Revue, Informatique et Statistique dans les Sciences humaines
XXI, 1-4, 1985, pp. 83-107

Etendue du lexique et distribution des classes de fréquence chez quelques auteurs latins

Essai d'analyse factorielle

Michel DUBROCARD

Parmi les différents critères statistiques qui permettent de caractériser le style d'une oeuvre, ou d'un auteur, l'étendue du lexique suscite, depuis fort longtemps, un intérêt particulier. Certains écrivains auraient recours à un vocabulaire riche et varié, d'autres, au contraire, utiliseraient un lexique très réduit.

Les développements récents de la statistique linguistique ont permis de revoir ces jugements très subjectifs, et parfois mal fondés (1), et de proposer une évaluation beaucoup plus rigoureuse de l'étendue lexicale, désormais étroitement liée à la notion de classes de fréquence.

Dès ses premières publications, l'équipe du L.A.S.L.A. a jugé nécessaire de compléter l'index traditionnel par des relevés statistiques qui comprenaient un tableau de distribution du vocabulaire. Sur ce point, comme sur beaucoup d'autres, le L.A.S.L.A. faisait ainsi figure de précurseur.

Nous souhaitons ici utiliser quelques-uns de ces relevés pour comparer entre eux un certain nombre d'auteurs latins en retenant deux aspects de leur oeuvre, l'étendue du lexique et l'utilisation des classes de fréquence.

1. Etendue du lexique et distribution des classes de fréquence : aperçu théorique

1.1. *Mesure de l'étendue lexicale*

Sans revenir sur le détail de méthodes de calcul désormais bien connues (2), nous rappellerons simplement que si l'on représente par N le nombre total de mots, et par V le nombre de mots différents, que l'on désigne généralement sous le nom de vocables ou lemmes, le rapport V/N constitue une première évaluation de l'étendue du lexique. Une description plus précise de la structure lexicale fait intervenir l'effectif de chaque classe de fréquence.

Ainsi on dénombrera V_1 vocables utilisés une seule fois, V_2 vocables utilisés deux fois, V_n vocables utilisés n fois. On parviendra finalement à l'équation

$$N = 1*V_1 + 2*V_2 + 3*V_3 + \ldots + n*V_n$$

ou encore

$$N = \sum_{i=1}^{n} i\ V_i$$

1.2. *Comparaison de deux oeuvres ou de deux auteurs*

Il est clair que le vocabulaire V s'accroît moins vite que le nombre de mots N : un texte très court contiendra des mots tous, ou presque tous, différents; un texte très long comptera un grand nombre de répétitions. Si l'on veut comparer deux textes de longueur différente, il faut recourir au calcul d'un vocabulaire théorique, suivant la méthode proposée à peu près simultanément, il y a quelques années, par Charles Muller à Strasbourg, et par Etienne Evrard à Liège.

Nous nous dispenserons d'exposer ici une méthode devenue classique (3), pour n'en retenir que les résultats : partant du texte le plus long, on calcule le vocabulaire théorique d'un fragment de ce texte de même longueur que le texte le plus court. En rapprochant le résultat de ce calcul du vocabulaire réel du texte le plus court, on observe généralement un écart, positif ou négatif, qui traduit, du point de vue de l'étendue du lexique, la supériorité, ou l'infériorité, du premier texte.

Cette méthode exige la comparaison des textes deux à deux. Si nous l'avions appliquée aux 36 textes que nous avons l'intention d'examiner ici, nous aurions dû effectuer plusieurs centaines de comparaisons. Nous pensons avoir démontré ailleurs (4) qu'il était possible de comparer beaucoup plus simplement oeuvres et auteurs en calculant pour chacun d'eux le vocabulaire théorique d'une tranche de 1 000 mots.

Partant soit de l'ensemble de l'oeuvre de Juvénal, soit d'une seule satire, soit encore d'un fragment du texte arbitrairement découpé, nous avons calculé le vocabulaire théorique d'une tranche de 1000 mots, en utilisant successivement la formule binomiale et un programme de simulation aléatoire. Les résultats de ces calculs, à peu près identiques, quelle que soit la méthode utilisée, sont fort proches des effectifs réels, mesurés suivant différents procédés.

Plus précisément, lorsque les tranches de 1 000 mots sont découpées à l'intérieur du texte en respectant l'ordre des mots, on observe un vocabulaire réel légèrement inférieur au vocabulaire théorique, du fait du phénomène bien connu de la spécialisation lexicale.

En revanche, lorsqu'il s'agit de tranches aléatoires, où les mots sont réunis sans que l'on tienne compte de leur ordre réel, les effectifs observés sont extrêmement proches des effectifs théoriques calculés soit par la formule binomiale, soit par notre programme de simulation aléatoire.

Fort de cette expérience, nous nous proposons ici de comparer un certain nombre d'oeuvres ou d'auteurs en utilisant comme données l'effectif des vocables et la distribution des fréquences dans une tranche de 1 000 mots. Pour ce calcul nous avons choisi d'utiliser la formule binomiale; l'utilisation du programme de simulation aléatoire donne des résultats très voisins.

2. Présentation des données

2.1. *Tableau d'ensemble*

Le tableau 1 présente le détail des données que nous avons utilisées.

Tableau 1

Oeuvres	Voc.	F1	F2	F3	F4	F5	F6	F7-8	F9-12	F13-20	F > 20
SATIRE I	656,0	513,6	88,7	24,2	11,5	4,8	1,7	3,1	2,1	4,0	1,9
SATIRE II	650,7	507,0	96,0	20,9	7,7	2,4	2,2	2,8	7,3	3,1	1,0
JUVENAL	646,2	511,5	80,5	23,2	9,6	5,1	3,2	4,0	3,9	3,0	1,8
HORACE	643,4	492,2	96,2	27,6	10,0	4,4	2,3	2,5	3,1	2,5	2,0
PERSE	640,9	509,0	79,7	23,2	8,9	4,2	2,4	2,9	3,4	4,0	2,5
TACITE	640,8	495,8	85,6	27,3	11,8	6,1	3,4	3,4	2,4	2,4	2,0
SATIRE XI	638,7	510,4	73,6	23,3	9,6	4,8	3,1	4,5	3,4	3,9	1,7
SATIRE VIII	638,6	500,9	81,3	24,2	10,3	5,6	3,3	3,6	3,7	3,3	1,9
SATIRE XV	634,9	476,1	94,2	27,8	15,0	5,9	5,6	4,4	2,6	1,7	1,2
SATIRE III	633,9	499,3	76,0	23,9	10,9	5,8	3,6	4,4	4,7	3,7	1,2
SATIRE XIII	632,6	495,3	78,7	25,2	12,0	6,0	2,7	2,3	3,9	4,5	1,5
SATIRE VI	632,1	498,4	78,4	22,9	10,0	5,6	3,4	3,8	3,6	3,3	2,3
SATIRE IX	631,0	484,4	93,5	23,4	12,2	4,6	1,8	1,8	2,5	4,3	1,9
SATIRE VII	630,5	489,5	81,6	25,8	11,3	5,9	3,6	4,1	4,1	2,1	2,0
SATIRE X	626,5	484,9	84,6	25,4	9,8	5,0	3,2	4,0	3,4	3,9	1,9
SATIRE XIV	625,4	483,3	80,7	27,4	12,7	6,1	3,2	3,7	3,3	2,6	2,1
SENEQUE TH.	623,1	459,7	95,1	32,1	13,8	6,8	3,7	3,8	3,6	2,7	1,2
GEORGIQUES	617,4	467,3	89,5	27,7	11,6	6,2	3,9	4,6	3,3	0,9	2,0
SATIRE V	615,3	480,6	68,7	28,1	11,1	5,8	6,5	5,9	4,1	1,7	2,4
PHEDRE	602,0	450,7	84,3	28,7	11,0	9,7	2,7	4,0	5,0	4,3	1,7
TIBULLE	592,5	433,7	90,1	30,5	13,2	6,8	3,9	4,2	3,8	3,6	2,1
OVIDE	590,7	439,2	89,0	28,9	12,4	6,2	3,5	3,6	2,6	1,8	2,9
PROPERCE	588,3	439,8	83,5	27,9	12,1	6,2	3,6	4,2	4,4	3,6	2,7
ENEIDE	586,8	426,3	91,3	31,6	14,1	7,5	4,4	4,6	3,4	1,1	2,0
CATULLE	579,9	434,7	76,6	26,4	13,0	7,6	4,8	5,4	4,6	4,1	2,1
AD MARCIAM	552,8	411,3	72,5	25,8	12,8	7,6	5,0	5,9	4,4	3,5	3,5
DE BREVITATE	542,5	401,4	70,3	24,9	13,0	8,2	5,6	6,8	5,2	3,3	3,5
AD HELVIAM	537,5	393,4	74,5	26,1	12,2	7,1	4,8	6,2	5,5	3,8	3,6
BUCOLIQUES	529,9	368,7	82,8	30,3	14,7	8,5	5,4	6,3	6,1	4,4	2,2
TITE-LIVE	525,8	373,4	77,1	28,1	13,7	8,1	5,4	6,7	5,9	3,8	3,0
JUGURTHA	501,2	336,5	82,2	31,4	15,3	8,9	5,9	7,3	6,4	3,8	3,0
CATILINA	498,5	336,0	79,2	30,3	15,4	9,5	6,5	8,2	6,6	3,5	2,9
AD LUCILIUM	498,0	360,7	65,3	23,9	12,6	8,2	5,9	7,7	5,9	3,1	4,1
CICERON	477,4	338,5	64,9	23,9	12,5	8,0	5,9	8,0	7,2	3,8	4,3
AD POLYBIUM	472,3	329,4	64,9	25,5	13,7	8,8	6,2	7,5	5,7	5,6	4,3
CESAR	464,0	312,7	72,8	27,6	13,7	8,1	5,3	6,5	6,2	6,1	4,6

Chaque ligne du tableau correspond à une oeuvre ou à un auteur étudié par le L.A.S.L.A.. Cette source unique assure une parfaite homogénéité des critères de lemmatisation. La plupart des tableaux de distribution des fréquences que nous avons utilisés ont été publiés dans *Revue*, n° 1, 1974, pp. 34 *sqq*. Il s'agit du *Bellum Gallicum* de César, d'extraits des *Métamorphoses* d'Ovide, d'un choix de *Fables* de Phèdre, du *De Coniuratione Catilinae* et du *Bellum Jugurthinum* de Salluste, des *Bucoliques*, des *Georgiques*, et des six premiers chants de l'*Enéide* de Virgile, des *Odes* d'Horace, du *Corpus Tibullianum*, des consolations *Ad Marciam, Ad Helviam* et *Ad Polybium* de Sénèque, auxquelles s'ajoutent les *Lettres à Lucilius*, les *Carmina* de Catulle, des extraits de Cicéron, Tite-Live et Tacite.

Nous avons complété ces données par des relevés concernant les treize satires de Juvénal qui comptent plus de 1 000 mots, ainsi que par des tableaux relatifs aux oeuvres théâtrales de Sénèque, aux *Elégies* de Properce et aux *Satires* de Perse. Ces derniers textes ont été étudiés par le L.A.S.L.A., mais n'ont pas encore fait l'objet de publications.

Il est à noter que les oeuvres que nous avons utilisées ont une longueur comprise entre 1 000 et 118 000 mots, ce dernier chiffre correspondant aux *Lettres à Lucilius*. Comme les autres oeuvres dépassent rarement 30 000 mots, il nous a paru légitime de leur appliquer un procédé de calcul dont nous avions pu vérifier la validité sur l'ensemble des *Satires* ainsi que sur des tranches allant de 1 000 à 25 000 mots.

Les colonnes du tableau 1 présentent successivement l'effectif de l'ensemble des vocables, puis celui des mots employés 1 fois, 2 fois, 3 fois, 4 fois et 5 fois; nous avons ensuite regroupé les fréquences allant de 6 à 7, de 8 à 11, de 12 à 19, de 20 à 35, enfin celles qui sont supérieures à cette dernière valeur. Il est à noter que dans ces regroupements nous avons chaque fois doublé le nombre des classes de fréquence concernées afin d'obtenir des effectifs à peu près comparables.

2.2. *Classement par ordre d'étendue lexicale décroissante*

Pour rendre plus faciles d'éventuels rapprochements, nous avons choisi de classer oeuvres et auteurs en fonction de l'étendue du lexique (colonne 1).

Bien entendu ce classement n'a qu'une valeur indicative. En règle générale les écarts observés entre deux lignes successives sont trop faibles pour que l'on puisse en conclure à une réelle supériorité de telle ou telle oeuvre sur l'oeuvre suivante.

Si l'on admet que le mode de calcul que nous avons adopté donne une bonne approximation de ce que seraient des effectifs réellement observés, on peut appliquer à ces données un test de X^2 pour mesurer la probabilité que les écarts ne correspondent pas à de simples variations aléatoires. Au seuil de 5 % , l'écart significatif est de 25 pour les premiers de la liste, de 20 pour les derniers.

En d'autres termes, s'il est impossible de départager Juvénal et Horace, ou encore Juvénal et Perse, il existe un écart significatif entre Juvénal et Tacite, ou encore entre *Georgiques* et *Enéide*, entre *Enéide* et *Bucoliques*. Entre les deux extrêmes, Juvénal et César, l'écart atteint près de 200 vocables. La signification stylistique de l'étendue lexicale apparaît ici très clairement.

2.3. *Classement en fonction de l'effectif des classes de fréquence*

Le classement que nous venons de présenter ne tenait compte que de l'effectif total des vocables. L'effectif de chaque classe de fréquence permettrait d'établir un classement différent.

Pour la fréquence 1, par exemple, le classement obtenu serait assez proche de celui qui est fondé sur le nombre total de vocables. Rien d'étonnant à cela : les textes dont le vocabulaire est le plus étendu font appel à un fort pourcentage de mots de fréquence 1. Chez Juvénal, pour une tranche de 1 000 mots, les hapax représentent près de 80 % du total des vocables, et 51 % du total des occurrences. Chez César, ils représentent 67 % du total des vocables, et seulement 32 % du total des mots.

Ce classement est à peu près inversé si l'on prend en considération les fréquences élevées. Ainsi pour les fréquences supérieures à 20, la première place est occupée par la *Consolation à Polybe*, où ces vocables sont au nombre de 9,9 contre 4,8 chez Juvénal.

2.4. *Représentation graphique de la distribution des fréquences*

Pour permettre de saisir plus clairement la place qu'occupe chaque classe de fréquence dans une tranche de 1 000 mots, nous avons calculé non plus l'effectif des vocables, mais celui des occurrences correspondant à chaque classe de fréquence. Pour simplifier un peu, nous avons limité le nombre des catégories à 8, en groupant les fréquences qui vont de 6 à 8, de 9 à 20, ainsi que celles qui sont supérieures à 20. Le tableau 2 présente les résultats de ce calcul. Les différentes oeuvres ont été classées en tenant compte de l'effectif des mots de fréquence 1. On observera que ce classement est fort proche de celui que nous avions obtenu en partant de l'effectif total du vocabulaire (Tableau 1).

Nous avons ensuite utilisé un programme de représentation graphique qui permet de faire apparaître la part relative de chaque classe de fréquence sous la forme d'un secteur circulaire. C'est le fameux "camembert" qui illustre les journaux au lendemain des consultations électorales, ou encore le *pie-chart* des statisticiens anglo-américains.

Pour ne pas multiplier les graphiques, nous ne proposerons ici que deux illustrations de cette méthode.

La figure 1 permet de comparer la structure du vocabulaire chez deux auteurs tout à fait opposés, Juvénal et César. Le graphique met en évidence ce qui différencie les deux écrivains : les mots de fréquence 1 représentent chez Juvénal plus de la moitié des occurrences, moins du tiers chez César. Le poète conserve encore l'avantage sur le prosateur pour les mots de fréquence 2 (161 contre 146), mais dès la fréquence 3 César, à son tour, prend la première place. L'écart entre les deux auteurs, d'abord faible, ne cesse de grandir, et à partir de la fréquence 6 l'effectif des occurrences est presque deux fois plus élevé chez César : 81 contre 49 pour les fréquences 6 à 8, 162 contre 88 pour les fréquences 9 à 20, 119 contre 56 pour les fréquences supérieures à 20.

Il est difficile d'imaginer opposition plus nette.

Mais même lorsque les différences sont moins marquées, une représentation graphique met en évidence ressemblances et oppositions. Ainsi la figure 2 réunit quatre oeuvres poétiques, les *Bucoliques*, les *Géorgiques* et l'*Enéide* de Virgile, les *Odes* d'Horace.

Tableau 2

Oeuvres	F1	F2	F3	F4	F5	F6-8	F9-20	F > 20
SATIRE I	514	177	73	46	24	34	82	50
JUVENAL	512	161	70	39	26	49	88	56
SATIRE XI	510	147	70	38	24	53	98	59
PERSE	509	160	70	36	21	37	101	66
SATIRE II	507	192	63	31	12	35	122	38
SATIRE VIII	501	163	73	41	28	47	91	55
SATIRE III	499	152	72	44	29	55	107	42
SATIRE VI	498	157	69	40	28	49	90	68
SATIRE XIII	495	157	76	48	30	34	115	44
HORACE	492	193	83	40	22	33	72	64
SATIRE VII	490	163	78	45	30	52	76	66
SATIRE X	485	169	76	39	25	49	99	56
SATIRE IX	484	187	70	49	23	25	97	64
SATIRE XIV	483	161	82	51	31	47	76	68
SATIRE V	481	138	84	45	29	83	70	71
SATIRE XV	476	189	84	60	30	66	55	40
GEORGIQUES	467	179	83	47	31	58	48	86
SENEQUE TH.	460	190	96	56	34	51	81	31
PHEDRE	448	172	81	47	31	56	108	56
TACITE	446	182	91	51	32	48	76	73
PROPERCE	440	167	84	49	31	53	104	73
OVIDE	439	178	87	50	31	48	56	110
CATULLE	435	153	79	52	38	70	113	59
TIBULLE	434	180	92	53	34	56	100	52
ENEIDE	426	183	95	57	38	61	52	88
AD MARCIAM	411	145	77	52	38	75	101	100
DE BREVITATE	401	141	75	52	41	84	107	98
AD HELVIAM	393	149	78	49	36	76	117	102
TITE-LIVE	373	154	84	55	41	83	122	87
BUCOLIQUES	369	166	91	59	43	80	133	60
AD LUCILIUM	361	131	72	51	41	93	110	141
CICERON	339	130	72	50	40	95	134	140
JUGURTHA	337	164	94	62	45	90	126	82
CATILINA	336	159	91	62	48	101	123	81
AD POLYBIUM	329	130	77	55	44	94	150	121
CESAR	313	146	83	55	41	81	163	119

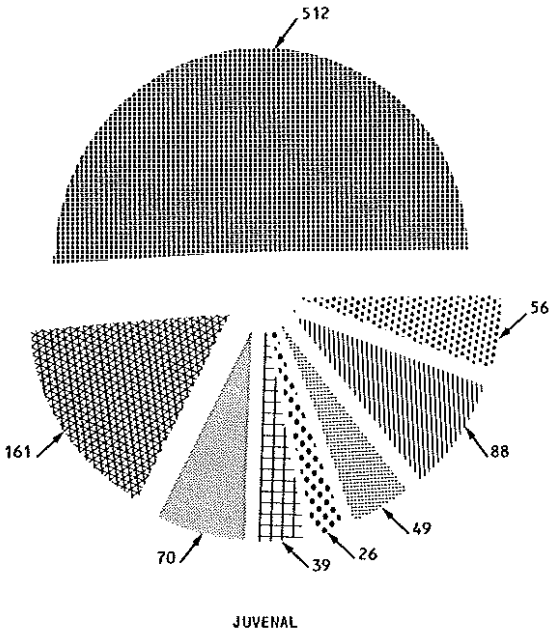
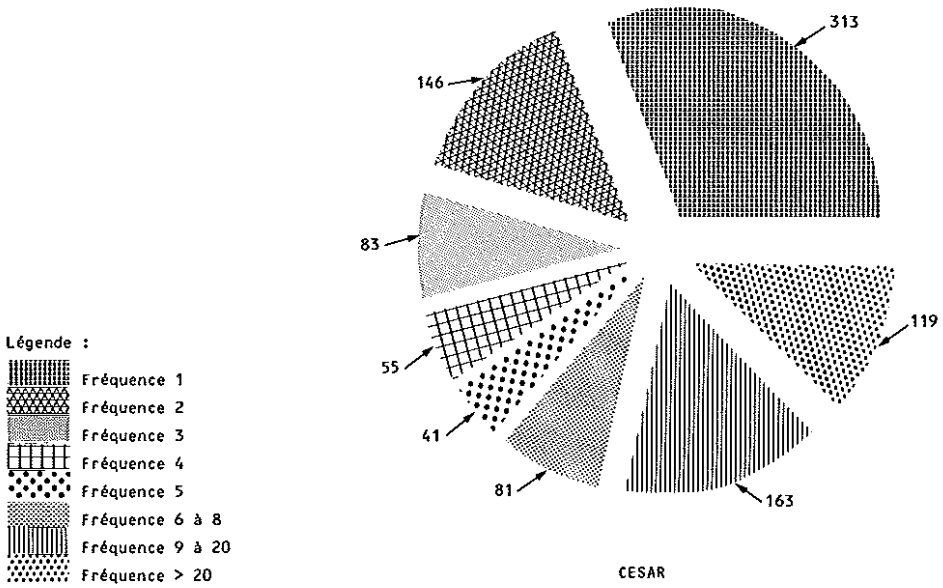


Figure 1



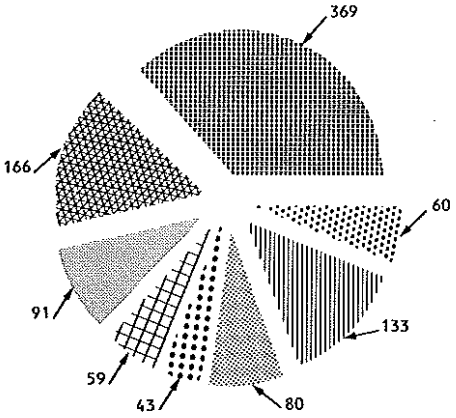
Même si les écarts sont moins importants que ceux que révélait la confrontation de Juvénal et de César, des oppositions se manifestent très clairement. Ainsi on observe chez Virgile une pauvreté relative du vocabulaire des *Bucoliques*, particulièrement marquée par le déficit des mots de fréquence 1. Cette particularité est sans doute due à l'unité thématique du recueil, peut-être aussi à une recherche délibérée de la sobriété.

A l'opposé les *Géorgiques* utilisent un lexique très étendu, qui comprend un effectif élevé de mots de fréquence 1. Cette richesse du vocabulaire est évidemment liée à la variété des thèmes abordés, sans doute aussi à l'utilisation d'un vocabulaire technique.

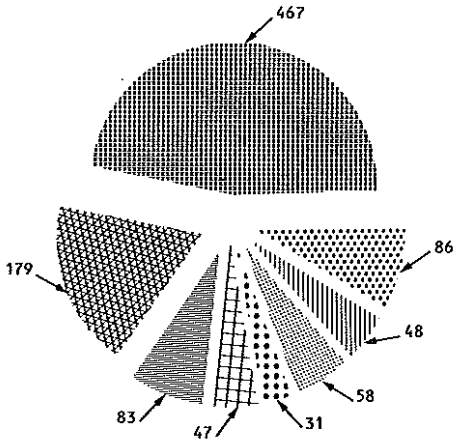
Entre ces deux extrêmes, l'*Enéide* occupe une position médiane. Il est à noter qu'au-delà de la fréquence 1 les oppositions sont beaucoup moins marquées, et les effectifs assez proches, sauf pour les mots de fréquence 9 à 20 qui sont exceptionnellement nombreux dans les *Bucoliques* (132 contre 48 dans les *Géorgiques* et 52 dans l'*Enéide*). Cette abondance est compensée par un déficit des mots (60) dont la fréquence est supérieure à 20. Cet effectif est comparable à celui que l'on avait observé chez Juvénal (56) et à celui que l'on va rencontrer chez Horace (64). Pour expliquer cette particularité il conviendrait d'examiner l'usage que fait Virgile des mots grammaticaux à fréquence élevée (conjonction de coordination et de subordination, en particulier). Il est intéressant de noter à ce propos que dans une étude consacrée à la subordination en latin (*Revue*, n° 1 à 4, 1981, p. 15) le pourcentage de verbes subordonnés par rapport au nombre total de verbes apparaît toujours plus faible chez les poètes que chez les prosateurs. Dans un classement en ordre décroissant, les dernières places sont occupées par les *Bucoliques*, l'*Enéide*, et les tragédies de Sénèque.

Les *Odes* d'Horace, qui complètent la figure 2, sont caractérisées par l'importance des mots de fréquence 1 (492, à rapprocher des 512 hapax des *Satires* de Juvénal). Pour le reste, la structure du lexique des *Odes* n'est pas très éloignée de celle que l'on observe dans l'*Enéide*.

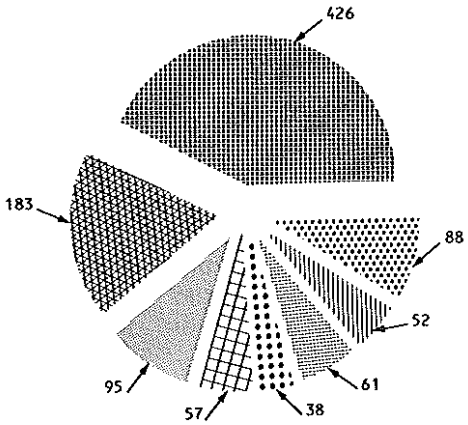
Nous aurions pu multiplier ces tableaux, qui permettent de rapprocher facilement des oeuvres ou des auteurs différents. Remarquons, d'ailleurs, qu'il n'est pas nécessaire de disposer



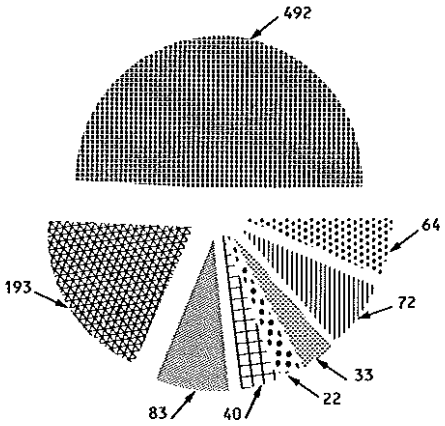
BUCOLIQUES



GEORGIQUES



ENEIDE



HORACE

Figure 2

d'un ordinateur pour construire de tels graphiques : un compas et un rapporteur suffisent.

Cependant nous avons préféré faire appel à une autre méthode pour comparer entre eux les auteurs et les oeuvres que nous avons choisi d'étudier ici; il s'agit de l'analyse factorielle.

3. Essai d'analyse factorielle

3.1. *Choix des données*

Ici encore on nous pardonnera de ne pas donner un exposé théorique de ces méthodes mathématiques dont l'utilisation, dans le domaine des sciences humaines et de la linguistique, devient de plus en plus fréquente (5).

Nous rappellerons seulement que l'analyse factorielle des correspondances permet de rendre compte d'un ensemble de données, présentées sous forme d'un tableau à double entrée, en rapprochant ou en opposant les différents éléments qui se trouvent en tête de ligne ou de colonne. Il s'agira ici d'une part des oeuvres ou des auteurs latins que nous avons choisi d'étudier, d'autre part de l'effectif des occurrences correspondant aux différentes classes de fréquence, en tenant compte de quelques regroupements. Ces données figurent sur le tableau 2.

Le programme d'analyse factorielle que nous avons utilisé (6) permet d'établir un graphique sur lequel chaque oeuvre et chaque classe de fréquence sera représentée par un point; au total le graphique comportera donc 44 points.

Notons cependant que 23 oeuvres ou auteurs seulement ont été pris en compte pour la construction du graphique. Comme l'oeuvre de Juvénal, qui apparaît à la fois dans sa totalité, et sous la forme de chacune des 13 satires comptant plus de 1000 mots, risquait d'exercer une influence excessive sur l'ensemble de la représentation, nous avons choisi de ne retenir que la totalité de l'oeuvre. Les 13 satires seront bien situées sur le graphique, mais n'auront pas contribué à sa construction.

3.2. *Etablissement du graphique*

Suivant une procédure bien connue, le programme d'analyse factorielle propose d'abord un tableau des valeurs propres de la matrice. Ce tableau, que nous nous dispenserons de reproduire ici, indique que le premier facteur (axe horizontal du graphique) représente 70,2 % de la variance de l'ensemble du système, le second facteur (axe vertical) 17,2 % . Au total ces deux facteurs permettent donc de rendre compte de 87,4 % de la variance du système. L'intervention d'un troisième facteur, représentant 9,35 % de la variance permet d'améliorer encore la représentation, en couvrant 96,8 % de la variance de l'ensemble. Nous examinerons plus loin s'il est nécessaire de faire intervenir des facteurs supplémentaires.

La figure 3 présente le graphique fondé sur la représentation du premier et du deuxième facteur.

Comment interpréter ce graphique ?

Si l'on examine d'abord les classes de fréquence, on constate qu'à l'exception des deux classes les plus élevées, qui sont isolées aux deux extrémités de l'axe vertical, les autres classes de fréquence apparaissent, en ordre régulièrement décroissant, à proximité de l'axe horizontal; la fréquence 5 se situe à la gauche et la fréquence 1 à la droite du graphique.

En ce qui concerne les écrivains, on note que, à l'exception des *Bucoliques*, toutes les oeuvres poétiques sont regroupées dans la partie droite du graphique, que tous les prosateurs, à l'exception de Tacite, figurent sur la partie gauche. Parmi les poètes, Horace, Perse, Juvénal et Sénèque auteur tragique se situent à l'extrême droite; parmi les oeuvres en prose, les extraits de Cicéron, la *Consolation à Polybe*, la *Guerre des Gaules* sont les plus proches de la marge de gauche.

Il s'agit là des oeuvres ou des auteurs qui présentent entre eux les plus fortes oppositions. Mais le graphique révèle aussi des similitudes par la proximité de certains points.

La plupart de ces ressemblances étaient prévisibles : rien de surprenant dans la proximité de *Catilina* et de *Jugurtha*, dans celle de la *Consolation à Marcia*, de la *Consolation à Helvia* et du *De Breuitate*, dans celle de Perse et de Juvénal. De même on voit les différentes satires se rassembler autour du point

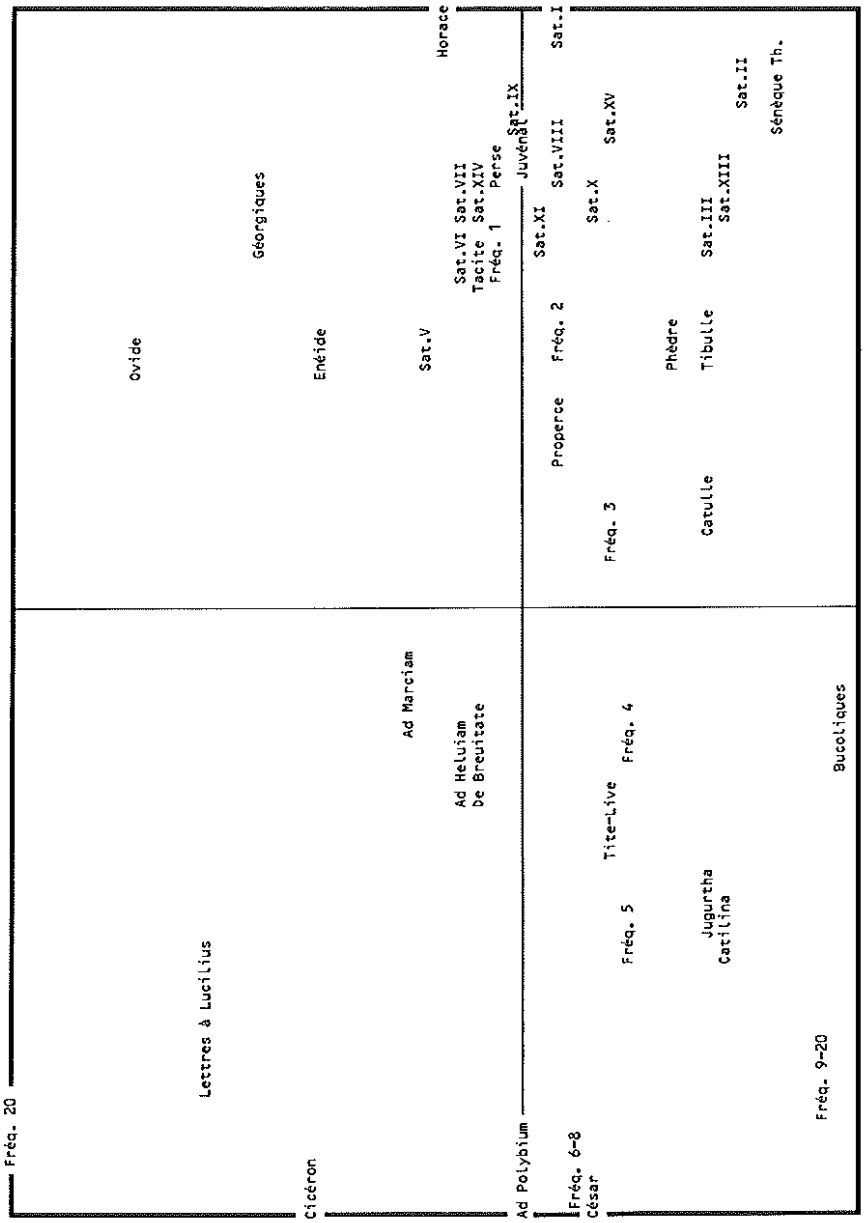


Figure 3
(Facteurs 1 et 2)

qui représente l'ensemble de l'oeuvre de Juvénal. Toutes ces observations montrent bien que la distribution des classes de fréquence peut caractériser un écrivain ou une forme littéraire.

D'autres rapprochements s'expliquent plus difficilement, par exemple celui de Phèdre et du *Corpus Tibullianum*, et justifieraient une étude plus approfondie.

On sait, par ailleurs, que sur les graphiques d'analyse factorielle la proximité des points-lignes et des points-colonnes est également significative : il apparaît très clairement que les poètes sont regroupés autour des fréquences 1 et 2, que les prosateurs sont plus proches des fréquences élevées.

De même les *Bucoliques*, que caractérise, ainsi que nous l'avons observé plus haut, l'abondance des mots de fréquence comprise entre 9 et 20, se trouvent placées en bas du graphique, comme cette classe de fréquence, alors que les fréquences supérieures à 20, déficitaires dans cette oeuvre, sont situées à l'opposé, sur la marge supérieure.

On pourrait multiplier ainsi les observations, mais il est nécessaire, pour aller un peu plus loin dans l'analyse, de faire intervenir des données numériques, grâce aux tableaux qui accompagnent le graphique.

3.3. *Utilisation des tableaux*

Les tableaux 3 et 4, un peu austère, reconnaissons-le, apportent des éléments essentiels pour interpréter le sens des éloignements et des proximités, la signification de la place occupée par tel ou tel point.

Chaque ligne du tableau 3 correspond à une oeuvre ou à un auteur, chaque ligne du tableau 4 à une classe de fréquence. Chaque ligne contient 18 nombres, répartis entre 6 groupes de 3 nombres.

Le premier groupe apporte des indications relatives à la signification de chaque point.

- La première colonne QLT définit la qualité de la représentation du point dans l'ensemble du tableau; plus QLT est proche de 1000, plus la représentation est satisfaisante. Il faut noter que QLT prend en compte les cinq facteurs

Tableau 3

	Oeuvres	QLT	POID	INR	1-F	COR	CTR	2-F	COR	CTR	3-F	COR	CTR	4-F	COR	CTR	5-F	COR	CTR
1	AD HELVIAM	999	43	10	-81	797	11	26	82	5	-30	116	12	0	0	0	6	4	13
2	DE BREVITATE	998	43	12	-79	622	11	29	83	6	-7	6	1	54	282	119	-6	5	22
3	BUCOLIQUES	999	44	30	-64	169	7	-141	801	141	27	28	9	-5	1	1	1	0	0
4	CATILINA	1000	44	48	-156	614	42	-85	184	52	83	172	89	31	24	40	16	6	108
5	JUGURTHA	1000	43	38	-137	603	32	-84	231	50	71	160	65	-10	4	5	7	2	22
6	AD LUCILIUM	999	43	73	-204	689	72	135	297	126	-7	1	1	27	12	31	2	0	1
7	AD MARCIAM	998	43	7	-44	351	3	49	425	17	-9	18	1	31	165	39	-14	39	95
8	AD POLYBIUM	1000	43	86	-263	978	119	1	0	0	-37	20	19	-6	1	2	-8	1	32
9	ENEIDE	1001	43	39	102	318	18	83	214	49	123	466	195	0	0	0	-8	3	37
10	GEORGIQUES	1001	43	50	158	594	43	112	297	87	59	84	45	29	20	34	16	6	111
11	SENEQUE TH.	1000	43	68	204	732	71	-108	208	82	58	59	43	4	0	1	-7	1	30
12	PHEDRE	999	43	20	103	617	18	-68	283	34	-39	93	20	-3	1	1	10	5	39
13	TIBULLE	999	44	23	109	608	20	-82	355	48	19	19	5	-17	17	14	-1	0	2
14	TITE-LIVE	999	43	14	-97	837	16	-40	145	12	11	10	2	8	6	3	3	1	4
15	CATULLE	999	43	13	35	118	2	-78	579	43	-29	83	11	46	200	88	-13	19	85
16	CESAR	999	44	92	-256	868	113	-29	11	6	-46	28	28	-83	92	291	-3	0	7
17	HORACE	1000	43	75	242	930	100	32	17	7	-27	12	10	-45	33	86	22	8	209
18	OVIDE	999	43	48	101	254	17	158	624	174	55	74	38	-41	45	75	-9	2	42
19	PROPERCE	997	44	10	76	735	10	-17	42	2	-31	132	13	-24	77	25	-8	11	38
20	PERSE	1000	43	65	186	640	59	10	2	1	-138	357	249	-4	1	1	-1	0	2
21	TACITE	999	43	26	138	869	32	21	21	3	36	61	17	-30	45	41	-8	3	31
22	JUVENAL	999	44	60	197	782	66	-4	0	0	-90	167	106	50	50	103	5	0	9
23	CICERON	1000	43	92	-257	874	114	89	103	55	-39	21	20	0	0	0	12	2	60
		22996.0	1000				1000			1000			1000			1000			1000

Tableau 4

	Fréquences	QLT	POID	INR	1-F	COR	CTR	2-F	COR	CTR	3-F	COR	CTR	4-F	COR	CTR	5-F	COR	CTR
1	FREQUENCE 1	1000	412	223	134	912	290	17	15	20	-32	57	135	18	16	122	-1	0	18
2	FREQUENCE 2	1000	161	59	95	686	58	-18	26	9	49	180	114	-34	93	189	14	15	307
3	FREQUENCE 3	995	83	22	19	38	1	-38	162	20	79	668	154	-33	123	91	-5	4	29
4	FREQUENCE 4	999	51	22	-69	310	10	-46	140	18	90	513	122	-1	0	0	-23	36	278
5	FREQUENCE 5	997	36	37	-160	695	37	-45	57	12	86	200	79	36	34	43	-20	11	148
6	FREQUENCES 6-8	999	68	155	-267	875	193	-27	9	8	54	35	58	79	76	407	18	4	217
7	FREQUENCES 9-20	999	104	226	-224	643	207	-127	209	274	-103	137	332	-28	10	82	0	0	0
8	FREQUENCES > 20	1000	85	256	-247	561	205	217	430	638	-15	2	6	-27	7	65	0	0	2
		22996.0		1000			1000			1000			1000			1000			1000

représentés sur le tableau. Le graphique que nous avons examiné plus haut ne faisait intervenir que les deux premiers facteurs.

- La seconde colonne POID exprime le poids numérique de la ligne ou de la colonne. Ce poids correspond à l'importance relative des effectifs de chaque catégorie. Sur le tableau 3, les différentes oeuvres ont sensiblement le même poids, puisque chacune d'elles est représentée par 1000 mots; que le tableau 4 les mots de fréquence 1 représentent, à eux seuls, plus de 40 % du poids total.
- La dernière colonne, INR, indique l'inertie du point, c'est-à-dire sa distance du centre de gravité du nuage. Plus un point est éloigné de la norme, plus il est différent, plus son inertie sera grande.

On constate la faible inertie des classes de fréquence et des oeuvres proches du centre du graphique (Fréquence 3, Fréquence 4, Properce, Catulle, *Consolation à Helvia, Consolation à Marcia, De Breuitate*), la forte inertie des oeuvres les plus "marginales" (Cicéron, César, Horace, Perse, Juvénal).

Les cinq autres groupes de 3 nombres concernent chacun un des facteurs déterminés par l'analyse, et décrivent donc un axe factoriel. Les deux axes qui apparaissent sur le graphique que nous avons présenté plus haut (figure 3) correspondent donc respectivement aux colonnes 4, 5, 6 et 7, 8, 9.

3.4. *Le premier facteur*

Pour l'axe horizontal, on trouve sous la colonne 1-F, la valeur de l'abscisse de chaque point : la *Consolation à Polybe*, avec une abscisse de -263 sera située à l'extrême gauche du tableau. En ce qui concerne les classes de fréquence (Tableau 4) on observe des valeurs décroissantes pour les six premières classes de fréquence, ce qui explique la disposition relevée sur le graphique.

La colonne suivante, COR, indique la contribution relative du point à l'axe. Ainsi la *Consolation à Polybe* apporte 97,8 % de sa contribution à l'axe horizontal. Si l'on y ajoute sa contribution aux autres facteurs (respectivement 0,20 , 1 et 1) on retrouvera le total 1 000 de la colonne QLT.

La dernière colonne, CTR, apporte les informations les plus précieuses. Il s'agit de la contribution absolue du point de l'axe

factoriel. Plus cette contribution sera élevée, plus le point, donc l'oeuvre ou la classe de fréquence considérée, aura contribué à l'élaboration du graphique.

Ainsi pour l'axe horizontal, et pour les oeuvres, les contributions les plus fortes correspondent à la *Consolation à Polybe*, à Cicéron et à César (partie négative), aux *Odes* d'Horace, aux oeuvres théâtrales de Sénèque, aux *Satires* de Perse et Juvénal (partie positive). Pour les classes de fréquence, l'apport essentiel vient des mots de fréquence 1 ainsi que des trois classes les plus élevées.

Il apparaît donc que l'axe horizontal est fondé, pour la plus grande part, sur l'opposition entre fréquence 1 et fréquences supérieures à 6, qui coïncide avec l'opposition entre des oeuvres tragiques, satiriques ou lyriques et des textes de prosateurs tels que la *Guerre des Gaules*, la *Consolation à Polybe* ou des extraits de Cicéron.

En fait cette opposition entre fréquences basses et fréquences élevées, entre poésie et prose, est directement liée aux variations de l'étendue du lexique, puisqu'il existe un rapport étroit entre le nombre des hapax et la richesse lexicale : un choix très large de vocables entraîne un fort contingent de mots de fréquence 1, et, en contre-partie un déficit des fréquences hautes; inversement plus le vocabulaire est limité, plus les répétitions sont nombreuses.

3.5. *Le deuxième facteur*

Les colonnes 7 à 9 des tableaux 3 et 4 apportent, pour l'axe vertical, des informations de même nature que celles qui étaient données par les colonnes 4 à 6 pour l'axe horizontal.

Pour les oeuvres, on constate que les plus fortes contributions sont le fait des *Bucoliques* pour la partie négative de l'axe, des *Métamorphoses* d'Ovide et des *Lettres à Lucilius* pour sa partie positive.

En ce qui concerne les fréquences, les fréquences comprises entre 9 et 20 pour la partie négative, les fréquences supérieures à 20 pour la partie positive, apportent une contribution décisive à la formation de l'axe.

Comment expliquer cette opposition ?

Si l'on revient au tableau 1 (trois dernières colonnes), on constate que les vocables les plus fréquemment utilisés sont en très petit nombre. Cependant, la multiplication de leurs occurrences leur donne une place importante dans le lexique. Un document publié par le L.A.S.L.A., *Dictionnaire fréquentiel et Index inverse de la langue latine* (7), permet de vérifier qu'il s'agit presque certainement de mots grammaticaux.

En effet, il est facile de calculer que pour qu'un mot apparaisse dans une tranche de 1 000 mots avec un effectif d'occurrences supérieur à 20, il faut, en théorie, que sa fréquence totale dans le corpus soit supérieure à 20/1000 du total, soit – ce dictionnaire portant sur un corpus de 794 662 mots – 15 893 occurrences. Trois vocables remplissent cette condition, ET, SVM et QVI. Pour les fréquences comprises entre 9 et 20, quatre nouveaux lemmes s'ajoutent à la liste, IN, QVE, NON et HIC.

Bien entendu ce classement est tout à fait théorique, et il est évidemment impossible de déterminer, à partir de l'ensemble du corpus, le classement des vocables dans chacune des oeuvres qui le composent. Il reste qu'il faut attendre la fréquence 3 709 qui correspond à un effectif théorique de moins de cinq occurrences dans une tranche de 1 000 mots, pour rencontrer le premier mot à sens plein, le verbe POSSVM.

Il apparaît donc que l'utilisation des vocables de fréquence élevée n'est pas directement liée au contenu même du texte. On comprend mieux dès lors qu'un texte poétique, les *Métamorphoses* d'Ovide, ait en commun avec un texte de prose, les *Lettres à Lucilius*, un nombre particulièrement élevé d'occurrences correspondant aux fréquences les plus hautes.

Il s'agirait là d'une opposition de caractère stylistique. Certains auteurs n'hésiteraient pas à utiliser très fréquemment les mêmes formes de coordination et de subordination; d'autres, comme Virgile dans les *Bucoliques*, s'efforceraient d'atteindre à plus de variété dans l'expression, et éviteraient de répéter trop souvent tel ou tel mot grammatical.

Il nous paraît difficile d'aller plus loin sans examiner dans le détail les oeuvres concernées; peut-être y a-t-il, dans l'opposition que nous avons relevée, un élément de différenciation

stylistique intéressant, peut-être aussi l'effectif important des occurrences mises en jeu contribue-t-il à grossir excessivement des écarts en eux-mêmes peu significatifs : une variation d'une unité dans le nombre des vocables se traduit ici par une variation de plusieurs dizaines d'unités dans le nombre des occurrences.

Après avoir examiné les deux premiers facteurs, qui étaient à l'origine du graphique que nous avons présenté (figure 3), nous pouvons prolonger notre analyse en faisant intervenir les facteurs suivants; ainsi pourrait-on présenter de nouveaux graphiques réunissant les facteurs 1 et 3, puis 2 et 3, 1 et 4, etc.

3.6. *Le troisième facteur*

En fait, et pour des raisons que nous justifierons plus loin, nous nous sommes arrêté au troisième facteur, qui apparaît, associé au premier facteur, sur un nouveau graphique (figure 4). Les colonnes 10 à 12 des tableaux 3 et 4 apportent des indications numériques qui permettent de mesurer la contribution des oeuvres et des classes de fréquence à la construction de l'axe vertical, représentant ce troisième facteur.

Il est clair, sur le tableau 4 comme sur le graphique, que les fréquences moyennes s'opposent aux fréquences extrêmes, cette fois regroupées. Parmi les oeuvres (tableau 3 et graphique), on constate que les *Satires* de Perse et de Juvénal apportent une contribution majeure à la constitution de la partie négative de l'axe, tandis que pour la partie positive l'apport essentiel vient de l'*Enéide*, et, à un degré moindre, de *Catilina* et *Jugurtha*.

Ces oppositions semblent définir, à côté de la distinction traditionnelle entre les oeuvres à vocabulaire très étendu et les oeuvres à vocabulaire plus limité, une catégorie nouvelle, celle des oeuvres qui, éloignées des extrêmes, évitent tout aussi bien la multiplication des hapax que celle des vocables à fréquence très élevée, et se tiennent dans une honnête médiocrité.

Devons-nous faire intervenir les facteurs suivants ? En ce qui concerne les classes de fréquence, le tableau 4 montre bien que celles qui ont le plus grand poids, c'est à dire les catégories extrêmes, s'étaient largement exprimées à travers les trois premiers facteurs. Pousser plus loin l'analyse ferait

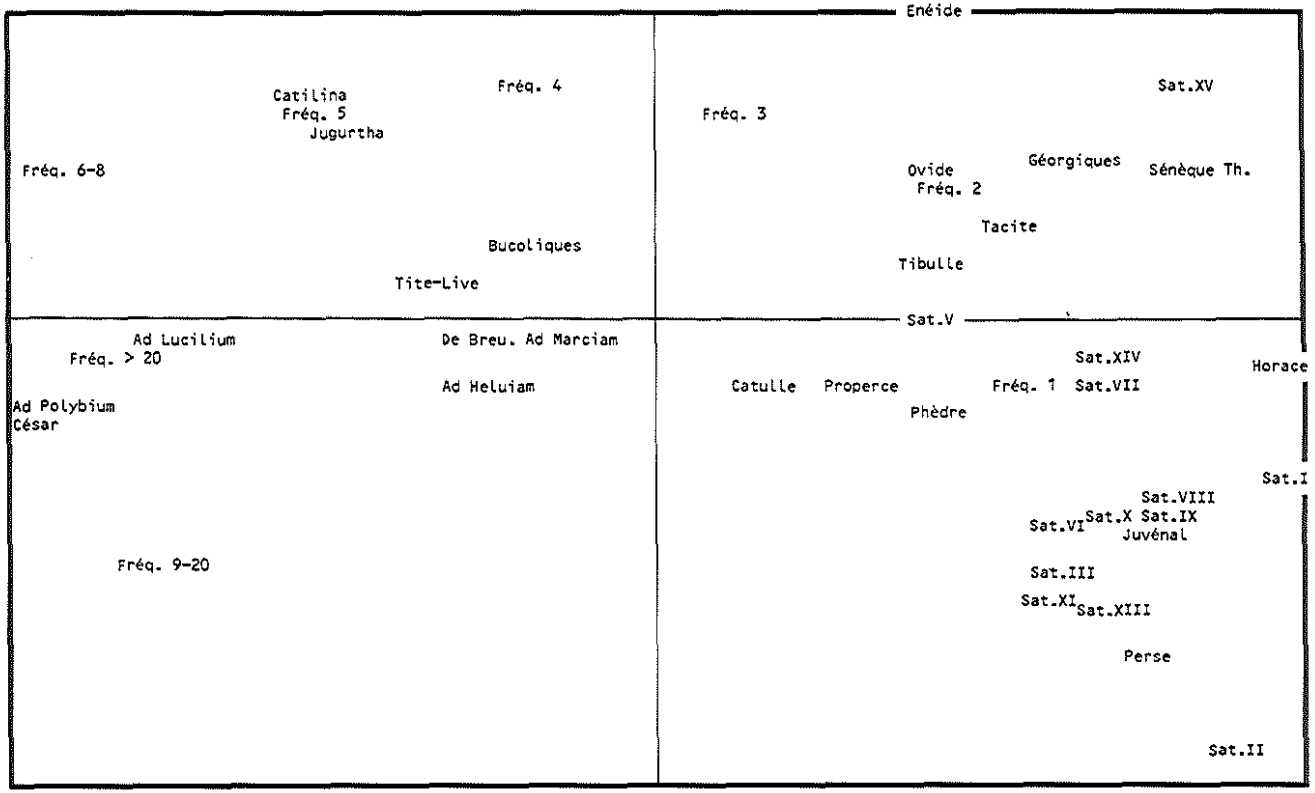


Figure 4
(Facteurs 1 et 3)

essentiellement intervenir les catégories moyennes, dont l'importance globale est faible, et qui, nous l'avons vu, ne présentent pas, d'une oeuvre à l'autre, des oppositions très significatives.

Quant aux auteurs, leur contribution relative cumulée aux trois premiers facteurs atteint, en moyenne, 95 % , et dépasse, dans tous les cas, 72 %.

Il semble donc qu'aller au-delà du troisième facteur n'apporterait pas des informations de grand intérêt.

Nous nous étions fixé un but très limité, comparer entre eux quelques auteurs latins, et quelques oeuvres, du point de vue de l'étendue du lexique et de la distribution des classes de fréquence.

Il conviendrait maintenant de revenir sur chacun de ces auteurs, et de proposer des explications aux écarts et aux ressemblances que nous avons relevés. Dans cette tentative d'élucidation, d'autres éléments stylistiques devraient nécessairement intervenir.

Si l'on se proposait, par exemple, de justifier la place particulière qu'occupe Tacite, exilé au milieu des poètes, il faudrait entrer dans le détail des oeuvres, faire appel à de nouveaux critères, examiner non plus les classes de fréquence, mais les vocables eux-mêmes. Ainsi, sur un point bien particulier, une étude très récente du L.A.S.L.A. (8) a montré que Tacite, dans le *Dialogue des Orateurs* et dans la *Germanie* fait appel à autant de types de subordination que les poètes.

Une telle recherche dépasserait très largement les limites d'un article; nous avons seulement voulu en montrer la possibilité et l'intérêt.

Déjà nous avons bien conscience d'avoir posé plus de questions que nous n'avons apporté de réponses. C'est que le but d'une analyse factorielle n'est pas de démontrer, mais de suggérer des analogies, des oppositions.

La présente étude nous a amené à compliquer un peu le schéma qui oppose traditionnellement les écrivains qui utilisent un vocabulaire très étendu, et ceux qui font appel à un nombre restreint de vocables.

Entre ces deux extrêmes se manifeste un autre type d'oeuvres, celles qui, évitant tout excès, donnent une place de choix aux vocables de fréquence moyenne.

Par ailleurs, il semble que l'on doive introduire une distinction entre les fréquences très élevées, et les fréquences un peu moins élevées : certains écrivains s'efforceraient de ne pas franchir un seuil bien précis de fréquence.

Pour vérifier ces hypothèses, il conviendrait de procéder à de nouvelles analyses, et, surtout, de revenir aux textes en étudiant plus précisément la structure lexicale de chacun d'eux.

Les travaux du L.A.S.L.A. mettent à la disposition des chercheurs une masse inépuisable de données. Pour en tirer le meilleur parti des outils statistiques puissants sont nécessaires. Parmi eux, l'analyse des correspondances paraît ouvrir d'intéressantes perspectives.

Notes

(1) On sait, par exemple, que l'opposition classique entre la richesse du vocabulaire de Corneille, et la pauvreté de celui de Racine n'est nullement vérifiée : le vocabulaire du *Cid* est moins étendu que celui de *Phèdre*. Sur ce point, voir Muller (Ch.) *Initiation à la statistique linguistique*, Paris, 1968, p. 157.

(2) Voir, en particulier, Muller (Ch.), *Principes et méthodes de statistique lexicale*, Paris, 1977. Les chapitres *La distribution des fréquences* (pp. 90-94), et *Modèle binomial d'une distribution théorique* (pp. 101-109) contiennent un exposé très clair de l'utilisation de la formule binomiale.

(3) Dans l'ouvrage de Charles Muller cité dans la note précédente, voir plus spécialement le chapitre *La richesse lexicale, calcul de l'étendue théorique* (pp. 127-129).

(4) Voir notre article *Evaluation de l'étendue du lexique, essai de simulation*, dans *Travaux de Linguistique et de Littératures romanes*, Strasbourg, 1983.

(5) Dans une très abondante bibliographie, nous retiendrons deux livres d'initiation :
- Cibois (Ph.), *L'analyse factorielle*, Paris, P.U.F., coll. "Que sais-je ?", 1983.
- Novi (M.), *Méthode d'analyse des "données", Technique et critique des composantes principales* dans les *Cahiers des Utilisateurs de Machines à des Fins d'Information et de Documentation* (CUMFID), n° 11, juin 1977.

Nous signalons également deux ouvrages fondamentaux :
- Benzecri (J.-P.) et al., *L'analyse des données*, 2 vol., Paris, 1973.
- Benzecri (J.-P. et F.), *La pratique de l'analyse des données*, 3 vol., Paris, 1980-1981.

(6) Programme d'analyse factorielle dû à N. Tabet, collaborateur de J.-P. Benzecri, et implanté au C.N.U.S.C. de Montpellier.

(7) Delatte (L.),Evrard (Et.), Govaerts (S.), Denooz (J.), *Dictionnaire fréquentiel et Index inverse de la langue latine*, Liège, L.A.S.L.A., 1981.

(8) Delatte (L.) Govaerts (S.), Denooz (J.), *Quelques observations statistiques relatives à la subordination en latin*, dans *Revue* n° 1 à 4, 1982. Voir, en particulier, p. 140.