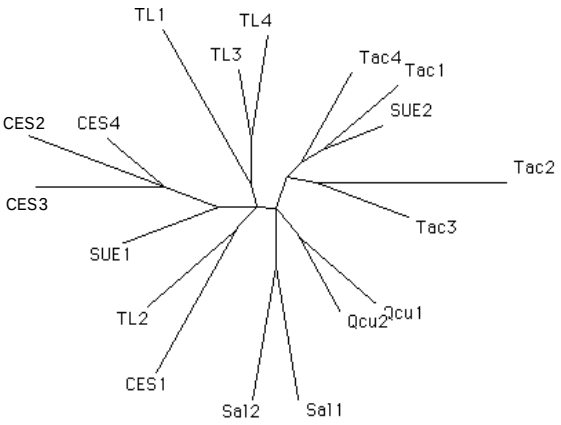


Figure 6-b : Arbre II-b, emploi des graphèmes à l'initiale de la forme, après filtrage

3. Emploi des graphèmes dans l'ensemble de la forme

L'arbre III (figure 7) présente les résultats obtenus en comparant l'utilisation des graphèmes dans l'ensemble de la forme, et prend en compte chacun de nos textes, après filtrage.
Les textes de Tacite sont réunis, les textes de César sont regroupés deux à deux, avec une opposition *Guerre des Gaules* (CES1, CES2), *Guerre civile* (CES3, CES4). Trois des textes de Tite-Live sont réunis, Tite-Live 1, emprunté au début de l'œuvre, restant isolé.

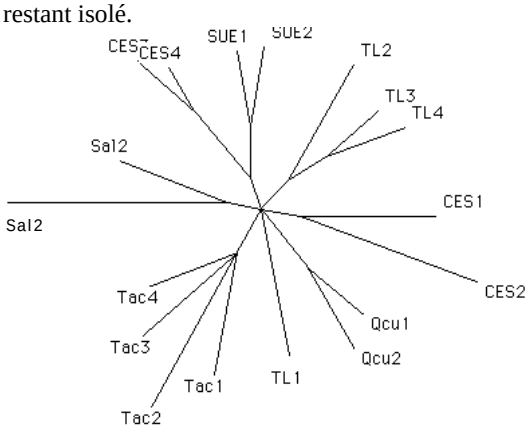


Figure 7 : Arbre III, emploi des graphèmes dans l'ensemble de la forme

4. Emploi des formes

L'arbre IV (figure 8) a été établi à partir de la fréquence d'emploi des formes. Dans un fichier global, réunissant les 21 fragments, les formes ont été classées par ordre de fréquence décroissante. Les 60 formes les plus fréquentes ont été retenues. On a ensuite décompté les occurrences de ces formes dans chacun des fragments.

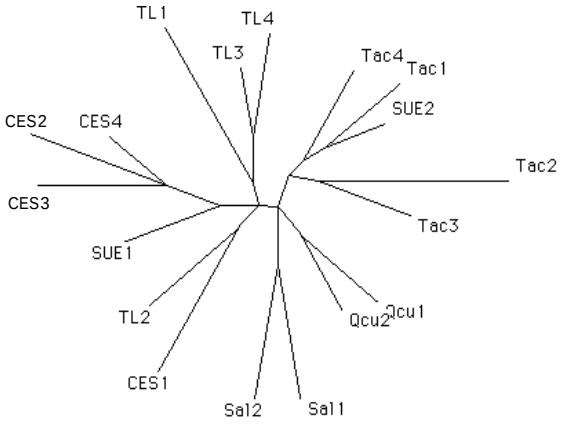


Figure 8 : Arbre IV, fréquence d'emploi des formes

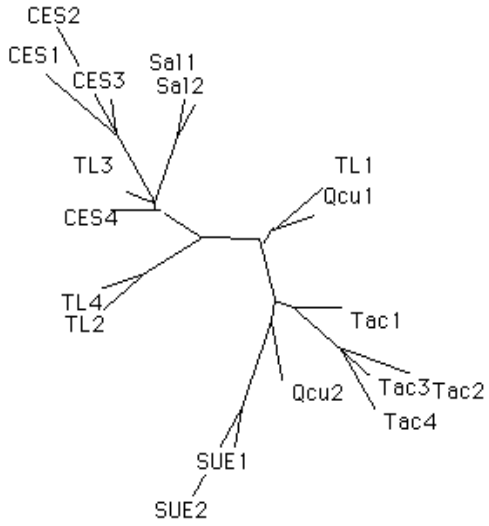


Figure 9 : Arbre V, répartition des classes de fréquences

Les résultats sont satisfaisants, puisque les extraits de tous les auteurs sont regroupés. On constate même, parmi les extraits d'un même auteur, des rapprochements intéressants. Ainsi les *Annales* de Tacite se distinguent des *Histoires*, les deux derniers quarts de l'œuvre de Tite-Live des deux premiers, la première moitié de la *Guerre des Gaules* du reste des *Commentaires.*

Un tri ne retenant que les formes associées à des khi-2 très élevés, et donc distribuées de la façon la plus irrégulière, donne des résultats tout à fait équivalents.

5. Distribution des fréquences

La répartition des classes de fréquence a été analysée en isolant, dans chaque texte, les 5 formes les plus fréquentes, puis en réunissant les autres formes d'une fréquence supérieure à 20, puis comprise entre 11 et 20, 6 et 10, 4 et 6, enfin en décomptant séparément les formes de fréquence 3, 2 et 1.

Les résultats, qui apparaissent sur l'arbre V (figure 9), sont un peu décevants. Seuls les fragments de Tacite, de Salluste et de Suétone sont regroupés, de même que trois des textes de César, et deux des textes de Tite-Live.

L'arbre utilisant des effectifs filtrés donne sensiblement les mêmes résultats.

6. Ensemble des données

Nous avons réuni toutes les données dont nous disposions sur un seul tableau, en ne retenant que les plus significatives. Ce tableau, de 18 lignes et de 60 colonnes, a permis de tracer l'arbre VI (figure 10). Comme pour l'arbre IV (figure 8) on observe un parfait regroupement des fragments empruntés à un même auteur. Ce test global paraît donc très satisfaisant.

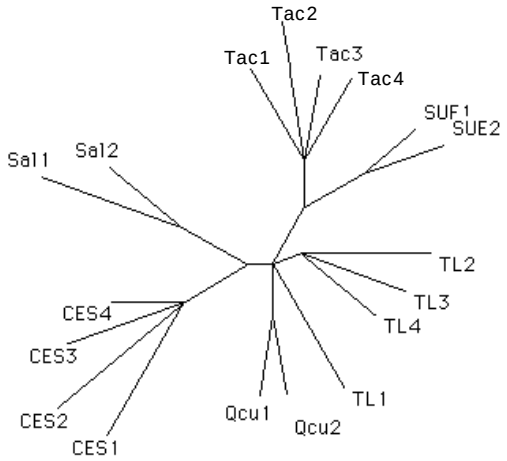


Figure 10: Arbre VI, ensemble des critères.

On constate donc que deux au moins des tests retenus, l'utilisation des formes les plus fréquentes, et la réunion des critères les plus performants, ont donné des résultats conformes à notre attente. Les autres tests ont certes permis de réunir les fragments appartenant à certaines œuvres, mais ont laissé subsister ailleurs des associations et des dissociations qui ne correspondaient pas à l'unité des ouvrages analysés.

On doit donc admettre l'efficacité de certains critères statistiques, non seulement lorsqu'il s'agit de résoudre des problèmes d'attribution, mais aussi, peut-être, pour analyser l'évolution d'un même auteur au fil de son œuvre.

Il reste cependant à essayer de déterminer les limites de la fiabilité que l'on peut accorder à ces méthodes.

Nos auteurs étaient relativement proches les uns des autres par leur inspiration et l'époque où ils ont écrit, mais offraient cependant un assez grand nombre de traits distinctifs. Il nous a paru intéressant de reprendre notre expérience en ajoutant à notre corpus des textes présentant de grandes ressemblances avec certaines des œuvres que nous avions analysées.

2. Les historiens latins et les continuateurs de César

C'est précisément le cas des *Commentaires sur la Guerre d'Alexandrie*, la *Guerre d'Afrique* et la *Guerre d'Espagne*, qui prennent la suite des *Commentaires* de César *sur la Guerre des Gaules* et la *Guerre civile*, même s'ils ne sont pas de la main de César.
Nous avons donc ajouté à nos tableaux trois nouveaux fragments de 4 000 mots, empruntés aux continuateurs de César.
Nous ne présenterons pas les différents arbres que nous avons pu construire en reprenant les cinq critères statistiques que nous avions déjà utilisés. Dans tous les cas, on constate que les trois extraits des continuateurs de César viennent se mêler aux quatre extraits de César.
Nous ne retiendrons que trois arbres, à titre d'exemple.

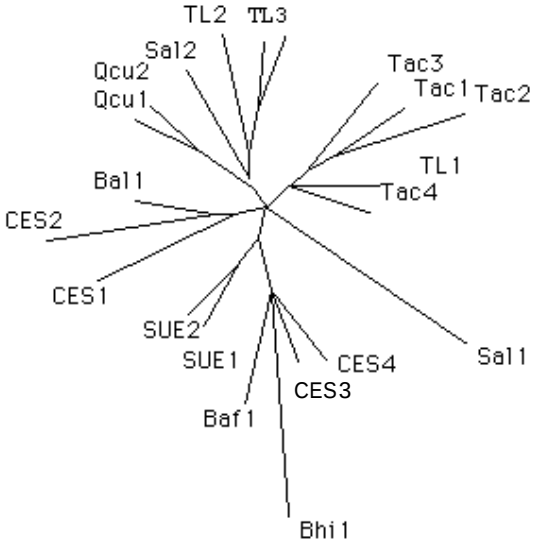


Figure 11 : Arbre VII, les graphèmes dans l'ensemble des formes les plus fréquentes

Ainsi l'arbre VII (figure 11) correspond à l'utilisation des graphèmes dans l'ensemble des formes. On constate que les textes empruntés à César constituent deux rameaux distincts, d'un côté les deux extraits de la *Guerre des Gaules*(CES1,CES2), rejoints par la *Guerre d'Alexandrie* (Bal1), de l'autre les deux extraits de la *Guerre civile* (CES3,CES4) associés à la *Guerre d'Afrique* (Baf1) et à la *Guerre d'Espagne*(Bhi1).

L'arbre VIII (figure 12) repose sur la répartition des formes les plus fréquentes dans l'ensemble des textes. Rappelons que ce critère avait donné un excellent résultat, puisque tous les extraits d'un même auteur s'étaient trouvés regroupés (Arbre IV, figure 8).

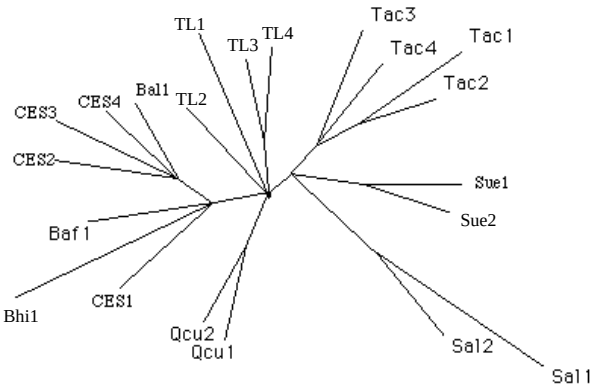


Figure 12 : Arbre VIII, répartition des formes les plus fréquentes.

Les continuateurs de César se rapprochent maintenant de leur modèle. Il est impossible de disjoindre la *Guerre d'Alexandrie* des extraits 2, 3 et 4 de César, tandis que l'extrait 1 est voisin de la *Guerre d'Afrique* et de la *Guerre d'Espagne*.

Il en va de même si l'on examine l'arbre IX (figure 13), que nous avons tracé en réunissant toutes les données dont nous disposions. Les *Guerres d'Alexandrie, d'Afrique et d'Espagne* viennent rejoindre le faisceau des *Commentaires* de César. Remarquons cependant que la longueur des arêtes est significative. Il peut être intéressant de noter que la *Guerre d'Espagne*, considérée, en général, comme la moins réussie des œuvres des continuateurs de César, est aussi la plus éloignée du sommet du groupement.

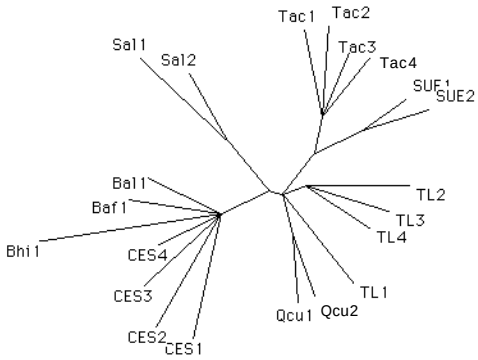


Figure 13 : Arbre IX, ensemble des critères.

Pour affiner notre analyse, nous avons effectué une nouvelle recherche en ne prenant en compte que les

textes attribués à César et à ses continuateurs. Nous avons ajouté aux textes déjà étudiés le livre VIII de la Guerre des Gaules, dont on sait qu'il a été écrit par Aulus Hirtius et isolé, à l'intérieur de chaque écrit, une ou plusieurs tranches de 5 000 mots, une tranche pour le livre VIII de la *Guerre des Gaules* (Bg81), et pour la *Guerre d'Espagne* (Bhi1), 2 tranches pour la *Guerre d'Afrique* (Baf1 et Baf2), ainsi que pour la *Guerre d'Alexandrie* (Bal1 et Bal2), 4 tranches pour la *Guerre des Gaules* (Bga1, Bga2, Bga3, Bga4) et 6 tranches pour la *Guerre civile* (Bci1, Bci2, Bci3, Bci4, Bci5, Bci6), soit, au total, 16 tranches différentes..
Nous avons soumis l'ensemble de ces textes aux critères d'analyse décrits plus haut.

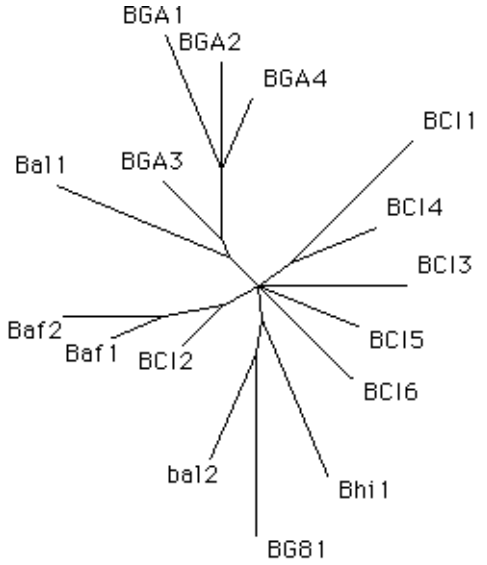


Figure 14 : Arbre X, les graphèmes à l'initiale du mot

L'utilisation des graphèmes à l'initiale du mot (arbre X, figure 14) donne des résultats assez intéressant, puisque sont regroupés les quatre extraits de la *Guerre des Gaules*, de même que cinq des six extraits de la *Guerre civile*. Le livre VIII de la *Guerre des Gaules* paraît isolé, de même que la *Guerre d'Espagne*. Seul échec, la disjonction des deux livres de la *Guerre d'Alexandrie*.

Nous ne commenterons pas dans le détail les autres arbres, dont l'interprétation n'est pas toujours très claire.
À moins d'imaginer, hypothèse contraire à toute la tradition philologique, que César ait pu mettre lui-même la main à ces ouvrages, il semble bien que l'on atteint là les limites de nos tests. Notons, toutefois, que les textes des continuateurs de César sont très proches de leur modèle, non seulement par les sujets traités et le vocabulaire utilisé, mais aussi, sans doute par le recours à des sources de documentation identiques. Il n'est cependant pas exclu que de nouveaux critères, fondés sur d'autres caractères statistiques, puissent permettre de distinguer une œuvre originale de ses imitations. Ce serait là un nouveau sujet de recherches, qui pourrait inclure le livre VIII de la Guerre des Gaules, dû à Aulus Hirtius.

Conclusion

Nous voudrions, pour conclure, insister sur deux points.
Tout d'abord la commodité et la pertinence des représentations arborées. Si les méthodes d'analyse spatiale, l'analyse factorielle des correspondances ou *Multidimensional scaling*, restent des outils extrêmement précieux entre les mains des linguistes statisticiens, l'analyse arborée offre une lisibilité, et une efficacité, qui méritent d'être reconnues et utilisées.
Ensuite la possibilité, en se fondant exclusivement sur des critères statistiques, de relier entre elles les œuvres d'un même auteur. Même si l'on est encore loin de cette caractérologie du style qu'attendait Charles Guiraud, du moins peut-on espérer progresser quelque peu dans cette voie, et apporter aux philologues de nouveaux arguments dans leurs recherches concernant des problèmes d'attribution et d'identification.

Bibliographie

BARTHÉLEMY J.P. & GUÉNOCHE A. *Tree and Proximity Representations* Wiley & Sons 1991
BARTHÉLEMY J.P. & LUONG X. « Représenter les données textuelles par les arbres… » JADT 1998, *4èmes Journées Internationales d'Analyse Statistique des Données Textuelles.* S. Mellet & al Ed.,pp49-71 Nice 1998
DUBROCARD M. « *César* dans César » in *Travaux du cercle linguistique de Nice*. Université de Nice 1994, **16**, 157-174
LUONG N.X. Ed. *Analyse arborée des données textuelles. Tree Analysis of Textual Data* CUMFID **16**, InaLF, CNRS Nice 1988