

Les affinités lexicales. Hommage à Etienne Evrard

Étienne Brunet

▶ **To cite this version:**

Étienne Brunet. Les affinités lexicales. Hommage à Etienne Evrard. Langues anciennes et analyse statistique. Cinquante ans après, Fialon S., Longrée D., Pietquin P., Oct 2011, Rome, Italie. pp.9-31. hal-01363232

HAL Id: hal-01363232
https://hal.univ-cotedazur.fr/hal-01363232

Submitted on 9 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Les affinités lexicales
Hommage à Étienne Évrard

Étienne Brunet

Ce n'est pas la première fois que l'occasion m'est donnée de rendre hommage à Étienne Évrard. Déjà en 1987 un colloque s'était tenu autour de lui, au Lasla, où s'étaient rassemblés ses disciples et ses amis. Et comme nous partagions le même prénom, cela m'avait permis, en levant mon verre, de dire - et non plus d'entendre- : « à la tienne Étienne! » Le présent Colloque au moment où il a été conçu était aussi un rassemblement où la latinitas s'apprêtait à fêter à Rome le cinquantième anniversaire du Lasla et l'œuvre de celui qui pendant cinquante ans avait été l'initiateur et le principal ouvrier de l'entreprise. C'est à dessein que j'emploie le mot modeste "ouvrier", plutôt que maître ou directeur. Car Évrard aimait travailler en intégrant toutes les contraintes de la recherche, même les plus techniques. Parmi les pionniers de la discipline (Roberto Busa, Pierre Guiraud, Bernard Quemada, Charles Muller), il est le seul à avoir osé mettre les mains dans le cambouis et s'attaquer à la programmation. Étienne hélas a manqué de peu son dernier rendez-vous. Son image s'impose sans doute devant vos yeux comme devant les miens et je ne puis m'empêcher de la fixer sur le mur, avec ses longs cheveux qu'il ne coupait plus et que j'ai vus, lors de notre dernière rencontre, flotter au vent du Cap de Bonne Espérance.

Figure 1. Étienne Évrard



Malgré mon âge avancé, je ne puis remonter jusqu'à la naissance du Lasla. Mais je puis faire la moitié du chemin en arrière et reprendre un sujet abordé à Liège en 1987 et soumis à l'appréciation d'Évrard et de ses disciples[1]. Il s'agit de la "connexion lexicale", notion que Muller avait mise en vogue et que je prétendais appliquer à Hugo, en lui donnant une appellation plus ambitieuse de "distance intertextuelle"[2]. Je ne suis plus très fier d'avoir proposé cette dénomination depuis qu'elle est au cœur de l'affaire Corneille-Molière. En réalité parler de connexion lexicale comme Muller[3], c'était envisager sous une forme positive ce que recouvre le mot distance, sous une forme négative et plus distanciée. Évrard qui était le premier à aborder concrètement ce problème[4], avait proposé le terme affinités, qui avait déjà servi dans les sciences avec le *Traité des affinités* de Bergman (1788) et les *Recherches sur les lois des affinités chimiques* de Berthelot (1801), mais aussi dans les lettres avec les *Affinités électives* de Goethe (1809).

I - Les affinités lexicales. Première application en 1966.

Les changements de terminologie perturbent souvent la chronologie et font oublier les origines. Arrêtons-nous sur le texte fondateur, proposé par Evrard, il y a cinquante ans, au Colloque de Strasbourg[5]. La plupart des pionniers qui ont illustré la statistique linguistique ou la linguistique tout court se trouvaient là autour des organisateurs Charles Muller et Bernard Pottier[6]. Le Colloque eut lieu

---

[1] Les actes de ce colloque ont paru en 1988, sous le titre: "Le nombre et le texte. Hommage à Étienne Évrard", in *Revue, Informatique et Statistique dans les Sciences humaines*, n° 1-4, C.I.P.L. Liège.

[2] Étienne Brunet, "Une mesure de la distance intertextuelle: la connexion lexicale", in *Revue*, ibid., p.81-116.

[3] G. Herdan se range du côté de Muller et estime « Vocabulary connectivity more appropriate than the conventionnal correlation methods », *Statistique et analyse statistique*, PUF, 1966, p.134.

[4] Cependant l'idée d'une prise en compte globale du contenu lexical est évoquée par Guiraud , dès 1959, dans *Problèmes et méthodes de la statistique linguistique* (p.129) : « On pourrait établir un tableau de corrélations lexicales entre les différentes œuvres en les prenant deux par deux pour voir les mots qu'elles ont en commun et ceux qu'elles ont en propre ; mais c'est un travail énorme ». Guiraud, dépourvu de moyens de calcul et acculé à des calculs manuels, n'a pu vérifier son hypothèse.

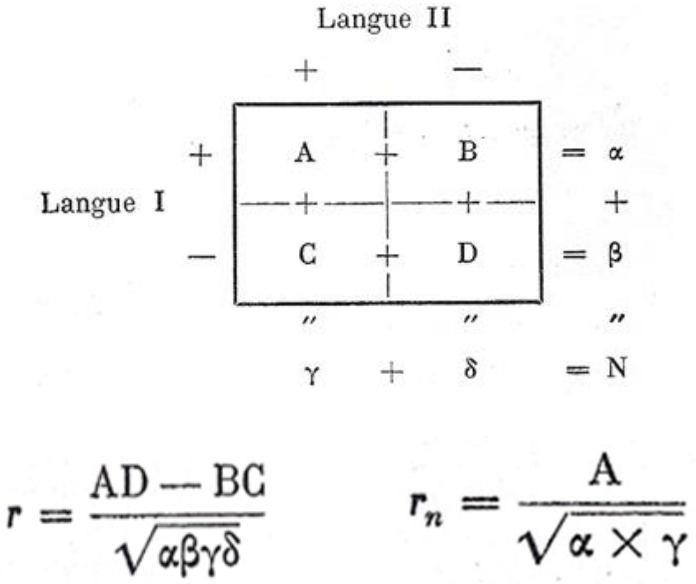
[5] É. Évrard, « Étude statistique sur les affinités de cinquante-huit dialectes bantous », Statistique et Analyse Linguistique, PUF, 1966, p.85-94.

[6] Parmi 70 participants on relève les noms de R. Busa, E. Coseriu, J.Dubois, É. Évrard, A. Greimas, P. Guiraud, G. Herdan, R. Martin, H. Mitterand, G. Gougenheim, R. Moreau, B. Quemada, K. Togeby, P. Wexler, A. Zampolli. Le LASLA, venu en voisin, était largement représenté, avec les communications de S. Govaerts et A. Bodson épaulant celle d'Évrard.

en avril 1964 et les *Actes* suivirent en 1966. À cette époque la question de la statistique linguistique était d'une actualité brûlante. P.Guiraud qui avait publié quatre ouvrages dans ce domaine nouveau, avait lancé le mouvement. Et l'année 1964 voit éclore en même temps trois publications fondamentales *Quantitative Linguistics* de G.Herdan, *le Français fondamental* de Gougenheim, Michéa, Rivenc et Sauvageot, et *l'Essai de statistique lexicale* de Ch. Muller.

1 - La communication d'Étienne Évrard avait de quoi surprendre, ce qu'il avoue avec humour. Ne connaissant rien aux dialectes qu'on parle en Afrique, il pouvait se prévaloir d'une impartialité parfaite en proposant une typologie objective, fondée sur des données numériques et cryptées. On lui avait fourni 58 listes codées, chacune correspondant à un dialecte et détaillant les mots, eux aussi codés, qui expriment dans ce dialecte une centaine de racines communes. Le dispositif qu'il met en place consiste, pour deux langues que l'on confronte, à compter le nombre de racines représentées dans ces deux langues (A), dans la première mais non dans la seconde (B) , dans la seconde mais non dans la première (C), et enfin dans aucune des deux (D). Et un tableau est ainsi constitué dont on calcule les effectifs marginaux pour en tirer le coefficient de corrélation de Bernoulli.

Figure 2. Le coefficient de corrélation de Bernoulli

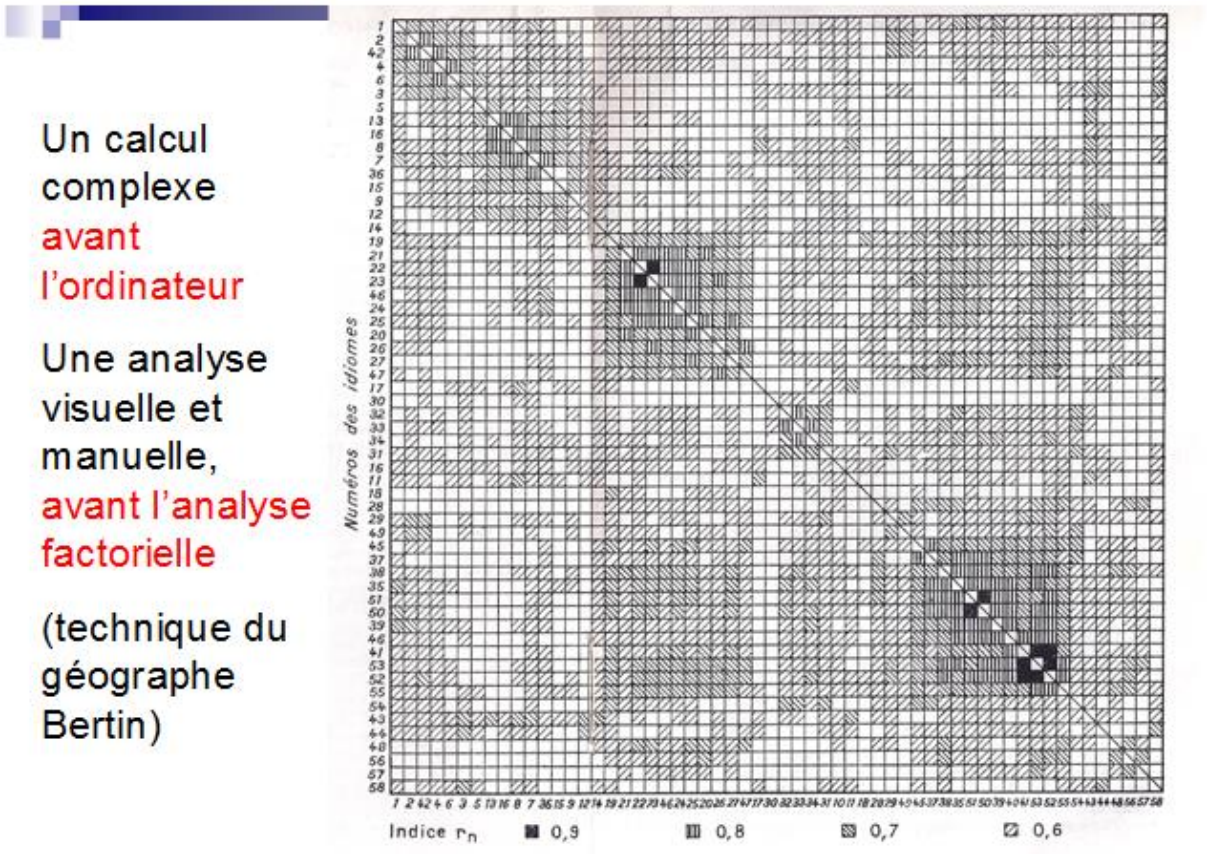


$$r = \frac{AD - BC}{\sqrt{\alpha\beta\gamma\delta}} \qquad r_n = \frac{A}{\sqrt{\alpha \times \gamma}}$$

On voit bien que le coefficient *r* augmente, c'est-à-dire la proximité des deux textes, quand les éléments communs présents (A) ou absents (D) l'emportent sur les éléments privatifs (B et C), le dénominateur servant de pondération pour contraindre le résultat entre les limites -1 et +1. Si l'on néglige les mots absents dans les deux textes (case D), la formule (rn) est simplifiée et les éléments qui dépendent de D disparaissent (BC, bêta et delta).

2 - Le résultat global est donné dans un tableau carré qui croise deux à deux tous les textes. La valeur plus ou moins forte du coefficient (entre 0.6 et 0.9) s'inscrit dans des cases plus ou moins ombrées, les plus claires étant réservées aux paires qui n'atteignent pas la valeur 0.6. Ne faisons pas grief à l'auteur de l'obscurité relative d'un tel tableau. On ne connaissait pas à l'époque les techniques multidimensionnelles. On savait seulement manipuler manuellement les lignes et les colonnes pour rapprocher sur la diagonale les affinités les plus fortes. Le procédé, purement graphique et visuel, avait reçu quelques perfectionnements du géographe Bertin, dont il n'est pas certain qu'Étienne Évrard ait profité. Sans doute le fournisseur des données avait-il suivi la proximité géographique dans l'attribution des codes. Car, malgré quelques ajustements, les constellations que l'œil repère sur la diagonale respectent d'assez près la série séquentielle des codes, et donc le voisinage sur la carte africaine (par exemple les séquences 1 à 6, 19 à 27, 30 à 34, 35 à 41 et 50 à 55). Mais en déplaçant certaines lignes dans un entourage inattendu (par exemple les lignes 42, 36, 47), Évrard invite à rattacher les dialectes correspondants à une famille que la localisation n'imposait pas.

Figure 3. Représentation graphique des affinités lexicales relevées par Évrard (*Statistique et Analyse linguistique, PUF, 1966, p.89*)



3 - Comme en cinquante ans la technologie a offert de nouveaux moyens d'exploration des données, nous avons voulu vérifier l'analyse évrardienne grâce à l'analyse arborée. À partir des cases ombrées de la figure 3, nous avons

reconstruit le tableau numérique original (reproduit partiellement dans la figure 4, avec une échelle de valeurs de 0 à 4 pour aller du blanc au noir). L'analyse arborée (figure 5) restitue alors une typologie très claire, où les affinités remarquables distinguées par Évrard sont respectées (à l'exception du dialecte 36 qui a changé de partenaires) et jusqu'à la parenté qu'Évrard souligne entre le groupe E et le groupe C (tous les deux se donnant la main au haut du graphique). Les groupements sont cependant plus étoffés et peu de dialectes sont laissés sans famille. L'analyse est ainsi plus complète et plus fiable et donne plus de force au coefficient de Bernoulli qu'Évrard a proposé.

Figure 4. Reconstitution du tableau des valeurs numériques

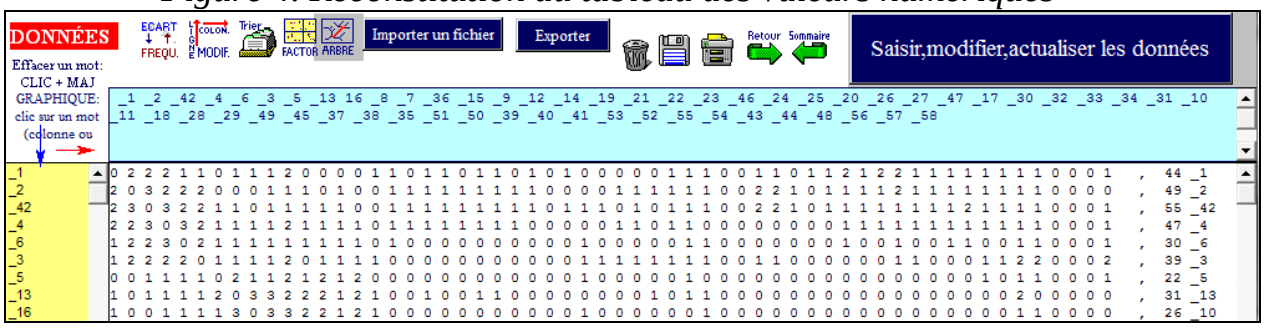
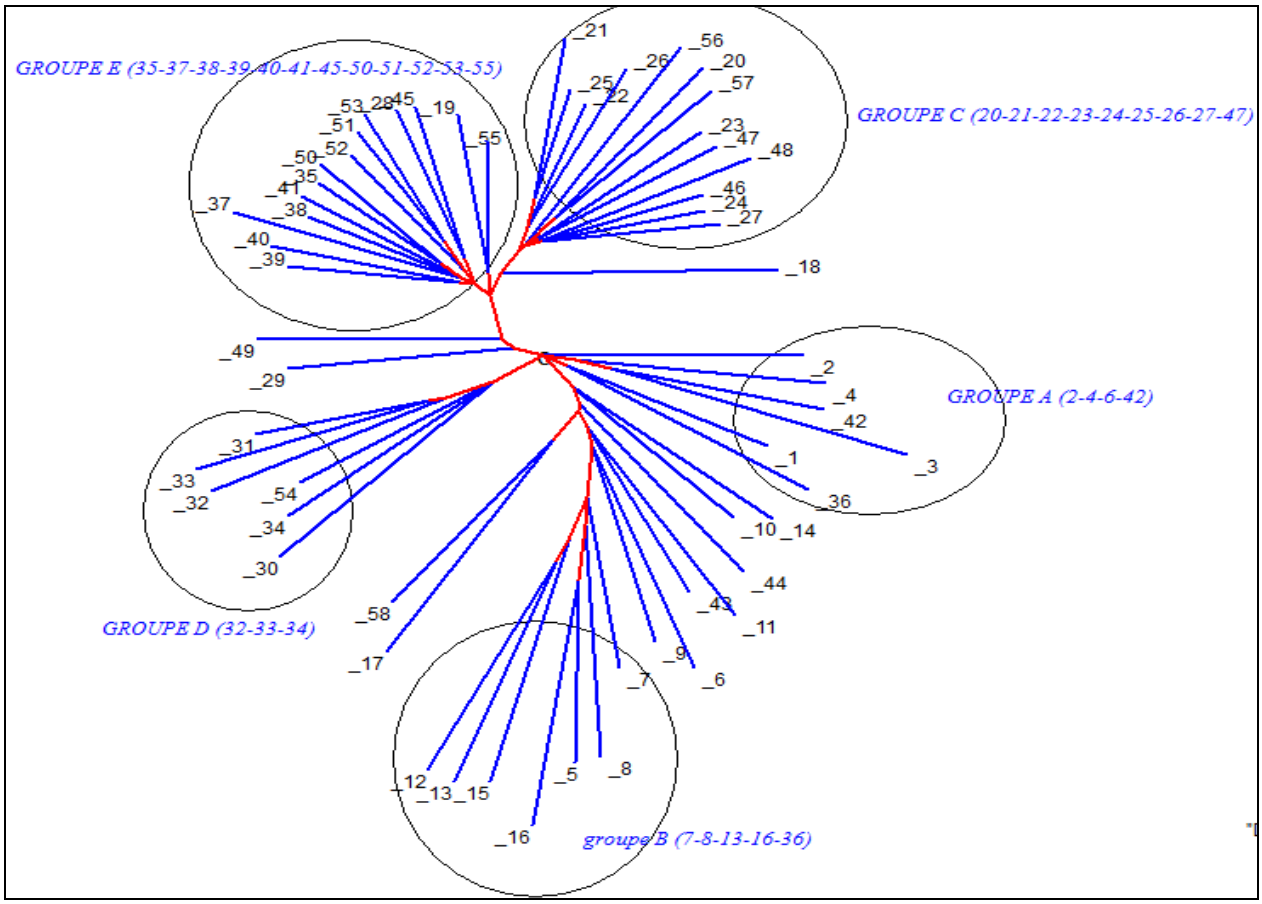


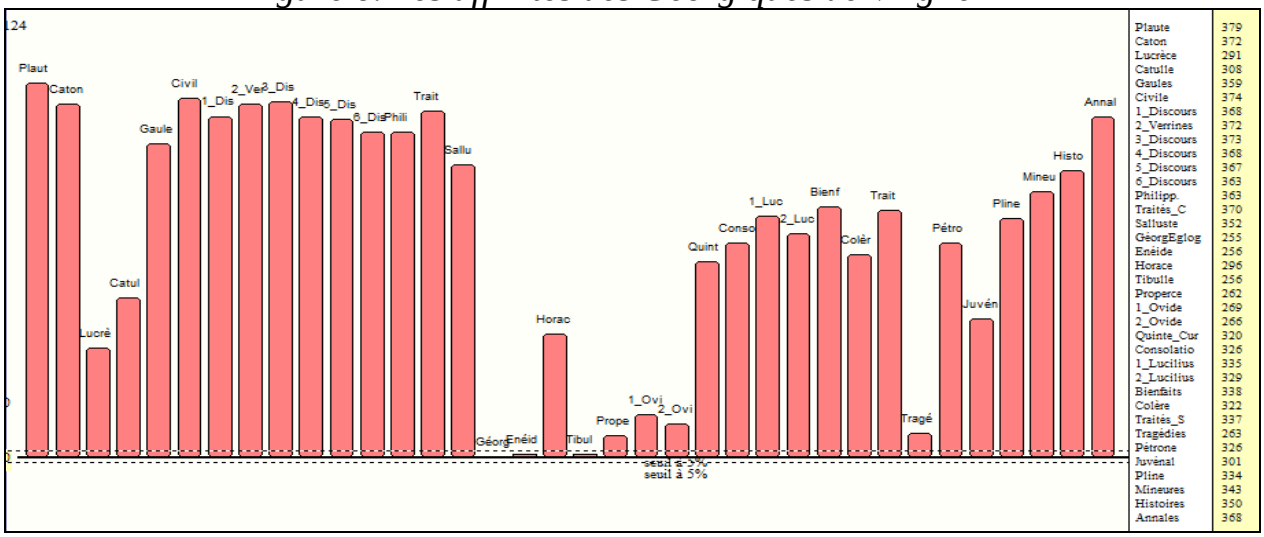
Figure 5. Analyse arborée des affinités de 58 dialectes bantous.
(entre parenthèses les dialectes qu'Évrard avait constitués en famille)



II - Les affinités lexicales. Application au latin

Mais les données sur les langues bantoues sont trop éloignées dans l'espace et dans l'expérience commune pour être immédiatement contrôlables. Appliquons la formule d'Évrard à des données plus familières, quoique éloignées dans le temps. Chacun les connaît ici, puisque ce sont celles du Lasla. Ce précieux dépôt a été transmis de main en main depuis Évrard jusqu'à Sylvie Mellet, en passant par Joseph Denooz, Dominique Longrée et Gérard Purnelle, ce dernier les ayant spécialement préparées pour le logiciel *Hyperbase*[7]. Il ne s'agit plus de racines comme dans la recherche précédente, mais de lemmes dûment désambiguïsés et regroupés selon les procédés mis au point au *Lasla*. Avant d'envisager le panorama d'ensemble, choisissons un texte au milieu de la chronologie, par exemple les *Géorgiques*. La dernière colonne à droite de la figure 6 rend compte du calcul d'Évrard, légèrement transformé : on a choisi le complément à l'unité et multiplié la résultat par 1000, soit (1-r)*1000. Les deux valeurs extrêmes sont aux deux bouts de la chaîne, soit avant (Plaute 379, Caton 372), soit après (*Annales* de Tacite 368). Et les plus proches sont au centre. Mais si la chronologie justifie le mouvement parabolique de la courbe, les violents à-coups qu'on y voit relèvent du genre. Les *Géorgiques* ont leurs affinités les plus fortes avec les œuvres en vers, d'abord avec d'autres œuvres du même auteur (l'*Énéide*), puis celles de poètes contemporains[8] (Tibulle, Properce, Ovide) et enfin avec des textes en vers plus éloignés dans le temps (Lucrèce et Catulle d'un côté, Juvénal de l'autre).

Figure 6. Les affinités des Géorgiques de Virgile

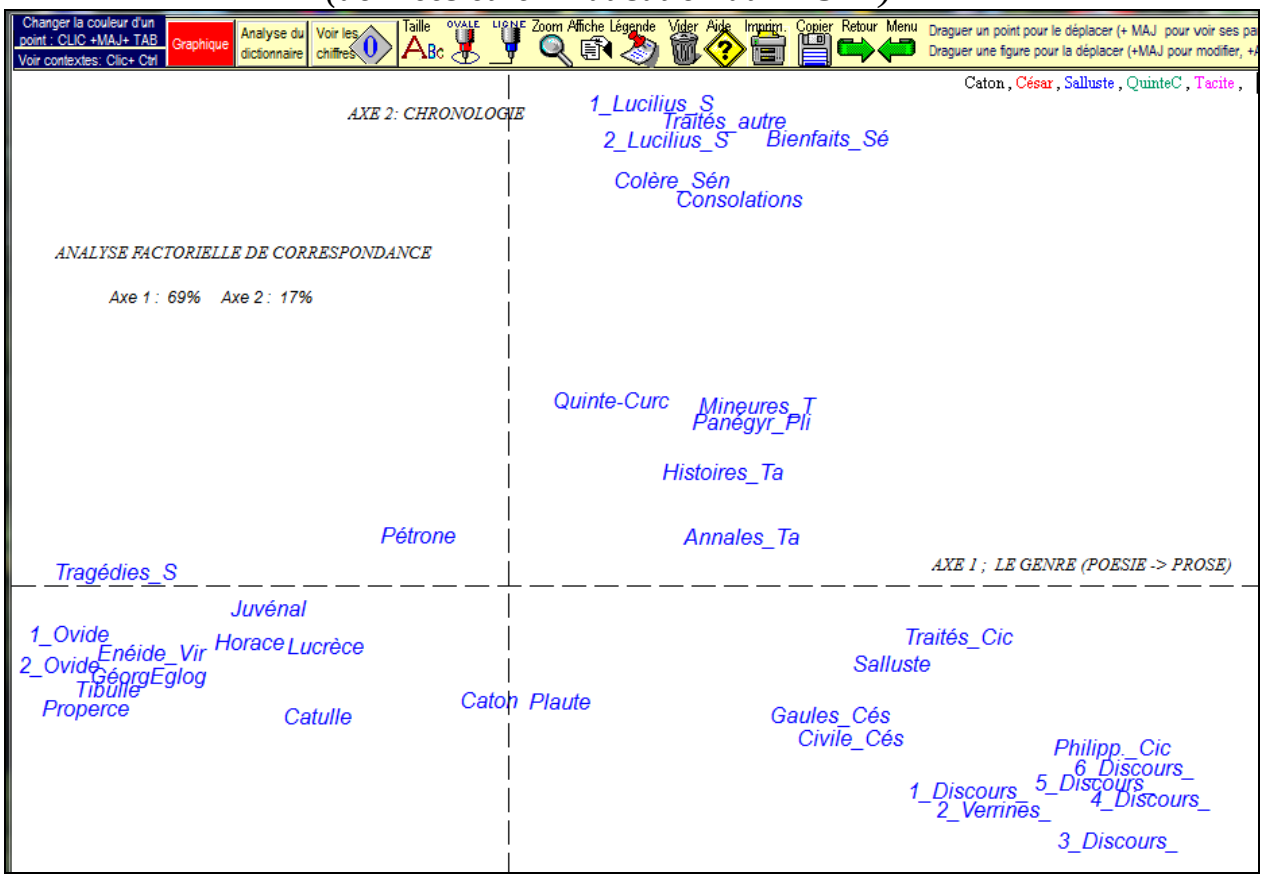


[7] MELLET, Sylvie et Etienne BRUNET 2004. *HYPERBASE, Manuel de référence pour la base de littérature latine* (Cédérom distribué par l'UMR 7320, s'adresser à Sylvie Mellet).

[8] Les textes en vers se répartissent en sous-genres, qui produisent à leur tour des écarts. Ainsi les *Odes* d'Horace s'accordent moins bien aux *Géorgiques* que les *Tragédies* de Sénèque.

Le clan serré des poètes se retrouve dans l'analyse factorielle de la figure 7 et y occupe fermement un territoire à gauche. Cette fois c'est tout le tableau des affinités qui est considéré, les *Géorgiques* n'ayant que leur juste part, égale à celle des autres textes. La dichotomie du genre est ici absolue, la poésie à gauche, la prose à droite. Et ce premier facteur (69% de l'inertie) domine nettement le second (17%) dans lequel on reconnaît la chronologie et qui oppose, de bas en haut, les textes les plus anciens aux plus récents.

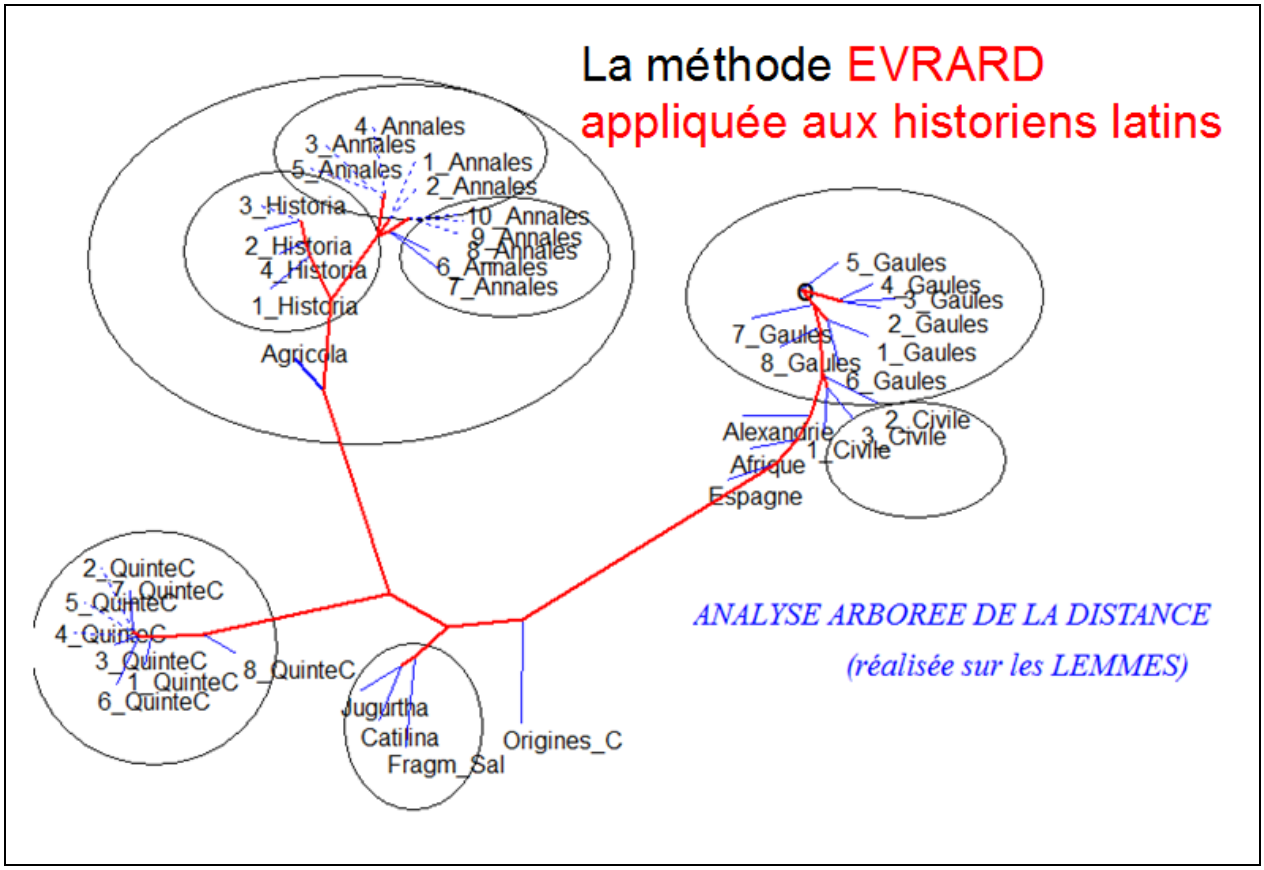
Figure 7. Analyse factorielle des textes latins (données et lemmatisation du LASLA)



On dira peut-être qu'un lecteur n'a pas besoin de calcul pour différencier le vers et la prose. Mais la mise en page qui saute aux yeux du lecteur ou la diction qui suffirait à son oreille sont inaccessibles à une machine sourde et aveugle, qui ne s'appuie que sur le comptage des mots. De même l'œil humain reconnaît facilement les visages de Pierre et de Paul et sans doute leur éventuelle parenté. Mais l'ADN déterminera plus sûrement si l'un est le fils de l'autre. Peut-on demander des analyses plus fines au coefficient d'Évrard ? Cette fois on écartera le genre, afin d'isoler la chronologie et la personnalité des écrivains. Ceux-ci ne se mélangent pas, chacun rassemblant ses œuvres dans un coin de la figure, César à droite, Salluste en bas, Quinte-Curce à droite et Tacite en haut. Un chemin est tracé qui parcourt la chaîne des auteurs selon la séquence

chronologique, de César à Tacite[9]. Plus finement encore l'analyse distingue chez César la *guerre des Gaules* et la *guerre civile*, et met à distance des œuvres dont la signature est moins sûre ou moins reconnaissable (successivement les textes relatifs à Alexandrie, à l'Afrique et à l'Espagne). Si les blocs dévolus à Salluste et à Quinte-Curce sont compacts, le territoire de Tacite est précisément compartimenté, la monographie sur Agricola se détachant des *Histoires*, celles-ci des *Annales* et dans les *Annales* les six premiers livres des derniers. On retrouve ici les conclusions que Sylvie Mellet, disciple reconnue et reconnaissante d'Évrard, a obtenues à partir d'autres indices et notamment à travers la distance grammaticale[10].

Figure 8. Analyse arborée des historiens latins selon le coefficient r d'Évrard



[9] Regrettons ici l'absence de Tite-Live, qui n'a pas été oublié mais qui a hérité d'un rang malheureux dans l'ordonnancement des travaux du *Lasla*.

[10] MELLET, Sylvie et Xuan LUONG 2003. « Mesures de distance grammaticale entre les textes ». *CORPUS 2*, 141-166.

III - D'autres indices de proximité. Le coefficient de Jaccard

On voit qu'en cinquante ans la technologie offre une meilleure lisibilité et une plus grande extension aux affinités définies par Evrard. Mais cette définition est-elle définitive ? Y a-t-il un progrès quand à partir des affinités on passe à l'indice de Jaccard, à la connexion de Muller, à la distance intertextuelle ou à quelque autre mesure visant à établir la parenté ou la similitude de deux textes? Du premier d'entre eux, l'indice de Jaccard, on ne saurait dire qu'il représente une avancée car les ingrédients sont les mêmes. Le coefficient de Jaccard est présenté, dans son expression la plus simple, comme le rapport entre la taille de l'intersection des deux ensembles A et B et la taille de l'union de ces deux ensembles[11] :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Le numérateur est le même que dans la formule simplifiée r_n d'Évrard, le dénominateur ayant aussi les mêmes éléments, mais assemblés de façon différente. On sait que Paul Jaccard a proposé son indice au début du 20e siècle[12]. Le nom de Bernoulli donné à l'indice d'Évrard paraît garantir l'avantage de l'antériorité. Mais le patronage de Bernoulli n'est peut-être qu'honorifique ; Évrard ne cite guère qu'un devancier, Czekanowski, dont les travaux sont publiés au milieu du même siècle. Il semble donc que, comme beaucoup d'autres indices, la formule d'Évrard soit dérivée de Jaccard.

C'est aussi le cas de celle qui est proposée dans Hyperbase sous le nom de Jaccard et qui met en jeu l'intersection ab de deux vocabulaires a et b, sous la forme d'un indice de distance et non de similarité :

distance = (((a-ab)/q) + ((b-ab)/b)) / 2

On ajoute ainsi la proportion du vocabulaire exclusif dans A à la proportion du vocabulaire exclusif dans B avant de diviser la somme par 2 pour obtenir une distance comprise entre 0 et 1. Nous avons eu la surprise de retrouver récemment dans une revue[13] datant de 1989 cette formule que nous avions

---

[11] On en déduit la distance Jaccard

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

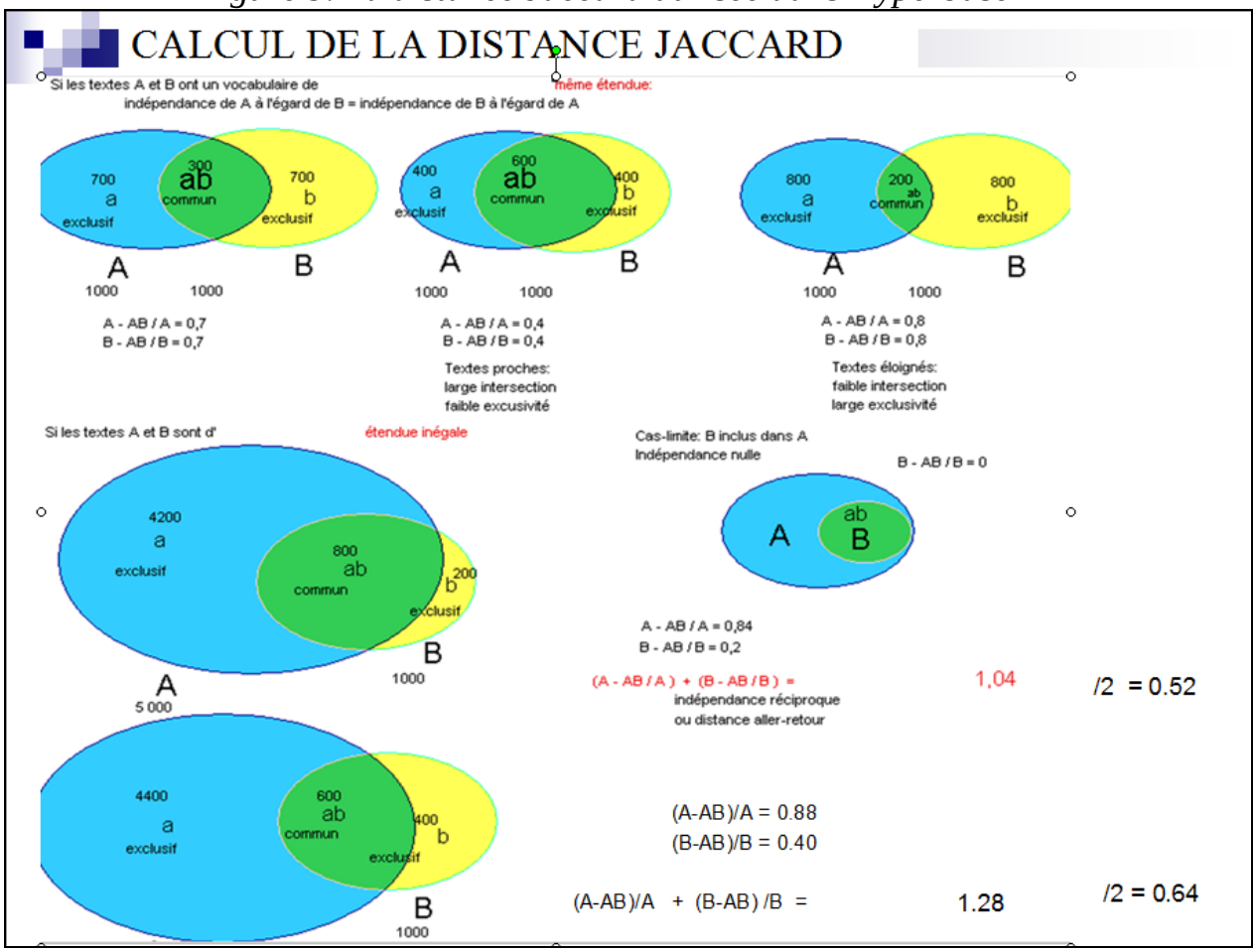
[12] Paul Jaccard (1901) *Bulletin de la société Vaudoise des Sciences Naturelles 37*, p .241-272.

[13] JF.B. Baulieu, « A classification of Presence/Absence Based Dissimilarity Coefficients », *Journal of Classification 6*, p.233-246 (1989).

aménagée à notre façon pour la rendre indépendante de l'étendue. Elle figure parmi une vingtaine d'autres, avec les mêmes ingrédients cuisinés autrement.

La figure 9 expose son fonctionnement. Dans le cas où les deux textes sont de longueur égale, il n'y a pas de difficulté. Mais s'il y a un déséquilibre, le plus petit perd son autonomie et son vocabulaire risque de se fondre dans le plus grand. Dans le même temps le plus gros arrondit son domaine privatif qui se rapproche de 1 quand l'autre tend vers 0. En faisant la somme des deux valeurs, puis leur moyenne, on annule les effets du déséquilibre.

Figure 9. La distance Jaccard utilisée dans Hyperbase



On a quelque scrupule à proposer une application de notre coefficient aux données du *Lasla*, puisque les résultats concordent avec ceux qu'on a obtenus avec le coefficient d'Évrard. La différence, assez mince, tient au fait qu'on néglige ici, comme dans la version simplifiée r_n d'Évrard, les mots qu'on ne trouve ni dans le premier texte, ni dans le second. Cet accord dans les refus ou les absences peut renforcer ou compléter la communauté des goûts et préférences dont rend compte l'intersection des deux vocabulaires (paramètre A dans la formule d'Évrard). Mais cette procédure exige qu'on soit dans un espace clos et qu'on puisse délimiter les absences conjuguées sans se perdre dans l'infini.

IV- L'analyse factorielle du tableau des fréquences (ou TLE)

On construit d'abord le dictionnaire des fréquences du corpus: un tableau ayant des milliers de lignes (autant que de mots) et autant de colonnes que le corpus compte de textes. C'est ce que Salem appelle le Tableau Lexical Entier ou TLE (tableau 10). Tous les logiciels de traitement statistique construisent à un moment ou à un autre ce tableau général, nécessaire pour l'application de la loi de Zipf ou le calcul des spécificités Les colonnes désignent les textes ou subdivisions du corpus et les lignes les différents mots (graphies ou lemmes) en ordre alphabétique. À l'intersection on trouve une fréquence et non plus une indication sommaire de présence/absence. C'est là un avantage incontestable. Car si la distribution d'un mot dans deux textes est très déséquilibrée (mettons 19 contre 1), cette inégalité devrait tendre à augmenter la distance entre les deux textes, au même titre que la distribution voisine 20 contre 0. Or dans le calcul de Jaccard qui ne considère que la présence la répartition 19 contre 1 contribue à rapprocher les deux textes quand la seconde les éloigne. Nul besoin de configurer un indice et de calculer des effectifs et des rapports. Le calcul s'exerce directement sur le tableau des fréquences brutes, même s'il est permis de les pondérer par quelque moyen : en les transformant en fréquences relatives ou en écarts réduits.

Tableau 10. Extrait du TLE des historiens latins

	Orig	Ga1	Ga2	Ga3	Ga4	Ga5	Ga6	Ga7	Ga8	1Ci	2Ci	3Ci	Alex
Afri	Espa	Cati	Jugu	Frag	1Qu	2Qu	3Qu	4Qu	5Qu	6Qu	7Qu	8Qu	Agri	1Hi	2Hi	3Hi
4Hi	1An	2An	3An	4An	5An	6An	7An	8An	9An	10An						
ab			39	100	109	60	77	96	66	111	48	125	73	102	63
130	98	32	57	56	46	38	34	39	30	35	23	55	11	24	23	11
15	30	23	27	0	33	18	17	31	24	8						
abdo			11	25	21	7	35	38	18	29	5	2	17	0	2	
1	5	2	5	6	4	1	4	3	3	11	4	5	4	29	19	58
15	12	22	13	12	15	6	15	4	15	27						
abdvco			9	14	5	5	5	3	4	1	3	42	30	43	12	
28	17	12	4	5	3	0	14	25	13	11	2	31	3	21	10	11
16	11	1	2	1	3	5	3	3	2	0						
abeo1			14	5	9	9	8	5	17	9	6	3	15	0	3	
1	7	10	57	9	14	39	13	12	26	44	5	44	6	16	35	9
28	24	46	4	11	37	18	14	14	37	8						
abhorreo			10	5	7	7	6	5	6	3	5	16	5	2	4	
14	5	3	0	7	36	14	34	47	17	41	4	5	5	3	15	16
22	16	4	4	4	5	6	5	35	19	3						
abicio			8	2	5	5	38	29	4	11	3	1	3	8	1	
20	3	1	21	5	29	0	15	2	2	12	2	32	3	21	0	22
17	23	12	2	1	15	4	3	3	15	20						
abnvo			13	5	8	9	8	5	7	2	6	3	6	0	3	
1	6	3	19	32	5	1	13	21	4	3	5	6	25	24	44	46
48	25	41	20	34	39	30	23	23	22	22	etc...					

Quand on a lancé un programme d'analyse factorielle selon des paramètres du tableau 11, on obtient un graphique (figure 12) où les textes prennent place selon le vote que les mots ont déposé dans l'urne. Il est à noter que l'analyse prend en compte 3015 lignes, chacune ayant 41 fréquences, soit plus de 100000 nombres à traiter. Cela semble considérable mais le tableau réel comporte 11272

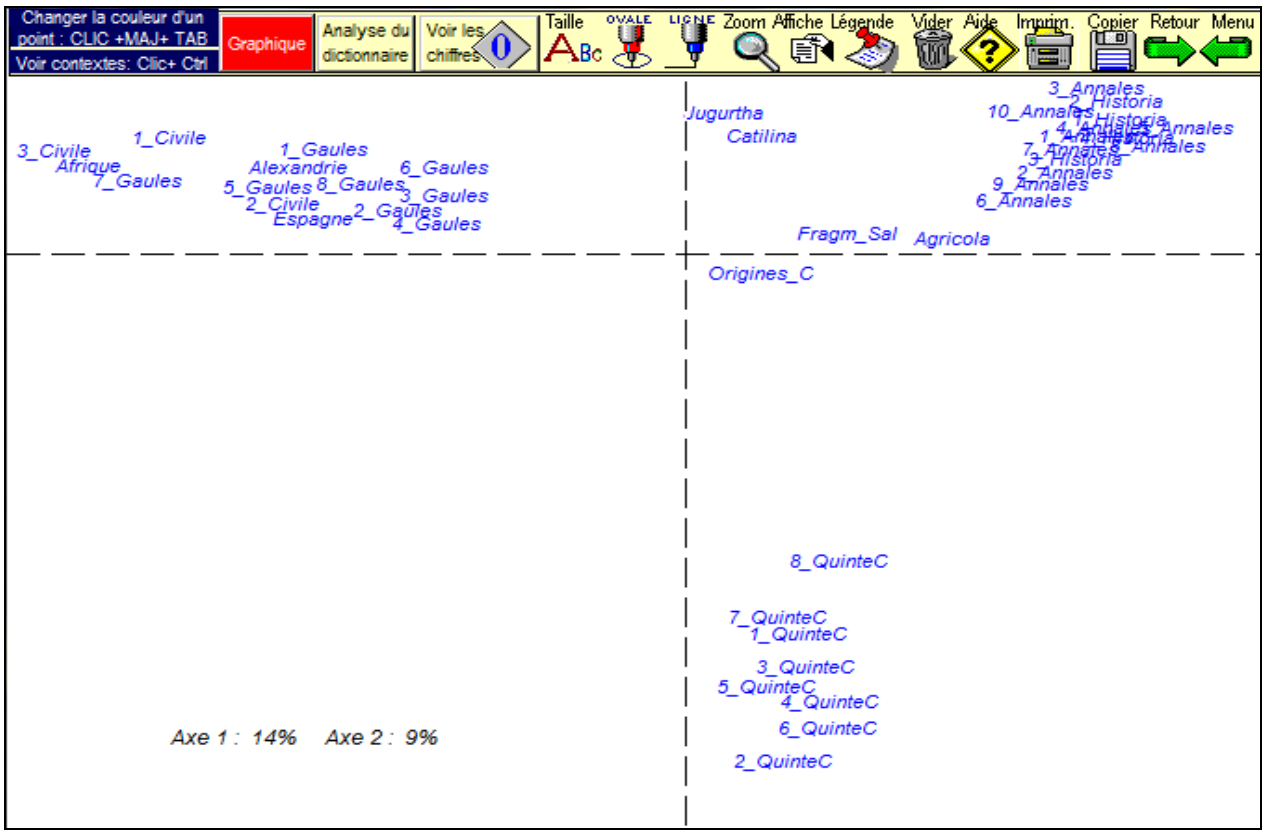
lignes, c'est-à-dire autant de lemmes, dont les trois quarts ont été négligés. Il s'agit des éléments moins fréquents, dont on peut faire l'économie parce qu'ils ont très peu d'influence sur le calcul. Au reste les 3000 mots intégrés dans le calcul ne peuvent trouver place, vu leur nombre, dans le graphique qui ne montre que les colonnes (les textes).

Tableau 11 Les paramètres du traitement

```
$RUN ANCORR
$L080
$F11=TABLEAU.afc
$PRT=ANALYSE.afc
$PAR=.
TITRE  ANALYSE FACTORIELLE ( écart réduit ) ;
PARAM NI = 3015 NJ = 41 NF = 4   ;
OPTIONS IMPFI=0 IMPFJ=1 NGR=2 ;
GRAPHE X=1 Y=2 GI = 0 GJ=3;
GRAPHE X=3 Y=4 GI = 0 GJ=3;
FLISTE Orig 1_Ga 2_Ga 3_Ga 4_Ga 5_Ga 6_Ga 7_Ga 8_Ga 1_Ci
2_Ci 3_Ci Alex Afri Espa Cati Jugu Frag 1_Qu 2_Qu
3_Qu 4_Qu 5_Qu 6_Qu 7_Qu 8_Qu Agri 1_Hi 2_Hi 3_Hi
4_Hi 1_An 2_An 3_An 4_An 5_An 6_An 7_An 8_An 9_An
10_A  ;
(12X,A4,120F5.0) ;
$END
```

Généralement les données sérielles ou chronologiques génèrent une courbe en forme de croissant où la chronologie va d'une pointe à l'autre. C'est bien ce qu'on observe ici : entre César à droite et Tacite à gauche se rangent les écrivains intermédiaires, Salluste dans la moitié haute, Quinte-Curce en bas.

Figure 12. Analyse factorielle du TLE des historiens



C'est, à grands traits, la disposition que proposait l'indice d'Évrard dans la figure 8. Mais le détail est ici moins fouillé et moins sûr. Si *Agricola* se détache des autres œuvres de Tacite, *Histoires* et *Annales* ne se distinguent plus. Et pareillement on observe un flottement chez César : on a perdu la vision claire qui dans l'analyse d'Évrard séparait *guerre des Gaules* et *guerre civile* et isolait les autres textes attribués ou non à César. En ne considérant que les fréquences hautes, quelques milliers au maximum, et en ignorant la majorité des mots pleins, on donne une plus-value aux mots grammaticaux, et donc aux phénomènes stylistiques, au détriment des particularités thématiques. Dans ce régime censitaire où ne votent que les riches, la loi du plus fort l'emporte non seulement dans la population des mots mais aussi dans la contribution inégale des textes.

V - La distance intertextuelle de Labbé

La formule de Labbé tient compte aussi de la fréquence, mais.de façon plus subtile. Pour chaque mot dans chaque couple de textes, on calcule la fréquence théorique ($theo_{jk}$) du mot dans le texte le plus court (j), compte tenu de sa fréquence dans le plus long (k). Cela donne un écart avec l'observation. L'écart est cumulé et, pondéré par l'écart maximum, il donne une mesure de distance. On envisage en principe tous les mots. Quelques retouches sont pourtant nécessaires lorsque la fréquence calculée est inférieure à 1, ou lorsque les écarts sont inférieurs à 0.5. Ainsi les hapax sont-ils écartés. Cela enlève un peu de généralité et de fiabilité à la démarche, sans en ruiner le crédit, pourvu qu'on écarte un barème absolu qui n'a pas été prévu pour le latin.

Appliqué à l'ensemble des données du *LASLA*, la distance intertextuelle de Labbé produit une carte satisfaisante (figure 13), où la poésie ne se compromet pas avec la prose, où les historiens font bande à part tout en gardant entre eux leurs distances, où les traités enfin se détachent des discours. Visiblement le genre impose sa loi, que respectent les écrivains quand, comme Sénèque, ils cultivent plusieurs genres[14]. La chronologie est ici écrasée par la domination du genre. Mais elle apparaît mieux lorsqu'on fait appel au coefficient d'Évrard (figure 14) : les historiens n'y sont plus groupés mais se répartissent de bas en haut selon l'axe du temps.

[14] Cependant les traités moraux de Cicéron n'accompagnent pas ceux de Sénèque, trop éloignés dans le temps. Ils se maintiennent sur la branche propre à Cicéron, mais à l'écart de ses discours.

Figure 13. Analyse arborée de la littérature latine selon la distance de Labbé

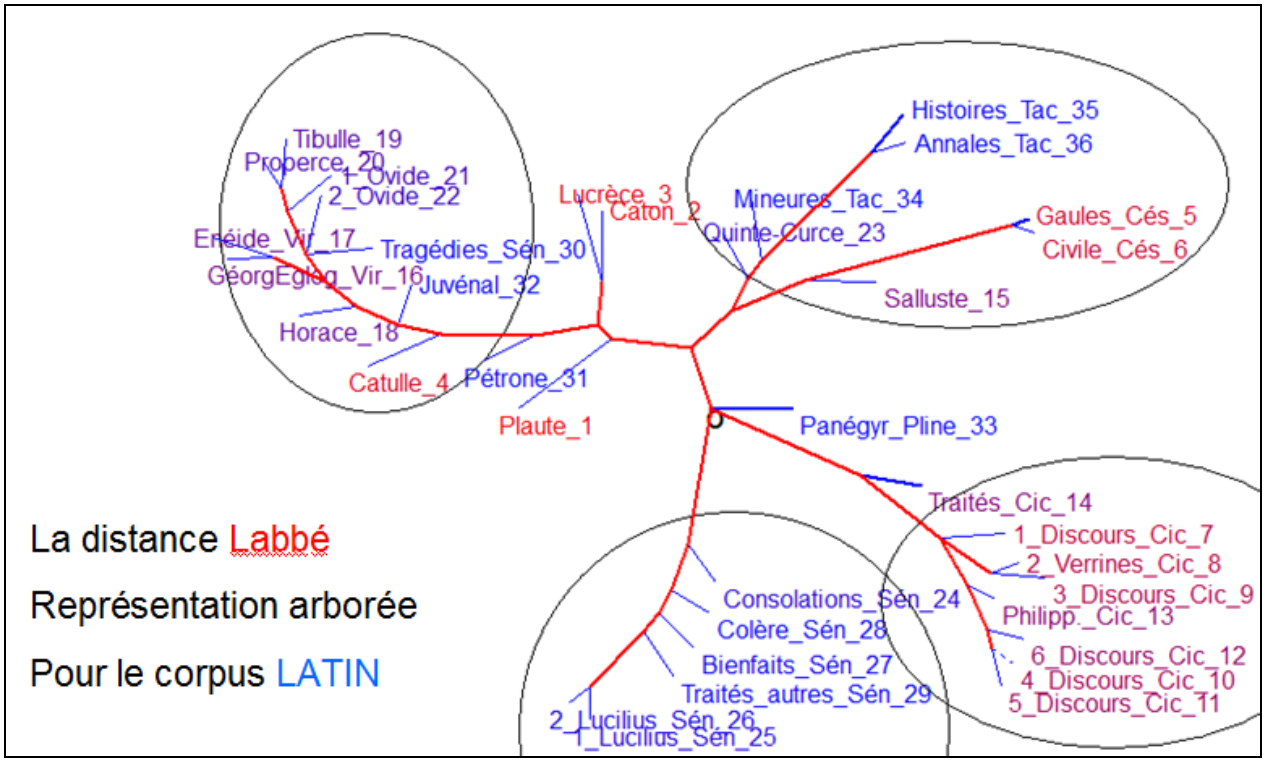
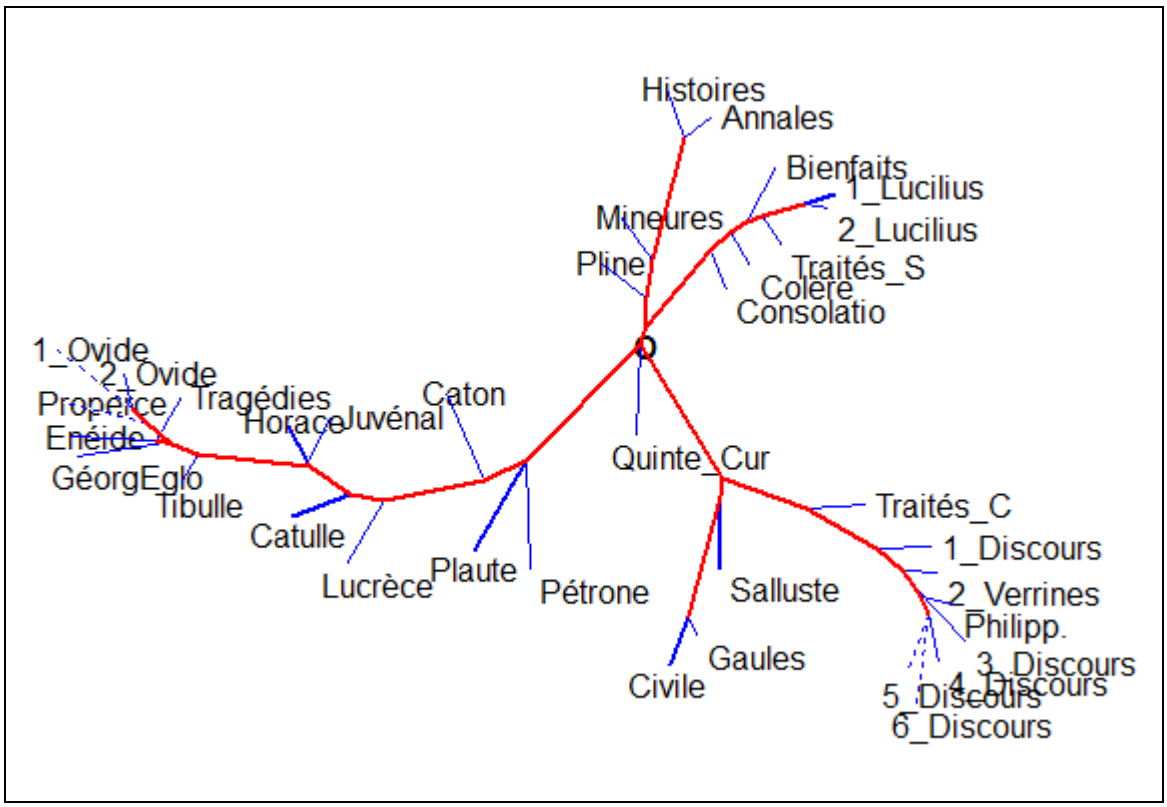


Figure 14. Analyse arborée de la littérature latine selon la mesure d'Évrard



VI - La connexion lexicale de Muller

En utilisant le mot *connexion*, Muller, comme Evrard, envisage plutôt le calcul positif, la similarité plutôt que la différence. Il est le seul à confronter l'observation à un modèle et à obtenir un test statistique, à partir du schéma d'urne. Pour les hautes fréquences, le calcul est simple. Quand on a cumulé les fréquences observées pour un même mot dans les deux textes que l'on confronte, la fréquence théorique du mot dans chacun est proportionnelle à la taille du texte, et l'écart avec la fréquence réelle s'apprécie par un CHI2 classique :

$$CHI2 = (reel - theo)^2 / theo$$

Les résultats individuels sont cumulés dans un CHI2 total tandis que le nombre des degrés de liberté est augmenté d'une unité. Mais pour les basses fréquences la démarche est plus délicate. Laissons Muller s'expliquer lui-même à la page 211 de son *Initiation à la statistique* (1968) :

FRÉQUENCES BASSES.

On dressera donc un tableau de distribution du vocabulaire de l'ensemble formé par la réunion des deux textes.

Connaissant l'étendue respective des deux textes A et B, qui est N_a et N_b, on calculera aisément la probabilité pour qu'une occurrence prise au hasard soit dans A ou dans B ; nous appellerons conventionnellement la première p et la seconde q, pour retrouver des notations déjà employées :

$$p = \frac{N_a}{N_a + N_b} \qquad q = \frac{N_b}{N_a + N_b} \qquad p + q = 1 .$$

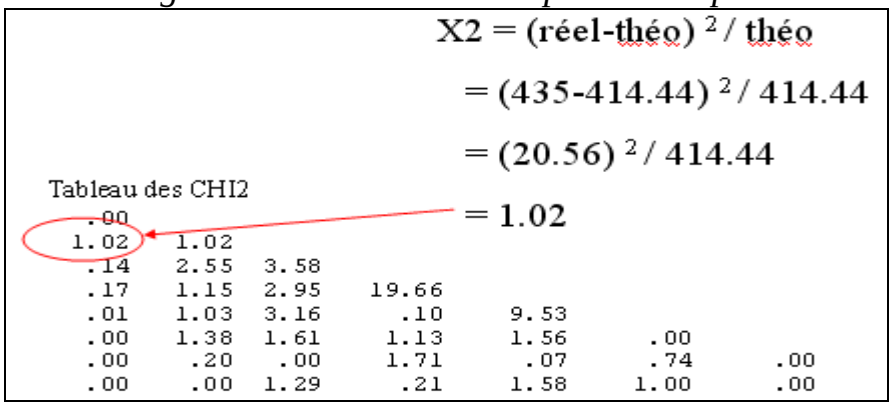
Il suffira ensuite d'appliquer les développements du binôme $(p + q)^f$ pour construire un modèle ; on appellera f la fréquence du vocable dans l'ensemble, et V_f l'effectif qui lui est associé ; f'_a et f'_b les sous-fréquences dans les deux textes, et V'_{fa}, V'_{fb} leurs effectifs.

La probabilité, pour un vocable de fréquence f, d'avoir les sous-fréquences 0, 1, 2... f dans l'un ou l'autre des textes est alors :

	Probabilité d'une sous-fréquence dans A					id. dans B				
f	0	1	2	3	4...	0	1	2	3	4...
1	q	p	0	0	0	p	q	0	0	0
2	q^2	$2pq$	p^2	0	0	p^2	$2pq$	q^2	0	0
3	q^3	$3pq^2$	$3p^2q$	p^3	0	p^3	$3p^2q$	$3pq^2$	q^3	0
4	q^4	$4pq^3$	$6p^2q^2$	$4p^3q$	p^4	p^4	$4p^3q$	$6p^2q^2$	$4pq^3$	q^4
etc.										

Le tableau est à construire pour toutes les confrontations des textes deux à deux. Si on met la barre à 50 entre les hautes et basses fréquences, le tableau s'agrandit à la dimension d'un triangle de base 50, soit 1225 cellules. Nous ne remplirons que la première et la plus simple, celle des hapax rencontrés au croisement de deux textes de taille équivalente pris chez Corneille. On en trouve 435 chez l'un et 392 chez l'autre, alors qu'on devrait en compter 414 chez l'un et l'autre.

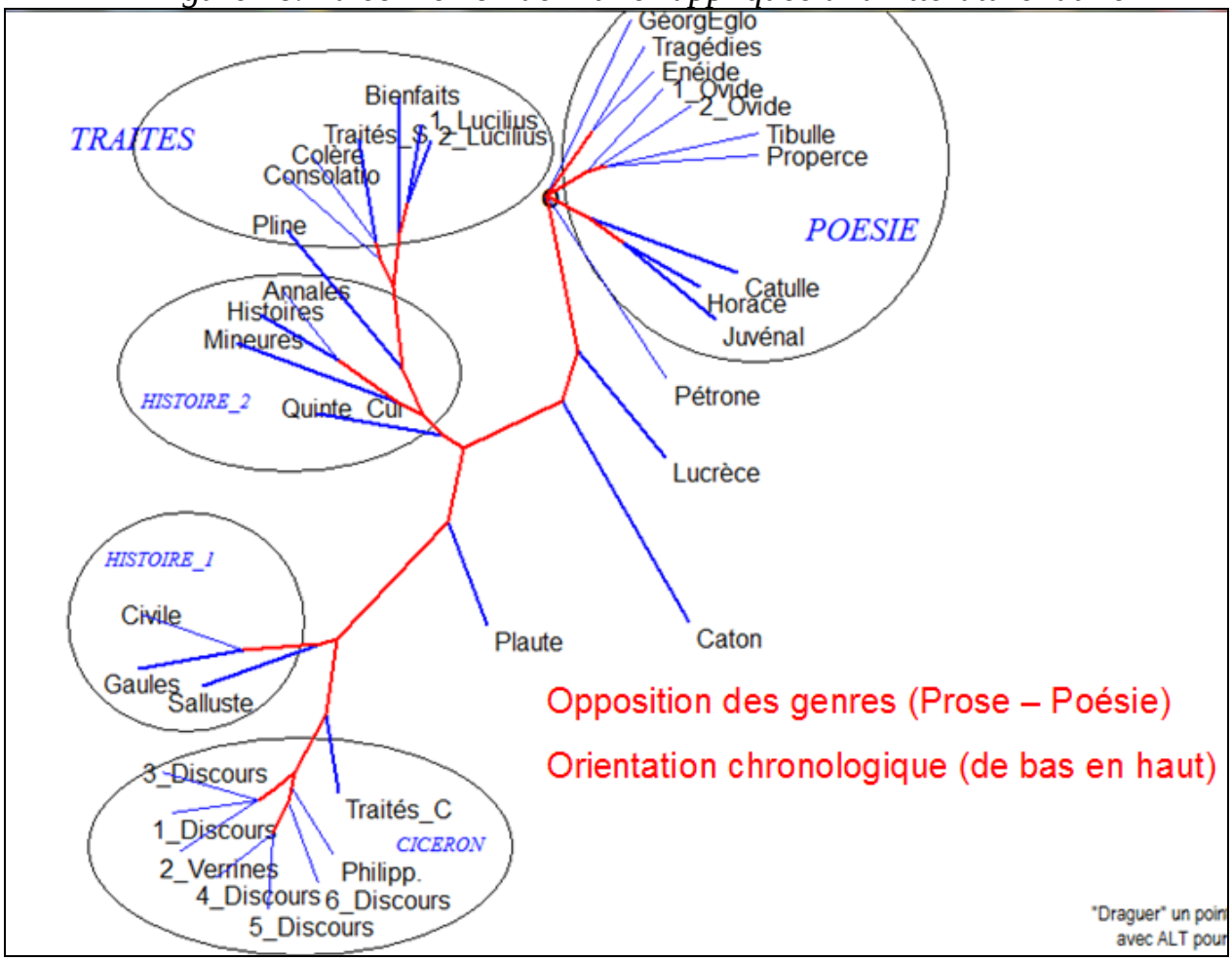
Figure 15. Détail du calcul pour les hapax



Dès la deuxième ligne le calcul se complique, avec les mots de fréquence 2, dont l'effectif doit être réparti en trois lots (p^2 pour deux occurrences en A, q^2 pour deux occurrences en B et *2pq* pour une occurrence dans A et dans B). Tant que la fréquence théorique est supérieure à 5, on détaille et cumule les CHI2 partiels. Au-delà, il y a regroupement.

Appliqué aux données du LASLA, le calcul propose une carte très lisible des auteurs latins, où l'opposition des genres se combine avec la marche du temps (figure 15).

Figure 16. La connexion de Muller appliquée à la littérature latine



Or dans les mesures de distance lexicale on ne sait trop quels facteurs agissent. Le choix des mots est gouverné par diverses influences qui se combattent ou s'appuient : le genre, l'époque, le sujet. Un indice global ne peut que les mêler sans permettre la décantation. Nous avons donc introduit dans le calcul de Muller un filtre qui aide à isoler les basses et les hautes fréquences, et peut-être, à travers cette distinction, les facteurs visés. On a vu précédemment que le calcul de la connexion lexicale se faisait en deux phases. La première ne tient compte que des fréquences basses (de 1 à 50). La seconde ne s'intéresse qu'aux autres fréquences. La figure 16 a fondé la synthèse sur le CHI2 final. Mais avant de réunir les deux lots, on peut examiner si les deux coupes se superposent. Abandonnons un instant la littérature latine où les oppositions de genre et d'époque sont trop violentes pour ne pas écraser les nuances discrètes. Prenons une œuvre homogène, comme celle de Molière, où les écarts de temps ou de genre sont relativement faibles. Or le spectroscope fournit deux images divergentes selon qu'on s'appuie sur les basses fréquences (figure 17) ou sur les hautes (figure 18). La première fait une distinction radicale entre les pièces écrites en prose, à droite, et celles, à gauche, où Molière utilise le vers, Car il y a des mots qui n'ont pas leur place dans une pièce en vers, d'autres qui sont permis dans une comédie mais non dans une tragédie. Ces interdits et ces privilèges concernent moins directement les mots fréquents et encore moins les mots-outils parce que leur emploi est inévitable dans tout discours et que l'ostracisme est plus difficile à leur endroit.

Figure 17. Connexion dans les basses fréquences. Influence du genre (prose- vers)

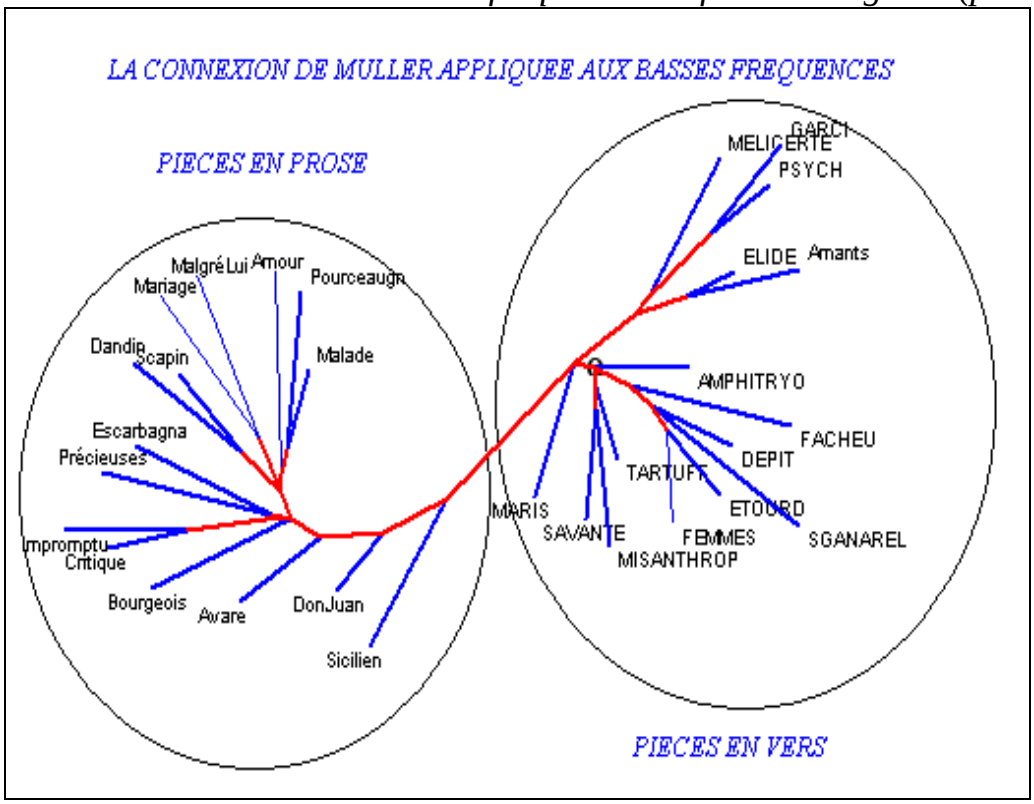
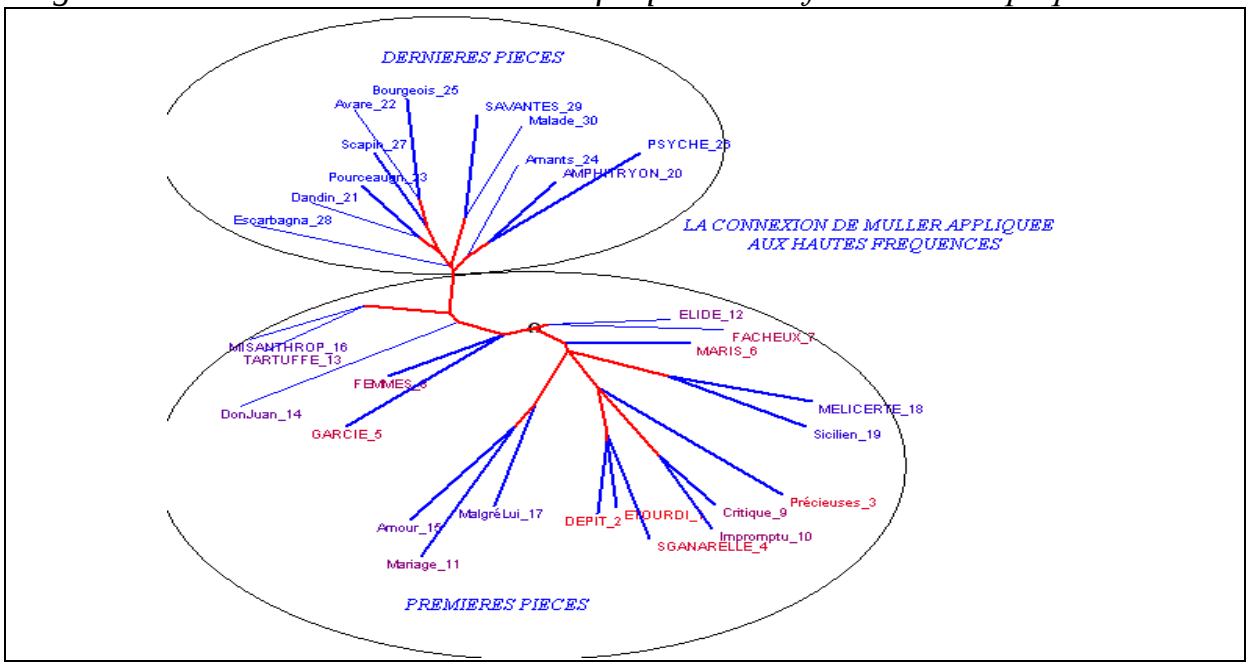
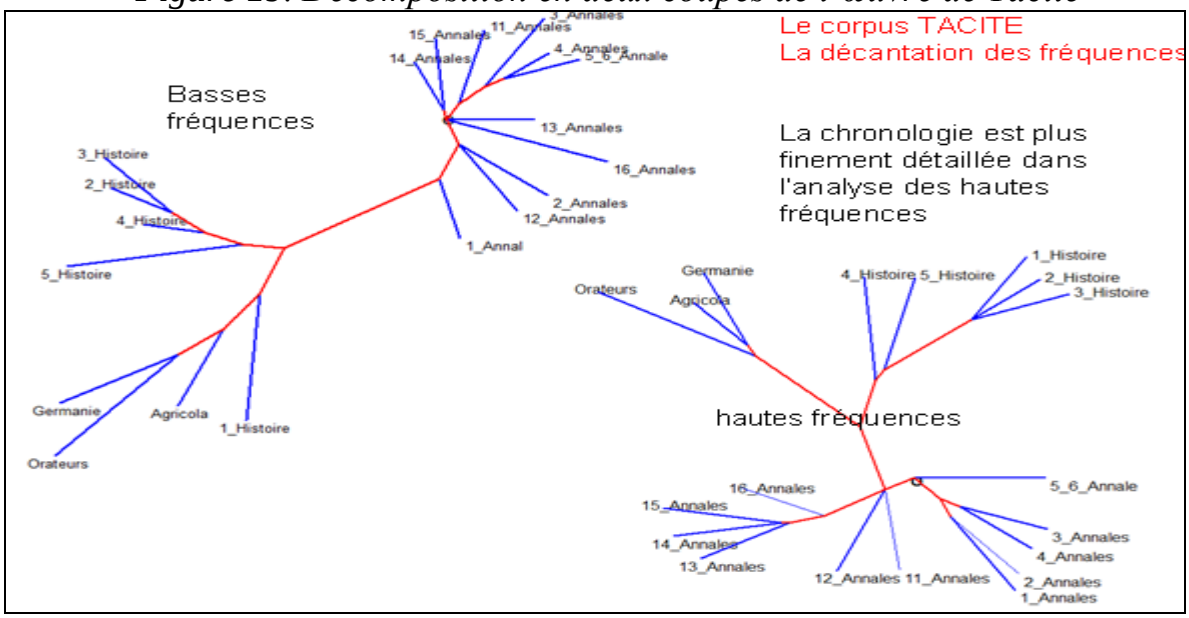


Figure 18. Connexion dans les hautes fréquences. Influence de l'époque



Mais les fréquences hautes n'en sont pas moins animées de mouvements qui paraissent plus lents mais plus profonds et qui décrivent sourdement l'évolution ou les variations de l'écriture. Ces mouvements de fond sont sans doute moins conscients ou moins volontaires que les choix clairs que l'écrivain fait parmi les genres et les sujets. Plus stylistiques que thématiques, ils sont davantage le reflet de la structure que du contenu. Le cas de Tacite se prête à cette analyse. Car son œuvre, mis à part le *Dialogue des Orateurs*, ne s'éloigne pas du genre historique. Les basses fréquences sont sensibles aux particularités qui s'attachent aux lieux, aux évènements et aux acteurs, et qui distendent le graphique 19 dans sa partie gauche. Les hautes fréquences (à droite) rendent compte plutôt de l'évolution de son écriture. On y distingue des phases plus détaillées et plus progressives dans la composition des Histoires et des Annales.

Figure 19. Décomposition en deux coupes de l'œuvre de Tacite



Conclusion

Les méthodes qu'on vient d'explorer viennent d'horizons différents et se fondent tantôt sur les fréquences absolues (Salem), tantôt sur des écarts pondérés (Labbé), sur des probabilités (Muller) ou des relevés de présence/absence (Évrard, Jaccard). Tantôt les occurrences ont le même poids, les mots fréquents ont alors l'avantage (Salem) ; tantôt l'égalité règne au niveau des vocables ; donnant aux basses fréquences l'influence déterminante. La figure 20 résume leurs défauts respectifs et les corrections proposées.

Figure 20. Récapitulatif des méthodes employées pour les calculs de proximité

Jaccard et Evrard:	Salem	Labbé:	Muller :
aucune correction	sondage partiel, limité aux hautes fréquences.	- exclusion des hapax dans le texte le plus long - exclusion des écarts inférieurs à 0.5 - exclusion des inégalités trop fortes	déformation propre au CHI2 réel théo écart chi2 9 10 1 0.1 90 100 10 1 900 1000 100 10 Correction:pondération par la racine carrée des deux vocabulaires CHI2 corr. = CHI2 / (√Va * √Vb)
les mots fréquents sont hors-jeu, étant toujours dans la zone commune	la loi du plus fort (les textes les plus longs, les mots les plus fréquents)		

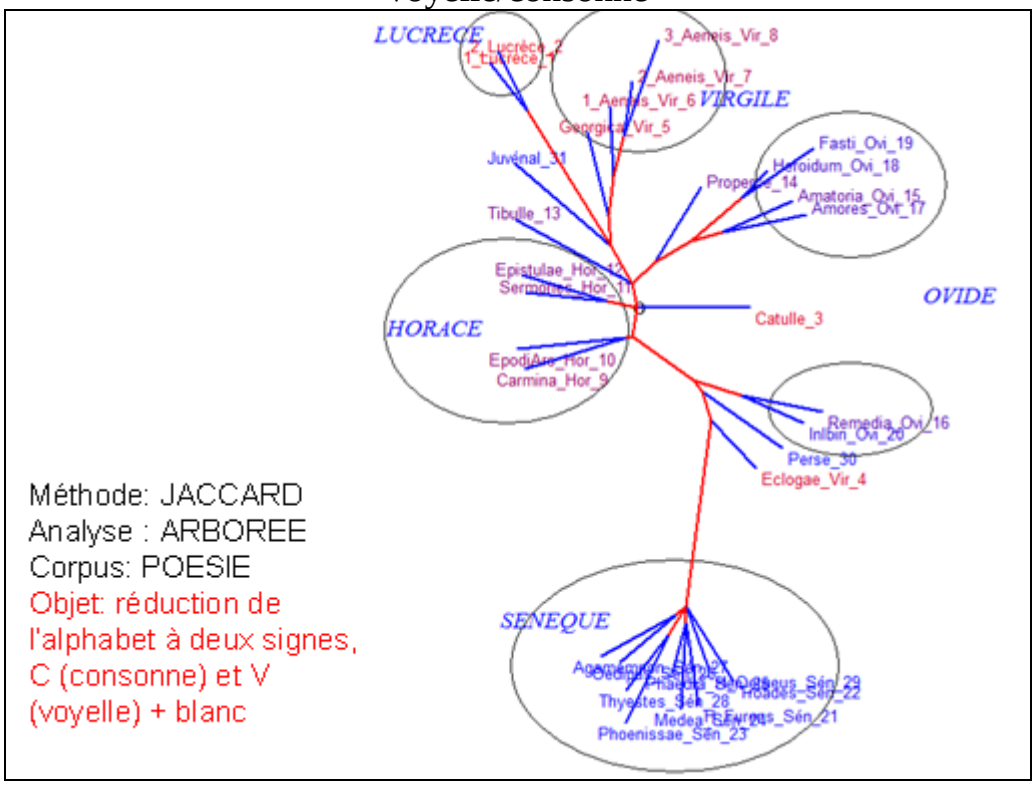
L'étude du vocabulaire est la voie la plus favorable à l'automatisation et aux recherches sur l'attribution ou la datation des œuvres littéraires. Mais affinités et distances ne sont pas propres au lexique. Elles peuvent se calculer sur bien d'autres objets linguistiques : des faits de syntaxe ou de sémantique, des rythmes, des sonorités, des figures de style. La garantie des mesures n'est pas dans l'infaillibilité d'un hypothétique ADN lexical mais dans la convergence des approches et des résultats.

Un dernier exemple, inattendu, tend à confirmer la solidité et la constance des forces qui structurent l'aimantation ou la gravitation des mots et qui s'exercent pareillement sur les graphies et les lemmes, ou même sur les codes grammaticaux d'où le sens est exclu, ou enfin sur les ngrammes, où le lexique est corrompu et les lettres couchées sur un lit de Procuste, par groupe de quatre. Procédons à une expérience ultime qui va jusqu'à détruire l'alphabet pour n'en retenir que la distinction voyelle/consonne. Les premiers vers du *De Natura*

rerum reproduits ci-dessous font l'objet d'une réduction drastique qui ne conserve que deux codes – à quoi s'ajoute le blanc.

> *Aeneadum genetrix hominum diuom que uoluptas Alma*
> *Venus caeli subter labentia signa Quae mare nauigerum*
> *quae terras frugiferentis Concelebras per te*
>
> *VVCVVCVC CVCVCCVC CVCVCVC CVVVC CVV VVCVCCVC*
> *VCCV CVCVC CVVCV CVCCVC CVCVCCVV CVCCV CVVV*
> *CVCV CVVVCVCVC CVVV CVCCVC CCVCVCVCVCCVC*
> *CVCCVCVCCVC CVC CV.*

Figure 21. Les affinités des poètes latins survivant à un codage voyelle/consonne



Tous les textes de la poésie ont subi ce traitement réducteur, qu'il ne faut pas confondre avec la transcription binaire où les combinaisons des codes 1 et 0 reproduisent les caractères sans perte d'information. Dans un texte réduit en cendres, la perte est ici radicale et l'ambiguïté maximale : les mots de deux lettres n'ont le choix qu'entre trois combinaisons : CV, VV et VC. Ceux de trois lettres ont des solutions à peine plus ouvertes. Les mots plus longs ne sont pas mieux traités : avec un alphabet aussi pauvre, les combinaisons rares se raréfient encore et il ne reste plus beaucoup d'hapax.

On peut douter qu'Étienne Évrard se fût prêté à une tentative aussi saugrenue. Pourtant son algorithme dénoue ce rébus opaque avec autant de facilité que la typologie des langues bantoues. Lucrèce et Sénèque tiennent les deux bouts de la chaîne chronologique, où s'échelonnent les autres poètes, Virgile, Horace, Ovide, chacun rassemblant ses œuvres autour de lui. Dans un brouillard intense où l'œil humain ne distingue rien dans le paysage, la statistique reconnaît les sentiers et les troupeaux.