LALIES 18 (1998)

LES MÉTHODES QUANTITATIVES EN LANGUES ANCIENNES

Étienne ÉVRARD & Sylvie MELLET

RÉSUMÉ

§1. L'histoire de la statistique linguistique et littéraire est retracée à grands traits. Quelques notions fondamentales sont ensuite définies et illustrées d'exemples (population et individu, fréquence absolue et fréquence relative, traitement quantitatif des données).
§2. La probabilité que les écarts entre effectifs observés de phonèmes, soit entre deux langues, soit entre plusieurs textes d'un même domaine linguistique, soient aléatoires ou significatifs peut être mesurée par plusieurs tests qui sont expliqués à partir d'exemples concrets.
§3. S'agissant de lexique, l'unité de compte est soit la forme graphique, soit le lemme. L'analyse factorielle permet de mesurer les distances et les proximités entre deux ou plusieurs œuvres, à partir des effectifs des mots qui composent leur vocabulaire. Plus classique, le test de l'écart réduit mesure la probabilité que l'effectif d'un mot dans une œuvre soit aléatoire ou significatif.
§4. La diversité du vocabulaire tient au taux de répétition des divers vocables (mots employés 1, 2, x fois). Un test fondé sur l'entropie permet de caractériser cette diversité. A partir de là, on peut développer un test qui précise par des paramètres quantitatifs la manière dont le vocabulaire d'une œuvre s'enrichit au fur et à mesure que cette dernière s'allonge.
§5. Les effectifs observés relativement à des phénomènes syntaxiques se prêtent à des tests qui font voir avec netteté des différences que la simple observation n'aurait pas laissé soupçonner (par exemple deux conjonctions considérées comme synonymes et les modes qu'elles régissent). Lorsque les données forment des tableaux de grandes dimensions (par exemple le croisement des formes modo-temporelles avec les personnes verbales), l'analyse factorielle est tout indiquée.
§6. La métrique de l'hexamètre dactylique permet de compter les effectifs des 16 formes possibles dans différentes œuvres. Divers modes de calcul, y compris l'analyse factorielle, fondés sur ces effectifs permettent de déterminer dans quelle mesure ils sont aléatoires ou significatifs, fournissant ainsi une indication précieuse à l'analyse textuelle.

ZUSAMMENFASSUNG

§1. Die Geschichte der linguistischen und literarischen Statistik wird kurz dargestellt. Einige Grundbegriffe (Testmenge und Individuum, absolute und relative Häufigkeit, quantitative Datenverarbeitung) werden erklärt und mit Beispielen erläutert.
§2. Die Wahrscheinlichkeit, daß die Abweichungen zwischen den Häufigkeiten der Phoneme sei es zweier verschiedener Sprachen, sei es verschiedener Texte ein und derselben Sprache zufällig oder signifikant sind, kann mit Hilfe verschiedener Tests berechnet werden. Diese Tests werden mit anschaulichen Beispielen erklärt.
§3. Wenn es sich um Wortschatz handelt, ist die Rechnungseinheit entweder die graphische Gestalt oder das Wort (lexikalische Einheit). Durch die Faktorenanalyse ist es möglich, den jeweiligen Abstand bzw. die Nähe zwischen zwei oder mehreren Werken zu messen, und zwar gemäß der Häufigkeit der jeweils verwendeten Wörter. In klassischer Weise mißt dann der Test der reduzierten Abweichung die Wahrscheinlichkeit, ob die absolute Häufigkeit eines

Wortes in einem Texte zufällig bzw. signifikant ist.
§4. Die Verschiedenartigkeit des Wortschatzes hängt von der Wiederholungshäufigkeit der Wörter ab (einmal, zweimal usw. verwendete Wörter). Mit einem auf der Entropie basierenden Test wird diese Verschiedenartigkeit bestimmt. Darüberhinaus untersucht ein weiterer Test mit quantitativen Parametern die Art und Weise, wie der Wortschatz eines Werkes mit dem Fortgang desselben reicher wird.
§5. Auf die Häufigkeiten der syntaktischen Konstruktionen werden Tests angewendet, die einige mit bloßem Auge nicht wahrnehmbare Unterschiede deutlich hervortreten lassen (z. B. zwei angeblich gleichbedeutende Konjunktionen und die von ihnen jeweils regierten Modi). Wenn die Daten allzu umfangreiche Tabellen bilden (z. B. wenn die Zeitformen mit den grammatikalischen Personen zusammengenommen werden), empfiehlt es sich, die Faktorenanalyse zu verwenden.
§6. Die Verslehre des lateinischen daktylischen Hexameters unterscheidet die sechzehn (bzw. zweiunddreißig) Formen dieses Verses. Auf die Häufigkeiten dieser Formen werden verschiedene Tests angewendet (einschließlich der Faktorenanalyse), die bestimmen, in welchem Maße diese Häufigkeiten zufällig bzw. signifikant sind, wobei sich nützliche Hinweise für die Textanalyse ergeben.

1. HISTORIQUE (Sylvie Mellet)

1. 1. Le développement des applications de la statistique en linguistique est relativement récent: les premiers dictionnaires de fréquence, établis à la main à partir de textes littéraires, datent de la première moitié de ce siècle.

Pour situer plus précisément les choses, on peut signaler quelques étapes importantes et citer quelques grands noms qui ont marqué l'histoire de la discipline:

– à la fin du XIX^{e} siècle, J.B. Estoup étudie les distributions lexicales dans un but très précis, celui d'améliorer les méthodes de transcription sténographique, et met ainsi en évidence des lois purement empiriques; son livre est réédité jusqu'en 1916;

– vingt ans plus tard, en 1935, G.K. Zipf reprend l'étude de ces lois de distribution dans un tout autre cadre, celui de la "psycho-biologie du langage".

Ce sont là deux précurseurs dont les travaux ont eu un certain retentissement. Parallèlement sont parues des études, peu nombreuses, portant sur le décompte et la distribution des phonèmes dans diverses langues (cf. en particulier l'article de Förstemann 1852) ou sur des questions de métrique (cf. les travaux de Drobisch 1866, 1868 et 1871 et Hultgren 1872); mais elles restent sans suite dans l'immédiat et ne suscitent aucun débat; ce n'est qu'en 1952, soit un siècle plus tard, que Pfister ou Herdan reprendront les analyses de Förstemann.

Viennent ensuite les travaux de Yule, statisticien de formation, qui auront une importance indéniable; mais leur impact se fera sentir, lui aussi, à retardement dans la mesure où ils furent publiés pendant la deuxième guerre mondiale.

Ce n'est donc qu'après 1950 que la statistique linguistique va connaître son véritable essor, avec l'aide de promoteurs tels que M. Cohen, P. Guiraud et Ch. Muller. L'absence d'études quantitatives dans les études textuelles est alors perçue comme un retard qu'il s'agit de combler: "Mais dans quelle mesure cette voie ouverte a-t-elle été suivie et élargie? On est honteux en y pensant. Un siècle s'est écoulé et la grammaire est toujours enseignée comme un cadre normatif en soi, sans souci des proportions

d'emploi"[1] ; et, bien sûr, rappelons la célèbre formule de P. Guiraud : "La linguistique est la science statistique type ; les statisticiens le savent bien, la plupart des linguistes l'ignorent encore". P. Guiraud, puis Ch. Muller vont donc essayer de familiariser les littéraires avec la statistique, d'une part en reprenant à leur compte les préoccupations et les thèmes de recherche traditionnels des études littéraires et philologiques, tels que la comparaison de la richesse du vocabulaire chez deux auteurs différents, la mesure de l'évolution du vocabulaire d'un même auteur au fil du temps, les analyses permettant d'authentifier l'attribution d'une œuvre à un auteur, etc. ; d'autre part en publiant des manuels de statistique à l'usage de chercheurs peu familiers avec les chiffres, l'un des plus connus étant l'*Initiation aux méthodes de la statistique linguistique* de Ch. Muller.

Le rayonnement de ces grands noms a fini par imposer la reconnaissance de la discipline qu'ils défendaient et illustraient, surtout dans le domaine de la lexicologie. Depuis, le développement des recherches s'est fait conjointement dans deux directions complémentaires, soutenu aussi par les progrès et la vulgarisation de l'informatique :

– amélioration des outils statistiques dans le sens d'une meilleure prise en compte de la multiplicité et du caractère souvent qualitatif des variables qui entrent en jeu dans les analyses textuelles : développement des analyses dites multidimensionnelles (cf. J.-P. Benzecri) ;

– développement des banques de données textuelles qui permettent d'augmenter sensiblement la taille des corpus étudiés tout en soulageant le chercheur du travail fastidieux que représentaient les dépouillements manuels et tout en réduisant aussi les risques d'erreur dans les décomptes. Pour les langues anciennes, on signalera tout particulièrement le travail du L.A.S.L.A., de l'Université de Liège, qui dès le début des années 60 a entrepris de constituer une banque de données informatisée des textes latins et grecs, banque textuelle qui a la particularité tout à fait remarquable d'être lemmatisée. Nous reviendrons ultérieurement sur ce point de méthode. Le L.A.S.L.A. a été créé en 1961 et son intitulé (Laboratoire d'Analyse Statistique des Langues Anciennes) affiche clairement ses objectifs.

Telle est donc, tracée à grands traits, l'histoire de la statistique linguistique ; mais, en réalité, ce panorama trop rapide néglige tout ce qui n'est pas écrit dans l'état civil de la discipline, c'est-à-dire tout ce qui relève de son histoire embryonnaire. Or, comme chacun sait, cette période prénatale représente un temps de maturation indispensable, celui au terme duquel l'enfant pourra être accueilli.

1.2. Voilà longtemps en effet que, dans leurs études sur les textes, les philologues pratiquent les dénombrements, voire le calcul de pourcentages ou encore l'appréciation des écarts entre la langue d'un auteur et une norme plus ou moins explicitement définie : c'est ce qu'on pourrait appeler les affleurements de la préstatistique.

En effet, un lecteur assidu des textes latins par exemple finit par tirer de sa pratique le sentiment plus ou moins conscient qu'il existe des règles générales, des tendances fortes comme on dit maintenant, dont certaines sont enregistrées par les grammaires pour définir la norme, dont d'autres constituent plus implicitement une toile de fond sur laquelle seront portées les appréciations concernant le niveau de langue ou le style de tel ou tel auteur. La démarche reste plus ou moins inconsciente, elle n'en est pas moins déjà d'ordre statistique : les textes lus et relus sont considérés comme **des**

1. Cohen 1967 : 4

échantillons représentatifs d'un état de langue, la langue classique par exemple ; la porte est alors ouverte à des raisonnements par induction non contrôlés.

Ainsi, j'ai montré dans ma thèse (Mellet : 1988) et dans un article paru en 1990, que certaines de ces inférences méritaient d'être remises en question ; je rappellerai l'exemple de la distribution du parfait et de l'imparfait du verbe *esse* : de nombreux grammairiens remarquent que, dans l'état de langue ancien représenté par le théâtre de Plaute, la forme *eram* est parfois utilisée là où, en langue classique, on attendrait plutôt *fui*. Wheeler parle à ce propos d'"imparfait aoristique" et Szantyr se demande si, au stade préclassique, le verbe *esse* était vraiment sensible à l'opposition aspectuelle ; et d'en tirer des conclusions sur l'évolution diachronique du latin dans ce domaine. Or ces conclusions reposent typiquement sur une induction abusive, liée probablement au poids excessif de Cicéron dans notre connaissance du latin classique et dans la norme implicite qu'on en dégage.

En effet, s'il est vrai que la proportion des formes d'imparfait et de parfait n'est pas la même pour le verbe *esse* que pour l'ensemble des autres lexèmes verbaux (l'imparfait *eram* étant proportionnellement plus fréquent que le parfait *fui*), la distribution de cette caractéristique varie selon les auteurs et, de fait, elle est plus accentuée chez Plaute que chez Cicéron. Le raisonnement intuitif a donc consisté à prendre l'œuvre de Cicéron comme échantillon représentatif de la langue classique, à en généraliser les caractéristiques distributionnelles, à comparer Plaute à cette norme - qualifiée de façon à orienter l'analyse vers la dimension diachronique - et à conclure à une évolution dans le temps du phénomène étudié. Or une exigence plus aiguë sur la notion d'échantillon représentatif et des dépouillements complémentaires pour contrôler l'analyse font apparaître que la prédominance de *eram* sur *fui* est encore plus marquée chez Tite-Live que chez Plaute (facteur chronologique remis en cause), et encore davantage chez César que chez Tite-Live (norme supposée de la langue classique battue en brèche). La leçon provisoire de ce rappel est que, dans ce cas, des dépouillements plus nombreux, accompagnés de tests statistiques fiables qui permettaient de comparer des textes de tailles différentes, ont fait apparaître l'existence d'une autre variable (le genre littéraire, opposant les historiens aux autres) susceptible d'influencer le phénomène étudié et de modifier la norme implicitement établie.

On verra plus bas (fin du § 2 et note 10 en particulier) un autre exemple concernant les conclusions de N. Herescu sur les structures phoniques du vers latin.

Loin de nous l'idée de critiquer tous les travaux anciens sous prétexte que leur maîtrise de la statistique ne serait pas suffisante : la grande majorité d'entre eux manifestent au contraire une connaissance admirable de la langue latine et ont produit des outils d'analyse remarquables sous la forme d'index et de concordances, objets typiques de la statistique documentaire.

Mais il s'agit de montrer que les raisonnements utilisés s'appuient souvent sur des notions statistiques intuitives qu'il vaudrait mieux approfondir, car l'à peu près et l'implicite dans ce domaine sont dangereux : d'une part, on est ainsi tenté d'apporter une réponse à un problème dont les termes n'ont pas été clairement posés ; d'autre part, les psychologues ont montré que l'intuition était particulièrement mauvaise conseillère pour nous aider à évaluer si la distribution d'un phénomène est ou non aléatoire (cf. L. Lopes & G. Odin 1987 et W.A. Wagenaar 1972).

Il s'agira donc ici de préciser quelques notions fondamentales, de signaler quelques pièges grossiers à éviter et de présenter quelques outils de recherche parmi les plus accessibles.

Les efforts pour maîtriser ce domaine ne sont pas dépourvus d'intérêt car les textes littéraires se prêtent tout particulièrement à l'analyse quantitative: ils sont en effet composés d'éléments discrets (discontinus) susceptibles d'être comptés, ce qu'en statistique démographique – et, par extension, en statistique générale – on appelle des **individus**. Les individus composant un texte peuvent être définis à différents niveaux: les paragraphes, les phrases, les vers, les mots, les syllabes, les phonèmes; et chaque individu peut être comparé aux autres du même type sur la base d'un critère qualitatif stable, par exemple pour la phrase ou le vers le nombre de mots qui les constituent, pour le mot la catégorie grammaticale à laquelle il appartient, pour le phonème la qualité de voyelle ou consonne, etc.

Constamment, on l'a vu, les grammairiens et les stylisticiens se livrent à des dénombrements qui leur permettent d'établir la **fréquence** du caractère étudié: **fréquence absolue** qui correspond à l'effectif des individus présentant ce caractère dans le texte et **fréquence relative** qui permet de rapporter cet effectif à l'effectif global des individus sous examen.

Par exemple, si on s'intéresse aux adjectifs dans *Agamemnon* de Sénèque, on note que sur les 5 537 mots composant cette tragédie, 788 sont des adjectifs et que parmi ces derniers 23 sont au comparatif et 18 au superlatif. On dira donc:

que la fréquence absolue des adjectifs dans ce texte est de 788;

que celle des comparatifs est de 23;

que celle des superlatifs est de 18;

mais aussi que la fréquence relative des adjectifs dans l'ensemble du texte est de 788/5 537 = 0,14;

celle des comparatifs de 23/5 537 = 0,004;

celle des superlatifs de 18/5 537 = 0,003;

et que la fréquence relative des comparatifs parmi les formes d'adjectifs est de 23/788 = 0,03.

Ce type de calculs simples est constamment utilisé par les philologues et par les stylisticiens, et à juste titre: le pourcentage est le premier outil de la statistique. Mais aussitôt après se pose la question cruciale: comment comparer des fréquences entre elles? A-t-on le droit d'aligner des pourcentages, de les hiérarchiser et d'en tirer des conclusions sans autre précaution? Le statisticien répond 'non'; ne serait-ce que dans la mesure où ces comparaisons brutes ne prennent pas en compte les différences de taille entre les divers échantillons comparés. C'est pourquoi des calculs complémentaires sont nécessaires, qui vont être présentés ci-après à travers divers types d'exemples.

2. STATISTIQUE PHONÉMIQUE (Étienne Évrard)

Le fil conducteur de ces exposés n'est pas la série des techniques statistiques utilisées, mais bien les objets soumis à examen, c'est-à-dire les éléments de la langue et des textes, dans la mesure où des méthodes quantitatives peuvent leur être appliquées.

Nous commencerons donc par les sons, qui constituent les éléments les plus simples de la langue. Quels sont, de ce point de vue, les questions et les problèmes qui pourraient bénéficier d'un traitement quantitatif?

On sait, tout d'abord, que règnent à propos de certaines langues des lieux communs relatifs à l'impression qu'elles font à l'audition: on dit volontiers que l'italien abonde en *i*, tandis que certaines langues germaniques admettent facilement des accumulations de consonnes qui semblent difficilement prononçables pour des gosiers

habitués aux langues romanes (par exemple *angstschreeuw* en néerlandais); on est sensible aussi à la fréquence des aspirées dans certaines langues sémitiques, qui en possèdent d'ailleurs tout un éventail. Il serait aisé d'allonger cette énumération.

Indépendamment de ces caractères généraux des langues, on remarque dans certains textes des particularités sonores qui les singularisent, soit que soient mis à profit les caractères virtuellement expressifs des sons, soit que leur répartition plus ou moins régulière crée des figures reconnaissables à l'audition.

Certaines de ces figures sont d'un emploi suffisamment remarquable pour qu'elles soient devenues des *patterns*; tels sont, entre autres, les allitérations ou les effets liés à la répétition de sons aux mêmes places dans le vers ou dans la phrase.

Tous les phénomènes que j'ai ainsi énumérés sont implicitement statistiques: pour les décrire avec précision, on doit nécessairement passer par le quantitatif. Qu'une langue paraisse remarquable par l'abondance d'un son, ou par l'abondance de certains groupes de sons, cela ne peut se décrire précisément que par des dénombrements et par des comparaisons, lesquelles ne sont convaincantes que si elles se fondent sur des techniques éprouvées. De même, si l'on croit percevoir dans un texte une utilisation expressive de certains sons, on n'échappe au soupçon de jugement subjectif que si l'on chiffre les occurrences de ces utilisations et si, par des comparaisons quantitatives, on en fait apparaître le caractère plus ou moins exceptionnel. Et ainsi de suite.

Le profit que les études quantitatives peut apporter à l'étude de l'aspect sonore des langues a été perçu dès le milieu du XIX^{e} siècle, par exemple par E. Förstemann (1852: 161-179), dont un article a servi de base à des recherches de R. Pfister (1952 (1956^{2}): 161-179) et de G. Herdan (1956: 81-84).

Förstemann note à juste titre que les recherches quantitatives sont de nature soit à préciser certains phénomènes déjà bien connus mais décrits de manière tout approximative, soit de mettre au grand jour des faits et des relations qui, jusque-là, étaient à peine soupçonnés ou même interprétés à contresens.

Mais il faut avouer qu'à côté de qualités réelles, son travail souffre de graves imperfections, dues certes pour une bonne part à l'état embryonnaire des recherches de ce genre à son époque. Tout d'abord, il a conscience que son étude des sons repose entièrement sur l'examen des graphies mais ne paraît pas s'en inquiéter; cela entraîne des anomalies telles que l'absence de l'aspiration marquée par l'esprit rude en grec, alors que le *h* latin est relevé. D'autre part, les chiffres publiés par Förstemann donnent un total de 100 pour chacune des trois langues considérées (latin, grec, gotique): cela donne à penser qu'il s'agit de pourcentages (arrondis puisqu'aucune décimale n'apparaît), sans que l'on soit mis au courant ni du corpus sur lequel ont été opérés les dénombrements, ni de son importance numérique; on doit donc tenir compte du fait que ce corpus est peut-être insuffisant. Par ailleurs, il semble évident que le but de Förstemann est de découvrir des lois aussi générales que possible et qu'il n'accorde guère d'importance aux diversités que peuvent recouvrir des moyennes; c'est là une tendance assez générale au XIX^{e} siècle, influencé par les ambitions du scientisme. Enfin, il faut dire qu'à l'époque de Förstemann, la statistique ne disposait guère des ressources actuelles: de nombreuses méthodes devenues tout à fait courantes n'avaient même pas encore été découvertes; tel est le cas de la fonction du χ^2, découverte et explorée à la fin du XIX^{e} siècle et au début du XX^{e} par K. Pearson (1900), sur laquelle se fonde un test qui sera très utilisé et est, de fait, présent dans l'analyse des correspondances, comme l'a indiqué à diverses reprises J.-P. Benzecri.

Les réflexions précédentes concernant Förstemann m'amènent à préciser quelques règles dont l'observation est indispensable au bon déroulement des recherches quantitatives. Comme dans toute recherche, il convient de définir l'objet qu'on se propose d'étudier; dans le cas d'études quantitatives, cela suppose la définition d'une variable dont toutes les occurrences peuvent être considérées comme équivalentes pour la recherche en vue: on ne peut additionner, soustraire ou multiplier que des quantités formées d'unités interchangeables. En fait, les éléments des corpus sur lesquels on travaille ont chacun leurs particularités, mais on les ignore temporairement pour ne prendre en considération qu'un ou quelques traits communs qu'on sélectionne (je préfère parler ici de sélection plutôt que d'abstraction: le caractère non sélecté dans une recherche peut l'être dans une autre et il n'est en aucune manière perdu ou annihilé dans une perspective hiérarchisée, si bien qu'on peut à tout moment le réactiver et le prendre en compte). La variable une fois définie, on énumère celles de ses modalités qu'on désire prendre en considération. Prenons un exemple concernant les sons, puisque la présente section est consacrée à la statistique phonémique. L'ode I, 11 d'Horace (à Leuconoé) peut être considérée comme une suite de phonèmes, très divers sans doute les uns des autres mais dont, pour l'occasion, on ne retient que le fait qu'ils sont des phonèmes. On peut alors les dénombrer et dire que l'ode est formée de 291 phonèmes, renseignement qui ne va pas loin, mais qui pourrait prêter à des comparaisons intéressantes avec d'autres odes, concernant la longueur des poèmes lyriques et ses variations, par exemple selon les auteurs. On notera ensuite que les phonèmes de notre texte se réalisent soit comme consonnes, soit comme voyelles. Si l'on décide de tenir compte de cette différence sans prendre en considération les distinctions de timbre, d'articulation, etc. qui individualisent les diverses voyelles et les diverses consonnes, on dira que notre texte contient 125 voyelles et 166 consonnes. D'autres dénombrements de ce genre montreraient dans quelle mesure la supériorité numérique des consonnes se retrouve ailleurs et dans quelles proportions.

Dans les domaines qu'aborderont les paragraphes suivants, on se trouvera en présence de la même nécessité de définir des variables, unités qui puissent être tenues pour équivalentes, qu'il s'agisse de lexique, de syntaxe ou de tout autre matière.

Une fois définie la variable et ses modalités, il faut encore préciser la nature, les limites et la taille du corpus sur lequel porte l'investigation. S'il s'agit d'un texte littéraire, par exemple, on ne peut se dispenser d'indiquer l'édition suivie et, éventuellement, les points sur lesquels on s'en sépare. Il doit, en outre, être bien entendu que, dans ce corpus, les dénombrements seront exhaustifs et non sélectifs.

J'ajouterai que, pour ma part, je préfère choisir des corpus de dimensions restreintes, dans lesquels la probabilité d'homogénéité est généralement plus grande et en tout cas plus facilement vérifiable. On pourra toujours, ensuite, regrouper plusieurs corpus qui se seront révélés homogènes, alors que l'opération inverse serait plus difficile, voire impossible.

Les données ainsi collectées seront ensuite soumises à des tests statistiques, mais ces tests ne donnent aucune certitude, ils ne révèlent rien; ils se bornent à donner des idées: sur ce qui fait difficulté, sur ce qui demande une explication autre que le hasard, etc. Ils doivent donc être interprétés et, dans cette interprétation, le retour au contexte est indispensable: en particulier, les variables et modalités non sélectées pour la recherche statistique, la nature du corpus, non seulement ses limites matérielles mais aussi ses autres caractéristiques (éventuellement qualitatives), auront un rôle à jouer, l'interprétation conduisant souvent à repérer des convergences éclairantes.

Je vous proposerai un premier test. Nous avons vu que l'ode I, 11 d'Horace contient 125 voyelles et 166 consonnes. En relation avec ce que j'ai dit quant à la coloration diverse des sonorités des langues, il m'a paru intéressant de faire le même décompte sur le poème de Goethe intitulé *Lust und Qual*[2], dont je n'ai retenu que les deux premières strophes, de manière à atteindre un total de phonèmes à peu près égal à celui de l'ode I, 11.

On dispose les données dans un tableau à double entrée, une colonne pour Horace et une pour Goethe; une ligne pour les effectifs des voyelles et une pour ceux des consonnes; dans les marges, les totaux.

	Horace	Goethe	Total
Voyelles	125	117	242
Consonnes	166	209	375
Total	291	326	617

Si les deux corpus appartiennent à une population homogène pour ce qui concerne la proportion consonnes voyelles, on s'attend à ce que les proportions soient les mêmes dans l'un et dans l'autre d'entre eux (c'est l'hypothèse nulle). Cela permet de calculer des effectifs théoriques: le corpus *Horace* représente un peu plus de 47 % du total (291 / 617); il devrait donc contenir 47 % des 242 voyelles c'est-à-dire 114. Le même procédé pourrait s'appliquer aux trois autres données, mais, pour un tableau de 2 x 2, cela n'est pas nécessaire: comme les totaux marginaux ne peuvent pas changer, il suffit de connaître une des valeurs théoriques pour pouvoir les déterminer toutes. Ce sera 177 pour les consonnes chez Horace, et, chez Goethe, 128 pour les voyelles et 198 pour les consonnes. Pour la commodité, les voici mises en tableau.

	Horace	Goethe	Total
Voyelles	114	128	242
Consonnes	177	198	375
Total	291	326	617

Ceci aide à comprendre de manière intuitive une notion qui réapparaîtra dans divers tests, la notion de degrés de liberté, qu'on symbolise d'habitude par la lettre grecque *nu* et à laquelle les Américains renvoient souvent par l'abréviation *df* (*degrees of freedom*). Puisque, dans un tableau de 2 x 2, il suffit de déterminer une valeur pour que les trois autres le soient *ipso facto*, on exprime cette particularité en disant qu'il y a ici un seul degré de liberté. D'une manière générale, ce nombre s'obtient en multipliant le nombre de colonnes moins un par le nombre de lignes moins un.

Revenons à notre problème. Les valeurs théoriques sont celles que l'on s'attend à trouver si aucun facteur extérieur ne vient troubler l'action du simple hasard. Mais le hasard admet des écarts, la fréquence de ces derniers étant en raison inverse de leur grandeur. Si on considère une seule variable, la probabilité de tels écarts est souvent mesurable par la loi normale (courbe en cloche); on l'évalue alors par le calcul de l'écart-réduit[3]. Mais dans le cas d'un tableau tel que le nôtre, ce qui est en question, ce

2. On le trouve dans *Goethes Gedichte in zeitlicher Folge* 1982 (rééd. 1990): 844, Francfort/Main, Insel.

3. Sur l'écart-réduit, voir par exemple Ch. Muller 1992 (réimpr. de l'éd. de1973): 62-67.

n'est pas seulement la possibilité d'un écart, mais c'est aussi sa compatibilité avec les écarts observés pour les autres variables avec lesquelles il forme système. Le statisticien anglais Karl Pearson a montré que la probabilité de ces systèmes d'écarts obéissent à une fonction qu'il a appelée χ^2. La quantité χ^2 s'obtient de la manière suivante. Dans les données que nous examinons, l'effectif de voyelles observé chez Horace est 125 et l'effectif théorique est 114 ; l'écart est de 11 ; on porte cet écart au carré, ce qui a entre autres avantages celui d'éliminer les valeurs négatives. On divise ce carré par l'effectif théorique, ce qui donne 1,0614. Cette opération a pour but de relativiser le carré de l'écart en fonction du niveau où l'on se trouve dans l'échelle des valeurs : entre une donnée observée 10 et une valeur théorique 20, la différence est 10, le carré, 100 et le quotient par l'effectif théorique, 5 ; tandis que, pour un effectif observé 90 et un effectif théorique 100, l'écart est de nouveau 10 et le carré, 100, mais le quotient est 1. On voit qu'un écart de même valeur cardinale prend une importance très différente en fonction du niveau dans l'échelle des valeurs absolues. Revenons à nos calculs : on fait les mêmes opérations pour les trois autres données ; dans un tableau 2 x 2, l'écart est chaque fois le même en valeur absolue, mais comme le diviseur change, on obtient respectivement 0,6836, 0,9453 et 0,6111. On additionne alors tous ces termes, ce qui donne 3,3014. En notation algébrique la formule s'écrit (*o* représentant les données observées et *c* les effectifs théoriques calculés) :

$$\chi^2 = S\ (o - c)^2 / c$$

La fonction χ^2 étudiée par Pearson établit, pour chaque valeur de la quantité χ^2, et compte tenu du nombre de degrés de liberté, la probabilité d'un ensemble d'écarts au moins aussi important que celui qu'on teste[4]. Il existe des tables, plus ou moins détaillées, qui indiquent ces probabilités[5]. En voici un bref extrait, où, pour chaque degré de liberté (*df*, une ligne pour chaque degré) est indiquée la probabilité d'une valeur de χ^2 égale ou supérieure à celle reprise dans la table en cas de distribution aléatoire (c'est-à-dire la probabilité d'un ensemble d'écarts égal ou supérieur à celui qui est en examen) ; par exemple, s'il y a 5 degrés de liberté, une valeur de χ^2 égale à 15,086 correspond à une probabilité de 0,01 d'un ensemble d'écarts au moins égal à celui qui a permis de calculer cette valeur.

df / P	0,90	0,50	0,10	0,05	0,01
1	0,016	0,455	2,706	3,841	6,635
2	0,211	1,386	4,605	5,991	9,210
3	0,583	2,366	6,251	7,805	11,241
4	1,064	3,357	7,779	9,488	13,277
5	1,610	4,351	9,236	11,070	15,086

Abrégé d'une table de probabilité de *chi*^{2}

La valeur 3,3014, pour 1 degré de liberté, a une probabilité située entre 0,08327 et 0,06520, c'est dire qu'il y a entre 8,327 et 6,520 chances sur cent, soit approximativement 7 chances sur 100 que les écarts observés soient aléatoires, et, corrélativement, plus de 9 chances sur 10 (93 sur 100) qu'un autre facteur ait agi,

4. Voir Ch. Muller 1992 : 116-127.

5. Très détaillées sont les tables publiées dans E. S. Pearson et H. O. Hartley 1958. La plupart des manuels de statistique publient des abrégés reprenant les valeurs les plus caractéristiques.

empêchant ainsi l'homogénéité de l'ensemble. C'est tout ce que peut nous dire la statistique. Au spécialiste de décider laquelle des deux possibilités choisir. Le choix se fera en fonction des autres éléments (éventuellement qualitatifs) propres aux corpus: ils feront apercevoir des convergences dans l'un ou l'autre sens. Dans le cas de la répartition voyelles / consonnes chez Horace et chez Goethe, l'élément le plus significatif est vraisemblablement la différence de langues. On verra ici, en adoptant la thèse de l'hétérogénéité, une confirmation du sentiment général auquel je faisais allusion au début de cet exposé, à propos de l'abondance des consonnes en allemand. D'autres tests, portant sur d'autres corpus latins et allemands, pourraient soit confirmer soit infirmer cette interprétation.

Considérons maintenant non seulement la distinction voyelle / consonne, mais les divers phonèmes pris isolément. Il conviendra ici d'étudier séparément les voyelles d'une part, les consonnes d'autre part. Voici pourquoi: il y a une voyelle, et une seule, par syllabe; quant aux consonnes, certaines syllabes en sont complètement dépourvues; d'autres ont une consonne ou un groupe de consonnes avant la voyelle (*prae*) ou après elle (*abs*) ou avant et après (*post*). Il est clair que voyelles et consonnes ne fonctionnent pas de la même manière; pour les consonnes, il paraîtrait convenable de dénombrer, comme pré- ou postvocaliques, les occurrences de la *consonne zéro,* puis celles des divers groupes consonantiques, simples ou complexes.

Pour les voyelles, dont je me suis occupé il y a quelques années[6], j'avais mené mon enquête en tenant compte des timbres mais non des longueurs (celles-ci poseraient en effet certains problèmes); de plus, j'avais regroupé les diphtongues avec les voyelles qu'elles comportent et j'avais compté *y* avec *i,* en raison de sa rareté. J'ai alors pris au hasard, dans les textes de douze prosateurs et de douze poètes latins, chaque fois trois échantillons de 1 000 syllabes chacun. De plus, dans les échantillons de poètes, j'ai compté à part les voyelles sous ictus. Pour traiter cette masse de données, il m'a paru opportun de considérer d'abord l'ordre de préférence des divers timbres. Les données recueillies sont exprimées en nombres cardinaux, ce qui permet de mesurer l'écart entre la fréquence d'un timbre et celle d'un autre. Mais, avant même ces écarts, l'ordre des préférences paraît significatif (que l'on pense par exemple aux notes attribuées aux étudiants par divers examinateurs); on peut facilement convertir un tableau cardinal en un tableau ordinal; on utilisera alors un test ordinal, par exemple le coefficient de concordance *W* de Kendall[7], dont le but est de mettre en évidence le degré de concordance ou de discordance de plusieurs classements et d'en évaluer la probabilité aléatoire. Supposons que les *k* rangements (ici, les trois échantillons par auteur) des *N* objets à ranger (ici, les cinq timbres vocaliques) coïncident: l'objet le plus fréquent a trois fois le rang 1, soit, en additionnant les rangs, 3; le suivant a trois fois le rang 2, soit 6, et ainsi de suite. Mais, dans la majorité des cas, les classements diffèrent plus ou moins profondément, ce qui se traduit par le fait que les sommes des rangs attribués aux objets ont une tendance à se rapprocher. En additionnant ces sommes de rangs (dans notre cas, et si tous les classement sont identiques, $3+6+9+...$) et en divisant par le nombre d'objets (soit *N*), on obtient la moyenne des sommes des rangs, par rapport à laquelle on calcule les écarts de chacune de ces sommes; on porte ces écarts au carré et on additionne. Il est possible de montrer que le résultat obtenu à partir de classements rigoureusement identiques représente le maximum possible pour ce calcul et qu'il prend

6. É. Évrard 1983: 157-166.

7. M. G. Kendall 1948: ch. 6. On trouve un exposé de ce test dans S. Siegel 1956: 229-239 et 286.

alors la forme suivante: $s^{max} = 1/12\ k^2(N^3 - N)$. Dès lors, le coefficient W sera le quotient de la somme des carrés des écarts observés par la somme maximum; sa valeur se situe entre 1 (si tous les classements sont rigoureusement identiques) et 0 (s'ils diffèrent au maximum); une table précise la probabilité des divers résultats en fonction des valeurs de N et de k.

Voici un exemple de ce calcul. Il a trait aux trois échantillons de Cicéron prosateur. Les valeurs cardinales une fois converties en valeurs ordinales, le tableau se présente comme suit:

Voyelle	I	II	III	Total
a	3	3	3	9
e	2	2	2	6
i	1	1	1	3
o	5	4	5	14
u	4	5	4	13
Total	15	15	15	45

La moyenne des sommes des rangs par voyelle est égale à la somme totale des rangs attribués (soit 45) divisée par le nombre de voyelles (soit 5), c'est-à-dire 45 / 5 = 9. Le total des carrés des écarts entre la somme des rangs de chaque voyelle et la moyenne est égal à:

$$(9 - 9)^2 + (9 - 6)^2 + (9 - 3)^2 + (9 - 14)^2 + (9 - 13)^2 = 86.$$

La somme maximum est

$$s^{max} = 1/12 \times 3^2 \times (5^3 - 5) = 90.$$

Dès lors, on a $W = 86 / 90 = 0{,}96$ (par excès).

Le test appliqué successivement aux échantillons de prose, de poésie et des voyelles sous ictus prend les valeurs suivantes: W = 0,91 (prose), W = 0,84 (poésie), W = 0,82 (voyelles sous ictus). Calculé sur l'ensemble des échantillons, il prend la valeur W = 0,77. Tous ces résultats sont compatibles avec l'hypothèse d'homogénéité. Pour préciser, notons que dans tous les échantillons, les trois premières places sont occupées par *a, e* et *i*, dans l'un des ordres possibles, tandis que *o* et *u* sont confinés aux deux dernières places. Des 108 échantillons testés, deux seulement divergent légèrement: un de Cicéron, où, à l'ictus, *i* et *u* sont *ex aequo* et un de Lucrèce, où *o* apparaît au 3e rang, tandis que *i* est rejeté au 5e. Il ressort de là que le latin marque une préférence accusée pour les voyelles non-arrondies, si nous suivons la nomenclature de Troubetzkoy[8].

Comme le test appliqué à l'ensemble donne un résultat inférieur à ceux qui résultent de l'examen d'un groupe isolé (prose, poésie, ictus), il y a lieu de préciser les observations à ce niveau. En prose, 29 échantillons sur 36 placent *i* en tête, 26 ont *e* en 2e position et 32 mettent *a* en 3e. En poésie, c'est *e* qui l'emporte dans 29 échantillons; la seconde et la troisième places se partagent presque également entre *a* (17 fois en 2e position dont une *ex aequo* avec *i* et 16 fois en 3e) et *i* (16 fois en 2e et 16 fois en 3e).

8. N. S. Troubetzkoy 1949 (19572): 106.

Enfin, à l'ictus, *a* prédomine et est 20 fois en première position, *e* est 15 fois premier et 15 fois second, et *i* est 24 fois troisième. Ces résultats manifestent une différence assez constante du vocalisme selon qu'on est en prose, en poésie ou, plus particulièrement, à l'ictus. Interprétés dans les termes de M. Grammont[9], ils conduisent à dire que la prose se caractérise par son acuité et la poésie, par son éclat. On observera en particulier que Cicéron prosateur a toujours l'ordre *i, e, a*, tandis que, dans ses vers, c'est toujours *e* qui vient en tête.

Au sein de la triple régularité qui vient d'être exposée, il y a place pour des variantes qui ne dérangent guère les rangs, mais qui se marquent si l'on retourne aux données cardinales. Considérons les trois échantillons de Caton. Ils ont l'ordre, habituel en prose, *i, e, a, o, u*, sauf que le premier échantillon inverse les rangs de *i* et de *e*. Le coefficient de corrélation pour ces trois textes s'élève à $W = 0,96$, ce qui est particulièrement satisfaisant. Mais l'examen des valeurs cardinales montre, dans les limites de l'ordre attendu, des écarts très considérables : la variation la plus forte est celle de *o*, qui va de 157 à 200, avec un écart de 43. Pour apprécier correctement cette situation, le mieux sera de soumettre les valeurs cardinales à un test de χ^2 : pour chaque voyelle, la somme des trois effectifs, divisée par 3 000 (nombre total des voyelles des trois textes) indique une fréquence qui, multipliée par 1 000 (nombre de voyelles par échantillon) indique l'effectif théorique. On notera qu'une fois remplies les quatre premières cases des deux premiers échantillons, les sept effectifs théoriques restants sont déterminés d'office par les totaux marginaux ; il y a donc 8 degrés de liberté, soit (5-1)x(3-1).

Le tableau des valeurs cardinales pour Caton se présente comme suit.

Voyelle	I	II	III	Total	%	Théor.
a	209	216	191	616	0,2053	205
e	265	235	237	737	0,2457	246
i	240	241	279	760	0,2533	253
o	160	200	157	517	0,1723	172
u	126	108	136	370	0,1233	123

$\chi^2 = 17,801$ $df = 8$ $0,0253>P>0,02123$

Le total des carrés des écarts entre valeurs observées et valeurs calculées divisés chaque fois par la valeur calculée correspondante se monte à $\chi^2 = 17,801$, ce qui, d'après les tables, correspond à une probabilité inférieure à 0,03. L'hypothèse aléatoire peut donc être rejetée en toute sécurité. Quant à l'explication du phénomène, elle est fournie par le deuxième texte (*de Agr.*, LI - LVII), qui consiste en une série de conseils exprimés par des impératifs futurs (j'en ai compté 72).

Voici un autre exemple. Les trois textes des tragédies de Sénèque ne présentent pratiquement pas d'anomalie du point de vue ordinal (coefficient de concordance $W = 0,8889$). Mais les valeurs cardinales révèlent des écarts exceptionnels.

9. M. Grammont 1933 (1946[3]) : 384 et 385.

Sénèque, tragédies

Voyelle	I	II	III	Total	%	Théor.
a	222	219	197	638	0,2127	213
e	260	285	240	785	0,2617	261
i	256	212	266	734	0,2447	245
o	130	121	147	398	0,1327	133
u	132	163	150	445	0,1483	148

$\chi^2 = 18{,}2915$ $df = 8$ $0{,}02123 > P > 0{,}01777$

La valeur de χ^2 est de 18,2915, ce qui, pour une probabilité d'environ 0,02, conduit au rejet de l'hypothèse aléatoire. C'est surtout le troisième échantillon qui présente des singularités : c'est lui qui a l'effectif le plus faible pour *a* et pour *e*, et l'effectif le plus élevé pour *i* et pour *o*. Cet échantillon provient de *Médée* et est constitué pour une bonne part par les vers 579-613, dans lesquels le chœur, effrayé par Médée et par l'annonce de sa vengeance, se répand en considérations angoissées qu'exprime bien la sonorité sourde des *o*, tandis que les *i*, par leur acuité, soulignent l'intensité de son inquiétude.

On voit ainsi la complémentarité des divers modes d'investigation : un test ordinal mesurant un degré élevé de concordance, laquelle s'accommode cependant de singularités que le test cardinal de χ^2 met en évidence. On voit aussi que la statistique, après avoir indiqué ce qui mérite explication, laisse le champ libre au spécialiste des langues et des textes dans la recherche de cette explication.

Dans les développements précédents, il a été à diverses reprises question de probabilité. Ce n'est pas qu'on imagine les faits de langue et de textes comme probables en vertu d'un caractère fixe. Ce qu'on utilise sous ce terme n'est autre que la fréquence relative. Il me semble que, quand on précise, pour un phénomène de langue, la probabilité aléatoire, c'est là une mesure qui est souvent en corrélation avec les chances qu'un lecteur ou auditeur attentif a de remarquer un caractère particulier : une probabilité aléatoire faible correspond sans doute à une situation où un tel lecteur percevra une anomalie, tandis qu'en cas de probabilité très forte, il est vraisemblable qu'il ne sentira rien d'anormal.

Mais il se peut que le lecteur se laisse entraîner par de fausses apparences : dans ces cas, la statistique contribuera à redresser la situation. Je voudrais vous exposer un cas de ce genre, auquel il a été fait allusion dans le premier §. Il me donnera l'occasion de vous montrer une autre technique quantitative, celle des probabilités composées. Dans son livre sur *La poésie latine*, Herescu s'intéresse au timbre des voyelles sous ictus[10]. Il est particulièrement frappé par les homophonies, par exemple du premier et du dernier ictus d'un vers, de l'ictus précédant la césure et de l'ictus final. Il admire surtout "cette acrobatie sonore qu'est l'homophonie de la totalité des voyelles fortes d'un vers" (Herescu 1960 : 85).

En fait, chaque timbre a, dans une œuvre donnée, une fréquence sous ictus ; on peut même apporter plus de précision en comptant cette fréquence sous chacun des six ictus, avec d'éventuelles différences d'un ictus à l'autre. Dans l'hypothèse aléatoire, c'est-à-dire si l'on suppose que le timbre d'un ictus n'influence en aucune manière celui

10. N. I. Herescu 1960. J'ai étudié les assertions qu'y énonce Herescu dans Ét. Évrard 1984.

de l'ictus suivant, on peut utiliser le calcul des probabilités composées, en d'autres termes, de la multiplication des probabilités. Soit un événement de probabilité 0,05 dans un corpus de 1 000 unités ; son effectif probable est de 50 ; la probabilité du même événement dans une position suivant immédiatement celles de ces 50 occurrences (ou dans toute autre position strictement définie) est toujours de 0,05, ce qui donne un effectif de 50x0,05 = 2,5. On arrive au même résultat si on multiplie l'une par l'autre la probabilité des deux événements (soit, ici, 0,05x0,05 = 0,0025) et que l'on multiplie ce quotient par la taille du corpus (c'est-à-dire 0,0025x1 000 = 2,5). En somme, la probabilité qu'un événement ET un autre se produisent est égale au produit des probabilités de ces deux événements (observons qu'ici *ET* marque non une addition mais une multiplication). Revenons maintenant à nos voyelles sous ictus. Dans les *Géorgiques*, sous chacun des six ictus de l'hexamètre, la voyelle *a* présente les fréquences respectives 0,2756, 0,2349, 0,2144, 0,2253, 0,2724, 0,2943. Le produit de ces six fréquences est égal à 0,000251 qui, multiplié par la taille des *Géorgiques* (2 188 vers) donne 0,55. Cet effectif théorique montre le caractère improbable de la combinaison testée. En fait, Herescu ne cite, pour les *Géorgiques*, qu'un seul vers homophone, IV 466 (homophonie en *e*) et un pour l'*Énéide*, IV 254 (homophonie en *i*) qui compte pourtant 9 896 vers. On ne voit donc pas que les vers de ce type soient plus abondants que la prévision, d'où l'on peut conclure que le poète se livre rarement à l'acrobatie dont le loue le critique, ou plutôt qu'il ne la réalise que de manière aléatoire.

Faisons l'expérience sur des combinaisons plus restreintes : sur les trois premiers ictus des vers, ou sur le premier et le dernier. Sans entrer dans le détail des calculs, qu'il me suffise de dire que, tant pour les *Bucoliques* que pour les *Géorgiques*, les effectifs observés d'homophonie des trois premiers ictus ne sont jamais supérieurs aux effectifs calculés, et qu'ils y sont le plus souvent inférieurs, d'où l'on est tenté de conclure que, si Virgile est sensible aux homophonies de ce type, c'est plutôt pour les éviter. En revanche, les homophonies du premier et du dernier ictus, dans les deux mêmes œuvres, sont presque toujours au moins égales et même souvent un peu supérieures à la prévision. Ceci nuance les affirmations de Herescu, en indiquant que Virgile a une légère tendance à souligner l'unité du vers grâce à l'homophonie de ses deux ictus extrêmes, mais qu'en revanche il évite la monotonie qu'engendrerait l'homophonie d'éléments conjoints.

La méthode exposée en dernier lieu est féconde aussi en d'autres domaines. On parle souvent d'allitérations, et on a tendance à croire qu'elles sont un ornement du style, donc un procédé voulu. En fait, si on se limite aux mots commençant par le même phonème, ou, plus précisément encore, par la même consonne, des suites de mots ayant même initiale sont inévitables et il est possible d'en calculer la fréquence aléatoire. Si on connaît, pour un texte donné, la fréquence de telle initiale de mot (par exemple *p*), les probabilités composées permettent de calculer les chances (et, donc, les effectifs théoriques) de rencontrer par hasard une suite de deux, trois, *x* mots ayant cette initiale. Le calcul des écarts-réduits, dont il sera question dans le § suivant, autorise la comparaison entre cet effectif théorique et celui observé. Plutôt que d'en donner des exemples, il convient sans doute de considérer qu'en voilà assez sur la statistique phonémique.

3. STATISTIQUE LEXICALE (Sylvie Mellet)

Les études thématiques comparées sont classiques en critique littéraire : on se demande par exemple si la lyrique élégiaque romaine se manifeste de la même manière chez tous ses représentants ou si l'on observe des différences thématiques entre Tibulle

et Properce. On peut aussi vouloir comparer entre elles les différentes œuvres d'un même écrivain, par exemple les tragédies de Sénèque : celles-ci recourent-elles dans l'ensemble à un même vocabulaire ou chacune d'elles offre-t-elle des spécificités remarquables ?

Dans ce domaine comme dans d'autres, on voit s'exprimer des avis divergents. Il convient donc, pour trancher, de s'appuyer sur des données tangibles et, à cet égard, le dénombrement des mots et des expressions verbales de chaque texte étudié peut fournir des éléments de réflexion précieux. Mais que dénombrer exactement et comment traiter les données ainsi obtenues ?

3.1. Le choix des unités de décompte

Il paraît trivial de le rappeler, mais décompter des unités, les comparer entre elles, les additionner, c'est les considérer - au moins momentanément, le temps d'une hypothèse - comme des occurrences d'une même entité, d'une même catégorie générale : elles appartiennent à la même classe, car il est bien connu qu'on n'additionne pas des pommes avec des poires sauf à les considérer plus généralement comme des fruits à pépins.

Dans le cas d'une étude lexicale, il va donc falloir adopter un principe stable permettant de segmenter d'une manière récurrente les textes étudiés pour donner naissance à des listes de termes comparables. En statistique linguistique française, le principe le plus souvent adopté a longtemps été de s'appuyer sur les **formes graphiques** : l'unité de décompte était donc la chaîne de caractères comprise entre deux blancs. On voit tout de suite les avantages de cette méthode, et en premier lieu sa commodité pour le traitement automatique des textes : la forme graphique est immédiatement reconnaissable et isolable par l'ordinateur (n'importe quel traitement de texte y parvient), sans encodage particulier.

A côté de cet avantage pratique, cette procédure présente aussi un avantage théorique : elle a le mérite de ne susciter aucune intervention plus ou moins arbitraire sur le texte avant l'analyse et de laisser par conséquent subsister la totalité des informations qu'il contient ; ainsi le choix des formes graphiques interdira de rassembler sous une même entrée - dans une même unité de décompte - les formes singulier et pluriel d'un substantif ; or il est des cas où le maintien des deux formes en deux unités distinctes peut être important : L. Lebart et A. Salem[11] citent à ce propos les deux syntagmes *défense de la liberté* et *défense des libertés* qui, dans les textes récents, sont caractéristiques de deux courants politiques opposés.

Mais on voit également les inconvénients de cette méthode : l'impossibilité de rassembler dans une même unité les différentes formes d'un même lexème peut aussi être gênante, notamment dans le cas des verbes conjugués. Par ailleurs les cas d'homonymie ne sont pas pris en compte, engendrant souvent un "bruit" important sur le résultat de certaines recherches (cf. le problème que constitue le participe passé du verbe *être* dans une étude thématique sur les saisons chez tel ou tel poète ou romancier). Enfin, certaines entrées du dictionnaire se présentent sous la forme de syntagmes composés de plusieurs unités graphiques (*pomme de terre*, *chef d'œuvre*) et inversement des entités lexicales différentes peuvent sporadiquement fusionner dans une même unité graphique (les formes dites contractes de l'article par exemple).

---

11. Lebart et Salem 1994 : 34.

En latin, de tels inconvénients s'alourdissent encore dans la mesure où on a affaire à une langue flexionnelle provoquant de nombreuses homographies occasionnelles et dont l'orthographe n'est pas entièrement stabilisée (cf. les variations graphiques liées aux assimilations consonantiques, aux haplologies, à la débilité de certains phonèmes) ; sans compter que le concept même de forme graphique est peu pertinent sans doute pour une langue qui pratiquait la *scriptio continua*.

On prônera donc, pour le latin tout particulièrement, un autre choix méthodologique, celui de la lemmatisation. C'est le choix effectué dès l'origine par le L.A.S.L.A., qui a ainsi constitué une banque de données de textes latins entièrement lemmatisés, c'est-à-dire pour lesquels chaque forme a été analysée et rattachée à une entrée de dictionnaire ; le texte est certes conservé dans sa forme originale, mais on lui adjoint un index des lemmes classés par ordre alphabétique, comprenant bien sûr en outre la référence de chacune des occurrences de chaque lemme dans le texte. Moyennant un petit programme informatique complémentaire de va-et-vient entre les deux fichiers on obtient aisément une conversion du texte d'origine en un texte lemmatisé ; en voici un exemple :

> *Nihil ex his quae animum fortuito impellunt affectus uocari debet ; ista, ut ita dicam, patitur magis animus quam facit. Ergo affectus est non ad oblatas rerum species moueri, sed permittere se illis et hunc fortuitum motum prosequi* (Sén., *de Ira* II, 3, 1)
>
> NIHIL EX HIC1 QVI1 ANIMVS FORTVITO IMPELLO AFFECTVS1 VOCO DEBEO ISTE VT4 ITA DICO2 PATIOR MAGIS2 ANIMVS QVAM1 FACIO. ERGO2 AFFECTVS1 SVM1 NON AD2 OFFERO RES SPECIES MOVEO SED PERMITTO SVI1 ILLE ET2 HIC1 FORTVITVS MOTVS PROSEQVOR

Les atouts de la lemmatisation[12] d'un texte latin sont évidents, puisqu'elle permet de :

- rassembler au sein d'une même unité les différentes formes graphiques et morphologiques d'un même mot ;
- lever les ambiguïtés dues aux homographies occasionnelles et aux homonymies lexicales (au moyen d'indices numériques suivant immédiatement le lemme : par ex. le chiffre 1 qui suit *affectus* le caractérise comme un substantif par opposition au participe-adjectif qui serait codé 2 ; le pronom *hic* est différencié de l'adverbe 'ici' ; *dico, -ere* est affecté de l'indice 2 alors que *dico, -are* le serait de l'indice 1, etc.) ;
- conserver l'analyse morpho-syntaxique préalable sous la forme d'un code qu'on pourra éventuellement exploiter de manière automatique pour des recherches sur les catégories grammaticales.

On ne peut cependant pas passer sous silence non plus les difficultés d'établir une norme de lemmatisation la plus fiable possible. Citons en désordre quelques-unes d'entre elles :

- deux formes identiques de sens différents constituent-elles une paire homonymique nécessitant deux entrées dans l'index des lemmes ou représentent-elles deux facettes d'un seul et même mot ? La décision devra être prise avec l'aide d'un dictionnaire de référence choisi une fois pour toutes et dont l'*auctoritas* devra être acceptée pour l'ensemble de la banque textuelle de façon à ce que les choix faits en la

12. Cf. S. Mellet 1996.

matière restent contrôlables et reproductibles (par ex. si l'on insérait de nouveaux textes pour augmenter la base) ;

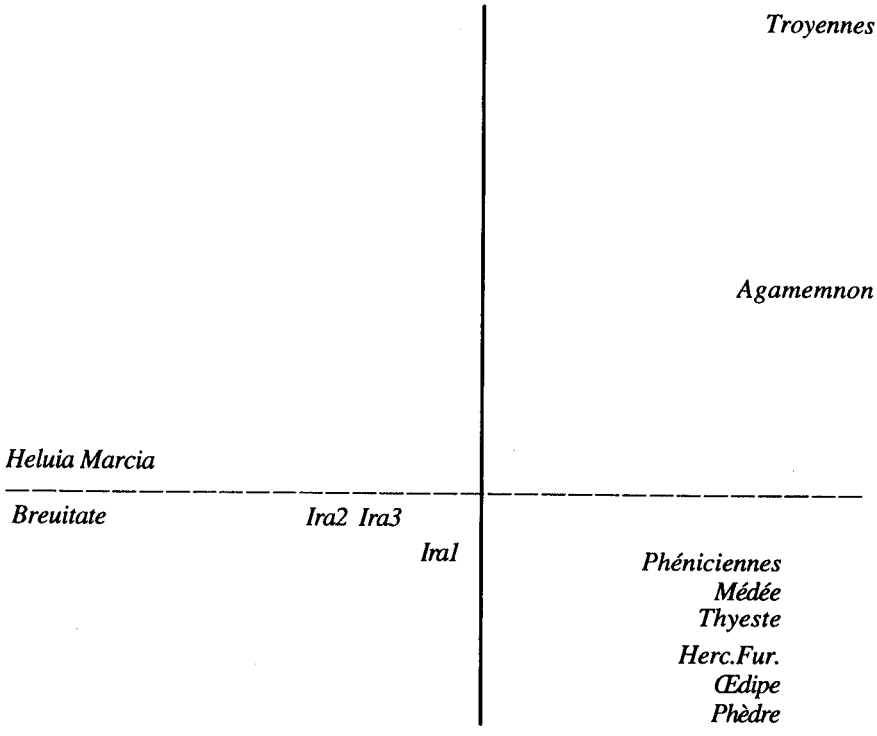
- on rencontrera le même type de choix difficile sur le plan grammatical. Prévoira-t-on une ou deux entrées pour les verbes *sum* et *eo* selon qu'ils fonctionnent comme auxiliaires ou comme verbes à sens plein ? Combien d'entrées accordera-t-on à des conjonctions comme *quod* et *ut* ?

- ponctuellement des hésitations pourront naître devant des syntagmes comme *et iam / etiam, quo modo ? / quomodo ?, animum aduerto / animaduerto*, etc.

De telles difficultés justifient la remarque de Ch. Muller (rapportée par L. Lebart et A. Salem 1994 : 37) :

> "La norme devrait être acceptable à la fois pour le linguiste, pour ses auxiliaires, et pour le statisticien. Mais leurs exigences sont souvent contradictoires. L'analyse linguistique aboutit à des classements nuancés, qui comportent toujours des zones d'indétermination ; la matière sur laquelle elle opère est éminemment continue, et il est rare qu'on puisse y tracer des limites nettes ; elle exige la plupart du temps un examen attentif de l'entourage syntagmatique (contexte) et paradigmatique (lexique) avant de trancher. La statistique, dans toutes ses applications, ne va pas sans une certaine simplification des catégories ; elle ne pourra entrer en action que quand le continu du langage a été rendu discontinu, ce qui est plus difficile au niveau lexical qu'aux autres niveaux ; elle perd de son efficacité quand on multiplie les distinctions ou quand on émiette les catégories ; opérant sur des ensembles très vastes elle tolère difficilement une casuistique subtile qui s'arrête à analyser les faits isolés."

Néanmoins, s'il est vrai que les catégories linguistiques se réalisent le plus souvent sur le mode d'un *continuum* dans lequel il va falloir créer plus ou moins artificiellement des discontinuités et opérer des simplificatons momentanées, la statistique peut cependant donner des résultats extrêmement suggestifs, ouvrir des pistes de recherche ou confirmer des hypothèses quand elle opère sur de grands corpus, quelles que soient les approximations que l'on a pu faire lors de la préparation du corpus. La preuve en est que les calculs aveugles de la statistique retrouvent les évidences de la comparaison entre plusieurs textes. En voici un exemple réalisé à partir d'un corpus comprenant deux consolations de Sénèque (*ad Heluiam* et *ad Marciam*), deux dialogues (*de Ira* et *de Breuitate uitae*) et huit de ses tragédies :



Analyse factorielle visualisant les distances ou les proximités entre les œuvres de Sénèque en fonction de l'ensemble du vocabulaire utilisé par chacune d'elle (vocabulaire partagé *vs* vocabulaire exclusif)[13].

3.2. Le vocabulaire d'un texte : mots clés, spécificités et distance lexicale

Est venu le moment d'introduire quelques notions simples, mais indispensables, et tout d'abord de se mettre d'accord sur la terminologie employée.

En statistique comme en linguistique, le mot "mot" est ambigu : il peut désigner les occurrences des formes graphiques qui se suivent pour constituer la chaîne linéaire d'un texte ou les termes différents du dictionnaire qui fournissent le vocabulaire employé dans le texte. L'ambiguïté demeure, que le texte soit lemmatisé ou non ; reprenons l'extrait du *de Ira* cité ci-dessus, sous ses deux formes :

Chacun de ces deux textes comprend 37 occurrences ; mais le premier, non lemmatisé, n'offre que 36 formes différentes (*affectus* est employé deux fois) et le second (lemmatisé) n'en offre que 34 (les lemmes AFFECTUS1, ANIMUS et HIC1 sont employés deux fois chacun).

13. Cette analyse a été réalisée grâce au logiciel *Hyperbase* d'Étienne Brunet (fonction : 'Analyse factorielle du dictionnaire'). On reviendra ultérieurement sur les principes de l'analyse factorielle.

En règle générale, on dira que ce texte lemmatisé est composé

- de 37 **occurrences** (n = 37) et ce nombre n définit la **taille** du texte
- et de 34 **mots** qui constituent le **vocabulaire** du texte (V = 34)

Le nombre d'occurrences d'un même mot dans un texte représente sa **fréquence absolue**. Dans le court extrait ci-dessus, nous avons essentiellement des hapax et trois mots de fréquence 2.

Ces paramètres suffisent pour calculer les éléments caractéristiques - ou spécificités - de chaque texte par rapport aux autres. On adapte pour ce faire des tests statistiques classiques qui s'appuient sur un modèle probabiliste. Le raisonnement est le suivant :

Admettons[14] que les huit tragédies de Sénèque étudiées offrent en tout 50 000 occurrences et que *Phèdre* en compte 7 000 ; l'index du corpus nous indique par ailleurs que le mot *animus* apparaît 150 fois dans l'ensemble des tragédies. La probabilité p de rencontrer *animus* plutôt qu'un autre mot dans cet ensemble de textes est donc de 150/50 000 = 0,003 (soit trois chances sur mille). En revanche la probabilité de rencontrer un autre mot que *animus* est égale à (50 000-150)/50 000 = 0,997 : on a 997 chances sur mille de rencontrer autre chose que *animus* ; c'est ce qu'on appelle la probabilité complémentaire de p, traditionnellement désignée par la lettre q et on remarquera que $p + q = 1$; c'est normal, dans un texte, on a mille chances sur mille de rencontrer *animus* ou autre chose que *animus* !

Si l'on suppose que ce mot *animus* est distribué très régulièrement dans l'ensemble du corpus (c'est-à-dire que sa distribution n'est en rien influencée par la thématique de chacune des œuvres qui constituent celui-ci, ni par la date de leur rédaction, etc.), on pourrait donc s'attendre à rencontrer 21 fois *animus* dans *Phèdre* (7 000 x 0,003). On appelle cette estimation (ou ce pari) l'espérance mathématique ($E = np$)[15]. Mais bien évidemment, même en cas d'une distribution aléatoire de la variable, on peut aussi s'attendre à observer certaines fluctuations autour de cette valeur : le hasard ne fait pas toujours bien les choses, et bien sûr si l'on répète plusieurs fois le tirage au sort de 7 000 mots dans une urne qui en contient 50 000, on peut avoir plusieurs tirages qui ne fournissent que 20, voire 19 ou 18 occurrences du mot *animus* sur les 21 attendues, ou au contraire des tirages qui en fournissent plus de vingt-et-une.

Ces fluctuations sont estimées au moyen d'un paramètre qu'on appelle la variance, très facile à évaluer puisque il est obtenu en calculant le produit npq ; la racine carrée de ce produit fournit **l'écart-type** : $s = \sqrt{npq}$.

Ces notions sont très importantes car elles permettent d'évaluer objectivement l'importance des écarts entre les effectifs observés réellement dans le corpus étudié et les effectifs espérés - ou du moins attendus en vertu de l'espérance mathématique : on estime en général que des écarts observés dont la valeur se situerait en deçà de la somme de deux écarts-types ne seraient pas significatifs, ne suffiraient pas à autoriser des déductions sur l'existence d'un phénomène particulier.

---

14. L'exemple proposé repose bien sûr sur des données numériques arrondies pour faciliter les calculs et la démonstration, mais reproduit néanmoins les ordres de grandeur réels du corpus.

15. On admet donc l'hypothèse que le vocabulaire du corpus étudié répond à la "loi normale", hypothèse qui se vérifie sur les corpus de grande taille.

Concrètement la mesure de la pertinence de l'écart se fait en divisant l'écart absolu par l'écart-type, ce qui donne **l'écart réduit**, et en se reportant à une table permettant d'évaluer la probabilité d'obtenir un tel écart par le seul jeu du hasard. Revenons à notre exemple du mot *animus* dans les tragédies de Sénèque.

On a donc vu que l'espérance mathématique de ce mot dans *Phèdre* était $E = np = 7\,000 \times 0{,}003 = 21$

Or un dépouillement du texte de *Phèdre* fait apparaître que cette tragédie contient en réalité 25 occurrences de ce mot. L'écart observé entre 25 et 21 est-il important compte tenu de l'étendue du corpus étudié et de la probabilité attachée au mot sous examen ? Considérera-t-on que cet écart est de 4 occurrences 'seulement' ou qu'il atteint 'déjà' 1/5 de l'effectif attendu ? Les deux perspectives sont fort différentes pour décrire la même réalité et il devient difficile de tirer des conclusions sur la place de ce mot dans la pièce. La seule façon scientifique de répondre à ces questions est de calculer l'écart réduit correspondant à cet écart absolu, soit : écart absolu / écart-type. Ce qui donne, avec les valeurs numériques :

$$25 - 21 / \sqrt{7\,000 \times 0{,}003 \times 0{,}997} = 0{,}87$$

On consulte alors la table des écarts réduits fournie par tous les manuels[16] et l'on constate qu'un tel écart a un peu moins de 40 chances sur 100 d'être dû au seul hasard : autant dire qu'il n'a aucune chance d'être significatif et que notre recherche, sur ce point précis, va s'arrêter là.

Si l'on ne dispose pas de table, on peut retenir qu'un écart réduit commence à être significatif à partir de la valeur 2 et devient fortement significatif à partir de 2,5.

De tels calculs répétés sur l'ensemble du vocabulaire constitutif d'un corpus permettent de mettre en évidence le vocabulaire spécifique d'une œuvre et, donc, généralement ses thématiques propres. Là encore on retrouve des évidences, comme, dans la liste des spécificités de *Phèdre*, la présence des noms propres, de *castus*, *amor*, *stuprum*, *furor* ; mais on peut aussi avoir des surprises ou simplement des pistes d'exploration plus suggestives, par exemple à partir de la présence dans la même liste d'éléments tels que le point d'interrogation, associé à la présence de *cur*, *quisnam*, des exclamations *en*, *heu*, *o*, qui peuvent orienter vers une étude des modalités énonciatives dans le texte.

16. Notamment dans celui de Ch. Muller 1973 : 175.

Hippolytus	14	14	9,05
Castus	20	15	7,65
Venus	17	13	7,22
Phaedra	6	6	5,92
Forma	15	10	5,71
Amor	46	19	5,13
Theseus	20	11	5,11
Fingo	10	7	4,96
En	48	19	4,90
Ambiguus	5	4	4,14
Amo	19	9	4,04
Saltus	8	5	3,83
Inuoco	8	5	3,83
Frenum	17	8	3,79
Cur	43	15	3,76
Cura	24	10	3,75
Per	166	41	3,68
Ritus	6	4	3,61
Laqueus	6	4	3,61
Libido	6	4	3,61
Filum	6	4	3,61
Dea	18	8	3,58
Pudor	33	12	3,54
Insanus	15	7	3,51

Liste des spécificités de *Phèdre*

4. DIVERSITÉ ET DIVERSIFICATION LEXICALES (Étienne Évrard)

Dans la section précédente, on a présenté des recherches quantitatives relatives aux mots (lemmes ou formes) dans lesquelles, outre les données quantitatives, il était tenu compte d'autres caractères, qualitatifs ceux-là, tels l'"individualité" du ou des mots, leur signification, leurs connotations, etc. (les mots-clés constituant un bon exemple de ce genre de démarche).

Il sera maintenant question de techniques où la sélection (d'ailleurs provisoire) des caractères pris en compte est plus sévère et plus complètement tournée vers le quantitatif.

Quand un corpus a été saisi en mémoire d'ordinateur, les mots-occurrences étant pourvus de descripteurs tels que (outre la référence complète) le lemme et les caractéristiques morphologiques et, dans une certaine mesure, syntaxiques, l'ordinateur peut, en un premier temps, trier les lemmes en ordre alphabétique, puis énumérer les occurrences de ces lemmes, avec référence, et en compter le nombre: on dispose alors d'un index. En un second temps, il peut sélecter les lemmes et leur fréquence, donnant ainsi une liste des lemmes en ordre alphabétique, avec indication de la fréquence. Ensuite, il va disposer les lemmes en ordre de fréquence croissante ou décroissante.

A partir de cette liste, il est possible de dresser ce que l'on appelle le tableau de distribution du vocabulaire. Cela consiste à compter le nombre de lemmes qui ont la même fréquence et, pour chaque fréquence représentée, à écrire en une ligne, d'abord (1ère colonne) la fréquence, puis (2e colonne) le nombre de lemmes qui ont cette fréquence, enfin (3e colonne) le produit de cette fréquence par le nombre de lemmes, ce qui correspond au nombre d'occurrences appartenant aux lemmes de la fréquence considérée. Le total de la 2e colonne est le nombre de lemmes du corpus; le total de la 3e, le nombre total d'occurrences de ce corpus. Comme on l'a remarqué, à ce niveau de sélection, seules des données quantitatives entrent en jeu, mais, au moment d'interpréter les résultats des tests appliqués à ces tableaux, les caractères "individuels" des lemmes ou des formes reprendront toute leur importance et fourniront souvent la clé interprétative adéquate.

Voici le tableau de la distribution du vocabulaire du *Moretum*[17].

Fréquences	Effectifs	Nombre d'occurrences
1	418	418
2	70	140
3	18	54
4	11	44
5	2	10
6	4	24
7	2	14
8	2	16
9	1	9
14	1	14
22	1	22
39	1	39
Totaux	531	804

La division du nombre total des occurrences par celui des lemmes donne un quotient qui est le nombre moyen d'emplois par lemme: dans le cas du *Moretum*, 804 / 531 = 1,151. Nous approchons ici de la notion de diversité (que l'on appelle souvent, mais à tort, me semble-t-il, richesse) du vocabulaire, puisque plus il y a de vocables représentés dans un corpus, plus faibles sont, pour chacun d'entre eux, les probabilités d'apparitions.

C'est un indicateur relativement peu fiable. Tout d'abord, il arrive que, pour deux textes de longueur à peu près identique, on trouve une même moyenne d'emplois par lemme, alors que dans le détail de la distribution se remarquent des différences appréciables: il suffit que ces différences se compensent pour que la moyenne n'y soit pas sensible. En voici un exemple; les 500 premières occurrences de la *Consolation à Polybe* et de la *Consolation à Helvia* appartiennent à 289 lemmes distincts, ce qui donne une moyenne d'emploi de 1,73; mais, dans la première, 77 occurrences appartiennent à des lemmes qui ont une fréquence supérieure à 8, tandis que, dans la seconde, ce nombre n'est que de 59.

D'autre part, la moyenne correspond à une situation qui, dans le cas de textes, n'est absolument pas parlante: une moyenne indique, en effet, la fréquence qu'aurait

17. Cf. Évrard 1982.

chaque lemme si tous les lemmes avaient la même fréquence ; c'est là, on l'avouera, une situation que l'on ne peut guère imaginer. Par ailleurs, on n'a aucune idée des limites entre lesquelles pourrait varier la moyenne d'emplois par lemme ; ceci la rend difficilement interprétable.

On notera encore que la moyenne est très sensible à la longueur du texte ; ainsi Sénèque, dans la *Consolation à Marcia*, emploie 2204 lemmes qui fournissent 8384 occurrences, soit une moyenne de 3,8040, tandis que dans les *Lettres à Lucilius*, 6749 lemmes donnent 119243 occurrences, soit une moyenne de 17,6682.

Enfin, les statisticiens considèrent que le sens d'une moyenne n'est clair que si l'on connaît aussi un indice de dispersion. Le plus fréquemment employé est l'écart-type, précédemment défini. Mais son emploi n'est légitime que si la variable considérée a une distribution normale (courbe de Laplace-Gauss, en forme de cloche), ce qui n'est certainement pas le cas d'une distribution de vocabulaire, puisqu'elle n'est pas symétrique. Toutes ces réflexions conduisent à penser que la moyenne peut donner une idée du texte, mais qu'on ne doit pas trop en espérer.

J'ai proposé il y a déjà longtemps (cf. Évrard 1967) de synthétiser un tableau de distribution du vocabulaire en utilisant une mesure empruntée à la thermodynamique, mais dont d'autres disciplines (par exemple la théorie de l'information) ont déjà fait usage. Il s'agit de l'entropie. L'entropie se définit comme l'inverse (d'où le signe "moins") de la somme (indiquée ici par S) des produits (un par lemme) formés par la multiplication de la fréquence p du lemme (le nombre d'occurrences de ce lemme divisé par le nombre d'occurrences du corpus) par le logarithme de cette fréquence. Soit

$$H = - S (p_i \times log\ p_i)$$

le S étant utilisé comme symbole de la sommation et p_i la fréquence du terme qui occupe le rang i dans la liste des lemmes.

Je rappelle que, dans un système de deux progressions, l'une arithmétique et l'autre géométrique, le logarithme d'un nombre de la première progression est son homologue dans la seconde ; comme les deux progressions ont un rythme très différent, deux écarts identiques, mais à des niveaux divers, dans la première correspondent à des écarts différents dans la seconde ; en voici un exemple : dans les logarithmes décimaux, le logarithme de 1 est égal à 0, celui de 10 est égal à 1, celui de 100 est égal à 2 (le logarithme étant l'exposant dont il faut affecter 10 pour obtenir le terme correspondant de la progression arithmétique, avec la convention que 10^0 est égal à 1). Dans ces conditions, on comprend qu'une différence dans la distribution, même si elle n'affecte pas les totaux des lemmes et des occurrences, modifie nécessairement l'entropie. A ce premier avantage s'en ajoute un autre. L'entropie d'un corpus verbal se situe entre deux limites facilement définissables et interprétables. Soit, par exemple, un texte de 100 occurrences (je prends des nombres ronds parce qu'ils sont plus parlants, mais cela ne modifie pas la démarche). Supposons que ces cent occurrences relèvent du même lemme (ce qui peut arriver si on conjugue un paradigme) ; ce mot unique a une fréquence de 100 / 100 = 1 et l'entropie sera

$$H = - (1 \times 0) = 0$$

Si un texte de même longueur est constitué de 100 lemmes apparaissant chacun 1 fois (par exemple si on lit les rubriques d'un dictionnaire), la fréquence de chacun de ces 100 lemmes est 1/100 = 0,01, avec pour logarithme -2, d'où

$$H = - 100 \times (0{,}01 \times -2) = 2$$

Généralisons : nous pouvons dire que l'entropie d'un texte varie entre 0 et le logarithme de la longueur de ce texte, ces deux extrêmes correspondant respectivement à une uniformité ou à une diversité maximales du vocabulaire. Soit le texte des 1 000 premiers mots des *Odes* d'Horace ; l'entropie en est H = 2,74, tandis que pour les 1 000 premières occurrences de la *Consolation à Polybe,* elle est H = 2,44. Pour cette longueur, le maximum est 3 ; on voit donc bien ce que révèlent ces nombres. En pratique, le calcul de l'entropie peut facilement être automatisé, si bien qu'il n'y a pas lieu de se laisser trop effrayer par les logarithmes.

Examinons maintenant la forme des distributions de fréquence. Les tableaux de fréquence, sans exception, semble-t-il, présentent une allure générale uniforme, avec, naturellement, des différences de détail : plus la fréquence est élevée, plus l'effectif est faible ; à part les inévitables petites dérogations, on constate qu'au fur et à mesure que les fréquences diminuent, les effectifs augmentent, et ce mouvement est surtout sensible quand on arrive à des fréquences relativement faibles (ainsi, dans la distribution des *Lettres à Lucilius*, il n'y a aucune dérogation à partir de la fréquence 12). Si l'on représente en un graphique un tel tableau, mettant en abscisse les fréquences et en ordonnée les effectifs, on obtient une courbe qui dessine plus ou moins franchement un L. La forme est du même type si on représente la relation rang-fréquence (le rang étant la position occupée par le lemme dans une liste de fréquence décroissante, le lemme le plus fréquent ayant le rang 1).

Certains, et surtout le psychologue Zipf, ont tenté de modéliser ce phénomène et d'en donner une expression mathématique rigoureuse. Zipf a soutenu que, si l'on donne aux lemmes un rang selon la fréquence décroissante, le produit du rang par la fréquence est une constante. Il pensait que le produit du carré d'une fréquence par le nombre de lemmes ayant cette fréquence est lui aussi une constante. P. Guiraud, partant de la première de ces constantes, pensait pouvoir exprimer une distribution de vocabulaire par une série harmonique qui, en graphique, est une hyperbole équilatère. Je n'insisterai pas sur ces constructions, parce que les constantes aperçues par Zipf sont plutôt de belles infidèles : l'exemple privilégié de Zipf est celui de James Joyce et de son *Ulysses*, où la constante peut varier de 20 000 à 29 000. Pour beaucoup d'autres exemples, les écarts sont de la même manière trop importants pour que l'on puisse accorder du crédit à la thèse de la constance. Je mentionnerai cependant les inférences psychologiques de Zipf parce que, débarrassées des excès mathématiques, elles expriment peut-être quelque chose de vrai. Pour Zipf, la communication verbale se réaliserait en fonction de la loi du moindre effort, considérée du point de vue du locuteur et de celui du destinataire. Le locuteur tente de se faire comprendre aux moindres frais, par exemple en multipliant les termes imprécis tels que "chose", "machin", etc. Mais le destinataire désire, lui, réduire son effort de compréhension et exige de recevoir des messages précis. Ce que traduirait une distribution de vocabulaire, ce serait un équilibre entre ces deux moindres efforts antagonistes. Dans la réalité, ces deux facteurs jouent sans doute un rôle, mais ils se combinent avec d'autres et ont une intensité variable, ce qui fait qu'on ne peut en attendre de représentation rigoureuse.

J'indique en passant un problème auquel un certain nombre de chercheurs se sont intéressés, mais pour lequel il n'y a pas de solution. Appelons vocabulaire d'un texte les lemmes qui y sont employés. Ce vocabulaire est un sous-ensemble de ce que l'on peut appeler le lexique de l'auteur, c'est-à-dire l'ensemble des mots qu'il a à sa disposition et qui pourraient donc entrer dans le vocabulaire d'un texte en cours d'élaboration. La distinction est intéressante, mais on ne voit pas de méthode

mathématique qui puisse, comme certains l'ont cru, établir l'importance numérique du lexique.

Abordons un autre problème, celui de la relation entre le vocabulaire total d'un texte et celui d'une partie de ce texte (mouvement inverse du mouvement habituel en statistique, qui, à partir d'un échantillon, tente d'évaluer de manière probable le tout). C'est un problème qu'il convient de résoudre avant toute affirmation concernant les caractères d'un texte relativement à son vocabulaire, parce que se pose toujours la question de l'homogénéité de l'ensemble. Une solution a été imaginée, indépendamment et à la même époque (1963-1964), par Ch. Muller 1964 et moi-même (Évrard 1964). Encore une fois, utilisons un exemple fictif composé de chiffres ronds. Soit un texte de 20 000 occurrences dont on connaît la distribution de vocabulaire. Le problème est d'évaluer ce que serait la distribution du vocabulaire d'une partie de ce texte, disons 5 000 occurrences, dans l'hypothèse où la répartition est aléatoire. Si, dans ce texte, on prend au hasard une occurrence, la probabilité p qu'elle appartienne à la partie prise en considération est de 0,25 et la probabilité q qu'elle y soit étrangère est de 0,75 ; on a, comme c'est normal, 0,25 + 0,75 = 1. Cette situation est celle de tous les lemmes employés une seule fois dans l'ensemble. Il suffit donc de multiplier ces deux probabilités par le nombre de lemmes de fréquence 1 pour obtenir deux effectifs théoriques ; supposons que ces lemmes soient au nombre de 1 000 ; les effectifs théoriques sont alors 250 dans la partie et 750 en dehors. Considérons maintenant les lemmes de fréquence 2. En raison de la loi des probabilités composées, rappelée précédemment, on peut représenter chacun d'entre eux par l'expression $(p + q)^2$ dont le développement est $(p + q)^2 = p^2 + 2pq + q^2$, chacun des trois termes correspondant à l'une des trois possibilités qui se présentent : les deux occurrences dans la partie, une occurrence dans la partie et l'autre en dehors, les deux occurrences hors de la partie. Dans le cas envisagé, on aurait

$$(0{,}25 + 0{,}75)^2 = 0{,}25^2 + 2 \times 0{,}25 \times 0{,}75 + 0{,}75^2 = 0{,}0625 + 0{,}375 + 0{,}5625.$$

De nouveau, en multipliant par l'effectif de la fréquence 2, on obtient des fréquences théoriques. Supposons que cet effectif soit de 500 ; on obtient comme résultat que 31,25 des lemmes devraient avoir leurs deux occurrences dans la partie, 187,5 une dans la partie et une en-dehors, et 281,25 leurs deux occurrences hors de la partie (les décimales ne sont évidemment pas à prendre au pied de la lettre ; elles expriment un ordre de grandeur). Pour chaque fréquence, on peut ainsi développer un binôme de Newton, avec, comme exposant, la fréquence totale considérée : si, par exemple, il y a x lemmes de fréquence 5, on écrira

$$x\,(p + q)^5 .$$

En fait, l'ordinateur peut se charger de cette mission fastidieuse, mais il convient que le chercheur sache ce que fait l'ordinateur. Pour chacune des fréquences globales, on a alors un nombre d'effectifs théoriques égal à la fréquence globale plus un. On totalise alors les termes en q^x, ce qui représente les lemmes absents de la partie. On totalise ensuite successivement tous les termes où p a le même exposant, ce qui donne l'effectif des lemmes ayant dans la partie la fréquence exprimée par cet exposant. Au terme, on dispose de la distribution théorique de la partie, que l'on peut comparer à la distribution réelle, de manière à vérifier qu'il y a ou non homogénéité.

On peut simplifier l'utilisation du test. On calcule uniquement les termes qui ne contiennent pas p, c'est-à-dire les derniers termes de chaque développement binomial, qui sont en fait les puissances successives de q. On multiplie chacune d'entre elles par

l'effectif de la fréquence correspondante et on totalise. Ce total correspond au nombre de lemmes présents dans la totalité mais absents de la partie ; en soustrayant, on obtient le vocabulaire théorique de chacune des parties que l'on compare au vocabulaire réellement observé. Comparons par exemple le *Moretum* aux deux *Bucoliques* non dialoguées les plus longues : la 6e compte 571 occurrences avec un vocabulaire de 367 lemmes ; le vocabulaire théorique des 571 premières occurrences du *Moretum*, calculé par la méthode que je viens de décrire, est de 403 (le vocabulaire réel étant de 407) ; la 10e *Bucolique*, pour sa part, a 530 occurrences et 316 lemmes ; la partie de même longueur du *Moretum* a un vocabulaire théorique de 380 lemmes (vocabulaire réel 377). Le sens de ces résultats apparaît de manière plus évidente si on les met en tableau : le *Moretum* n'est nullement homogène par rapport aux deux *Bucoliques*.

	Mor.,1-88	6e *Buc.*	*Mor.*,1-82	10e *Buc.*
Nbre d'occ.	571	571	530	530
Voc. réel	407	367	377	316
Voc. calculé	403	–	380	–

Pour caractériser l'enrichissement (ou diversification) d'un vocabulaire, divers chercheurs ont proposé des formules intéressantes étudiées par Cossette[18]. Je me limiterai à celle de Ch. Muller, dont j'ai déjà parlé et qui se fonde sur la loi binomiale, et à celle d'Étienne Brunet, qui s'exprime dans la formule $W = N^{1/v \exp.a}$, où a est un paramètre qu'il y a lieu d'estimer. Il me semble qu'on peut formuler une remarque d'ensemble. Tous les chercheurs étudiés par Cossette ont le souci d'éliminer l'influence du facteur "longueur du texte". Cela peut se comprendre mais me paraît propre à dénaturer les données. La longueur du texte (telle qu'elle est plus ou moins prévue au moment où on se met à l'écrire) règle pour une bonne part la manière dont celui qui écrit ou qui parle manœuvre son vocabulaire (ou lexique) : on ne réagit pas de la même manière quand on veut écrire un sonnet ou un long poème épique, une brève note de recherche ou une étude compacte sur un ensemble important de documents. Il m'a donc paru intéressant de définir des paramètres qui tiennent compte de cette action[19]. Je suis parti du calcul de l'entropie d'une distribution, tel qu'il a été exposé antérieurement[20]. Rien n'empêche de faire ce calcul non seulement pour la totalité d'une œuvre, mais aussi pour des parties de plus en plus longues : les 500 premiers mots, les 1000 premiers mots, etc. On se souvient qu'à chaque fois, le maximum possible est le logarithme de la longueur ; on a ainsi en graphique un angle dont le sommet est à l'origine, l'un des côtés se confondant avec l'abscisse et l'autre formant un angle avec cette abscisse. Dans cet espace s'inscrivent les valeurs réelles. Les points ainsi fixés dessinent une courbe qui ressemble fort à une parabole. De fait, dans beaucoup de cas, le calcul, à partir des données, d'une parabole théorique (selon la méthode des moindres carrés) donne un degré d'approximation tout à fait satisfaisant. La courbe parabolique s'exprime par une fonction du second degré. Dans le cas du vocabulaire, on sait que, de toute manière, la courbe commence à 0, puisque le premier mot donne une entropie égale à zéro. L'expression sera donc $y = ax^2 + bx$, où y est l'entropie du texte et x le logarithme de sa longueur. Les coefficients a et b se calculent pour chaque œuvre : a est toujours négatif et b toujours positif. Au vu de l'action de ces deux coefficients dans le

18. Cossette 1994 (cf. mon compte rendu dans *Revue, Informatique et Statistique dans les Sciences Humaines* 30, 1994 : 271-274).

19 . Cf. Évrard 1985.

20. Cf. ci-dessus, début du § 4.

calcul de *y*, on s'aperçoit que *b* provoque une élévation progressive (qui serait régulière s'il n'y avait *a*) par rapport à l'axe d'abscisse ; on y verra donc l'expression de la tendance de l'auteur à diversifier son vocabulaire. Le coefficient *a*, en revanche, en raison du fait qu'il est négatif, entrave cette montée et tend à rapprocher la courbe du niveau de l'abscisse ; de plus, par le fait qu'il est multiplié par un carré (x^2), son influence va croissant ; il est donc légitime d'y voir un "indice d'essoufflement", qui mesure le fait que l'auteur a de plus en plus de difficulté à renouveler son vocabulaire. Ces deux facteurs en sens opposé agissent plus ou moins selon les auteurs, si bien que les deux coefficients apparaissent comme un moyen de caractériser un auteur quant à ses manœuvres sur le vocabulaire. Le tableau suivant met en parallèle les entropies observées et les entropies calculées par cette méthode pour le *Moretum*. La première colonne indique la limite supérieure des parties, la seconde, les entropies maximales pour la longueur correspondante, la troisième, les entropies observées, la quatrième, les entropies calculées. On constate que la série observée et la série calculée sont très proches l'une de l'autre, ce qui est un indice de l'excellence, dans ce cas-ci, de l'approximation parabolique.

Limite supérieure	Entropie maximale	Entropie réelle	Entropie calculée
104	2,017003	1,95914	1,95721
209	2,32015	2,18491	2,18678
309	2,49000	2,31127	2,30802
412	2,61490	2,39225	2,39384
515	2,71181	2,44521	2,45844
615	2,78888	2,51301	2,50858
717	2,85552	2,56168	2,55106
804	2,90526	2,57868	2,58223

Voici maintenant un tableau donnant les valeurs de *a* et de *b* (calculées selon la méthode décrite) pour quelques œuvres latines. Elles sont classées sur la valeur croissante de *b*. On observe que les valeurs absolues de *a* sont aussi croissantes, sauf dans quelques cas intéressants, ceux d'Horace, du *Moretum* et de Virgile. Par ailleurs, les deux extrémités de la liste sont occupées par des œuvres de Cicéron, ce qui montre combien la notion d'auteur est un critère peu efficace.

Texte	Valeur de *b*	Valeur de *a*
Cic., *N.D.*, III	1,01	- 0,08
Cés., *B.G.*	1,01	- 0,09
Sén., *Ir.*	1,10	- 0,10
Hor., *C.*, I	1,15	- 0,08
Sén., *Marc.*	1,15	- 0,11
Moretum	1,16	- 0,09
Sén., *Helu.*	1,17	- 0,11
Sall., *Cat.*	1,17	- 0,11
Sén., *Ben.*	1,17	- 0,12
Sén., *Pol.*	1,21	- 0,13
Sall., *Jug.*	1,21	- 0,13
Virg., *En.*, I-VI	1,24	- 0,12
Cic., *Caec.*	1,27	- 0,16

J'ajoute qu'un philologue classique liégeois a appliqué ce test aux tragédies de Sophocle et qu'il a eu la surprise de voir apparaître ainsi un classement correspondant à la succession chronologique probable de ces œuvres[21].

J'insiste sur le fait que les tests décrits font l'objet de logiciels et que le chercheur ne doit donc se livrer à aucun calcul fastidieux. Il importe simplement qu'il comprenne ce qui se passe et comment les données sont traitées, de manière à choisir à bon escient les tests qu'il utilisera dans telle ou telle recherche. A quoi il faut ajouter que certains types de recherche ne s'accommodent pas du tout de méthodes quantitatives ; il ne doit y avoir, de la part de ceux qui utilisent ces dernières, nul absolutisme, ni pour eux-mêmes ni pour les autres chercheurs. Mais, réciproquement, il serait regrettable de rejeter ces méthodes dans ce qu'elles ont d'utile.

5. STATISTIQUE SYNTAXIQUE (Sylvie Mellet)

Comme on l'a signalé en introduction, il est rare que l'on fasse appel à la statistique dans les études de syntaxe[22]. Diverses raisons peuvent expliquer cet état de fait : conjoncturelles tout d'abord, le dénombrement automatique des 'individus' morpho-syntaxiques étant plus difficile à réaliser que celui des 'individus' lexicaux, surtout lorsque les banques de données textuelles ne sont pas lemmatisées - ce qui a longtemps été le cas, en français notamment ; conceptuelles ensuite : est-il raisonnable de traiter comme des unités abstraites et interchangeables des éléments dont la valeur linguistique est précisément liée à leur insertion syntagmatique et phrastique ou même interphrastique ?

Pourtant, on a pu expérimenter que le rapprochement des deux disciplines était fructueux : là encore la statistique peut suggérer des pistes de recherche, mettre en évidence des rapprochements intéressants ou rectifier des perceptions erronées. On en rappellera d'abord rapidement quelques exemples, puis on approfondira l'un d'eux pour présenter l'un des outils statistiques majeurs pour les études textuelles, à savoir l'analyse factorielle. En effet, ce type d'analyse, dite multidimensionnelle, permet de prendre en compte simultanément et selon leurs poids respectifs, les divers paramètres qui pèsent sur le texte ou sur le phénomène étudié.

5.1. Les différents types de problèmes

Prenons une question typiquement syntaxique : la distribution des temps et des modes verbaux dans un type de subordonnées en fonction de paramètres qui semblent devoir être linguistiquement intéressants. Il s'agit d'étudier une population spécifique, par exemple celle des subordonnées exprimant une relation causale, et d'évaluer si la répartition des principaux temps et modes verbaux est neutre par rapport au choix de la conjonction introductive. Le principe de cette évaluation est toujours le même : on s'appuie sur l'hypothèse d'une distribution conforme à la loi normale et on calcule l'espérance mathématique de la variable étudiée - en l'occurrence de telle ou telle forme verbale - dans les différentes propositions ; on compare ensuite ces effectifs théoriques calculés avec les effectifs réellement observés et on recourt à divers calculs statistiques pour mesurer la probabilité des écarts constatés (χ^2 ou test de Pearson) et/ou leur intensité (coefficient de Bernoulli ou coefficient de Yule).

---

21. Cf. Rigo 1994.

22. Citons néanmoins l'étude de R. Martin et Ch. Muller 1964.

Une étude des propositions causales introduites par *quod* et *quia* nous a laissé pressentir que, loin d'être synonymes comme on le prétend généralement, ces deux conjonctions manifestent deux attitudes énonciatives différentes : la prise en charge, par l'énonciateur, de la cause énoncée ne serait pas la même selon qu'il recourt à *quod* ou à *quia*. On a donc supposé que la distribution des modes verbaux après l'une et l'autre conjonctions pourrait apporter des indices complémentaires pour établir une telle différence. Le calcul est alors le suivant : à partir des effectifs globaux d'indicatifs et de subjonctifs relevés dans le corpus, une simple règle de trois permet de calculer les effectifs théoriques que l'on aurait pu attendre (par exemple pour les indicatifs après *quod* : (134 x 62) / 182 = 45,6). On obtient les deux tableaux récapitulatifs ci-dessous :

Effectifs observés

Annales, XI-XVI	Indicatif	Subjonctif	Total
Quod	26	36	62
Quia	108	12	120
Total	134	48	182

Effectifs calculés

Annales, XI-XVI	Indicatif	Subjonctif	Total
Quod	45,6	16,4	62
Quia	88,4	31,6	120
Total	134	48	182

Le calcul du χ^2 se fait selon la formule $\chi^2 = S\ (o - c)^2 / c$; il est ici égal à 48,33 et montre que de tels écarts ont largement moins d'une chance sur mille d'être aléatoires. La probabilité d'une répartition des modes conditionnée par la conjonction qui introduit la subordonnée est donc très grande. L'hypothèse linguistique se voit confirmée et la recherche peut être poursuivie (et, de fait, les mêmes tests pratiqués sur l'œuvre de Quinte-Curce donnent des résultats tout à fait comparables).

Prenons un autre exemple, traité lui aussi avec le test de Pearson. Il s'agit cette fois d'étoffer l'analyse des divers emplois de *dum*. Une lecture attentive des textes suggère que la place de la proposition temporelle dans la phrase varie selon le mode qui y est employé et tend par ailleurs à se stéréotyper : antéposition pour *dum* + indicatif et postposition pour *dum* + subjonctif ; une hypothèse linguistique surgit alors, prenant appui sur des analyses bien connues de l'ordre des syntagmes dans la phrase latine : les propositions à l'indicatif seraient thématiques tandis que celles où *dum* est suivi du subjonctif seraient rhématiques. Il convient toutefois de vérifier la corrélation perçue plus ou moins intuitivement ; le calcul du χ^2 est de nouveau tout à fait adapté, en choisissant pour variable la place de la subordonnée dans la phrase et pour paramètre susceptible d'influencer la distribution le mode de la subordonnée :

Effectifs observés

Tite-Live	*Dum* + Indicatif	*Dum* + Subjonctif	Total
Antéposition	83	6	89
Autres positions	29	51	80
Total	112	57	169

Effectifs calculés

Tite-Live	*Dum* + Indicatif	*Dum* + Subjonctif	Total
Antéposition	59	30	89
Autres positions	53	27	80
Total	112	57	169

Ici $\chi^2 = 61{,}15$ ce qui équivaut à une probabilité quasi nulle que la répartition observée soit l'effet du hasard.

Un autre exemple encore pourrait être donné par la distribution des corrélatifs *ita* et *sic* selon que la proposition qui suit, introduite par *ut*, est de sens comparatif ou consécutif: le test de Pearson montre que *ita* est largement préféré en contexte consécutif et *sic* en contexte comparatif.

5.2. Les analyses factorielles des correspondances

Nous allons maintenant aborder un type d'analyses un peu plus complexes, mais qui offrent des atouts supplémentaires importants par rapport aux outils présentés jusqu'à maintenant:

1) ces analyses permettent de prendre en compte simultanément plusieurs variables et de mettre ainsi en évidence les différents facteurs qui interviennent dans la structuration des données;

2) le résultat de l'analyse est donné sous forme graphique où les associations statistiques sont représentées par des proximités géométriques dans un espace à deux dimensions: la lecture en est donc aisée, plus n'est besoin ici de faire appel à des tables de probabilités.

Imaginons qu'à la lecture des *Catilinaires* de Cicéron il nous semble y avoir une correspondance assez fréquente entre l'emploi de telle personne verbale et l'emploi de telle forme modo-temporelle: par exemple, on remarque que les huit futurs du *perfectum* employés dans le livre IV sont tous à la deuxième personne du pluriel; qu'en revanche le futur de l'*infectum* est souvent associé à la première personne du singulier. Comment contrôler la portée de ces remarques et pousser plus loin l'analyse?

Bien sûr, on pourrait encore recourir au test de Pearson; et là deux méthodes s'offrent à nous. Ou bien l'on dresse un grand tableau récapitulatif composé de six colonnes pour les six personnes et de dix lignes pour les six temps de l'indicatif et les quatre du subjonctif, avec un calcul du χ^2 qui montrera sans doute que, globalement, la distribution n'a rien d'aléatoire; mais l'interprétation précise de ce tableau et la perception des correspondances particulières entre telle personne et tel temps seront malaisées.

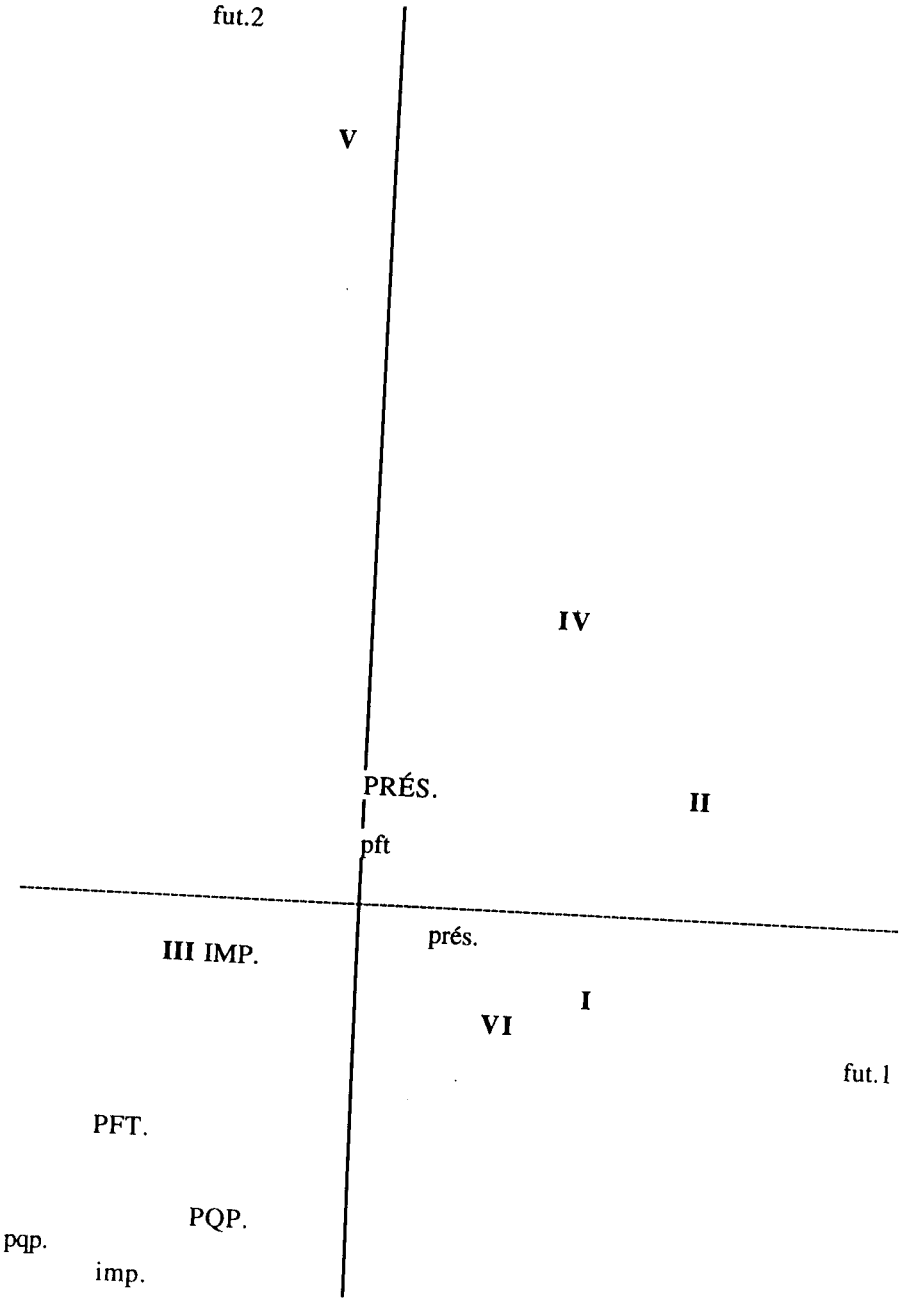
	I	II	III	IV	V	VI	Total
Ind. prés.	109	60	240	25	26	166	626
Ind. impft.	12	3	38	0	0	14	67
Futur 1	38	21	34	6	2	27	128
Ind. parf.	71	32	166	18	20	92	399
Ind. pqp.	2	1	19	0	0	6	28
Futur 2	4	4	22	3	8	2	43
Subj. prés.	31	23	92	9	13	57	225
Subj. impft	25	6	68	5	6	40	150
Subj. parf.	6	7	31	0	0	5	49
Subj. pqp.	10	3	35	2	0	19	69
Total	308	160	745	68	75	428	1784

Ou bien l'on établit une série de tableaux partiels, isolant par exemple chaque personne verbale successivement; mais c'est alors l'interprétation globale des relations de temps et de personnes qui deviendra difficile ainsi que l'évaluation du poids respectif de chaque association.

	1ère personne	Autres personnes	Total
Futur 1	38	90	128
Autres formes	270	1386	1656
Total	308	1476	1784

C'est dans pareil cas que l'analyse factorielle des correspondances est particulièrement utile: elle va permettre de donner une description facilement interprétable de ce grand tableau croisé, en visualisant les proximités entre les lignes et les colonnes. On observe en effet que certaines colonnes ont des profils similaires; par exemple la colonne de la personne IV et celle de la personne V ne se différencient vraiment que par leurs effectifs de futurs. De même les lignes peuvent offrir des séries numériques assez comparables, par exemple la ligne du subjonctif parfait et celle du subjonctif plus-que-parfait ou aussi celle de l'indicatif plus-que-parfait.

On extrait par une méthode mathématique les profils caractéristiques du tableau; les distances de chaque ligne et chaque colonne à ces profils vont être calculées, puis représentées sur un plan. Voici la première projection géométrique construite à partir du tableau à 6 colones et 10 lignes ci-dessus (les temps de l'indicatif sont indiqués en minuscules, ceux du subjonctif en majuscules):



Les principes de lecture de la figure sont les suivants :

- un angle au centre faible entre deux points figurant deux lignes ou deux colonnes indique une forte ressemblance ou proximité entre celles-ci (par exemple ici on retrouve la proximité déjà évoquée entre subjonctif et indicatif plus-que-parfaits ;
- au contraire un angle de 180° manifeste une opposition ou distance maximale entre deux éléments (c'est le cas ici entre les personnes II et III ou V et VI) ;

- l'éloignement des points par rapport au centre souligne l'intensité de la conjonction ou de l'opposition (par exemple, on remarquera ici l'intensité de l'opposition entre le futur de l'*infectum* et le futur du *perfectum*) ; en revanche des points proches de l'origine des axes figurent des éléments peu sensibles aux facteurs mis en évidence par l'analyse, des formes neutres en quelque sorte ;

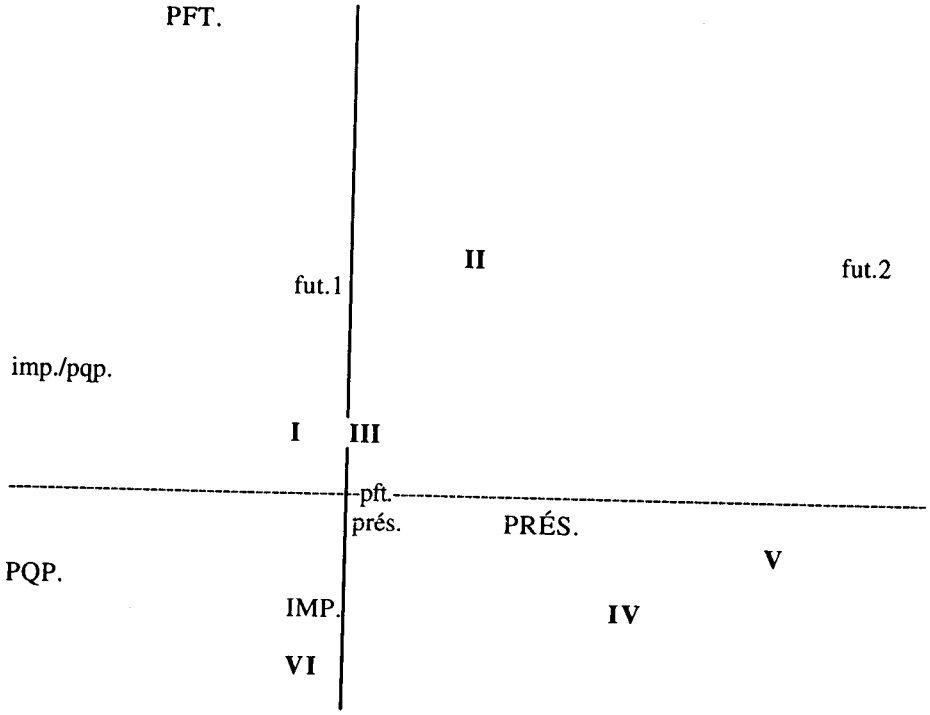
- en principe, il n'est pas licite d'interpréter une proximité ou une opposition entre un point représentant une ligne du tableau (temps verbaux) et un point représentant une colonne (personne verbale) ; la seule possibilité consiste à repérer un point-ligne par rapport à l'ensemble des points-colonnes ou vice-versa. Dans la pratique néanmoins, situer la personne V par rapport à l'ensemble des formes modo-temporelles revient bien à souligner sa proximité avec le futur du *perfectum* ; le retour au texte et les possibilités d'interprétation linguistique d'une telle conjonction semblent justifier une telle licence par rapport à la théorie. Il convient cependant de rester prudent et conscient que l'on s'aventure alors sur un terrain quelque peu glissant.

La méthode - mise en œuvre, bien évidemment par un logiciel parfaitement rodé - fait appel à la décomposition du tableau matriciel initial en une série de tableaux qui ont des propriétés particulières décrivant la complexité progressive de la matrice. Le nombre de ces tableaux partiels est égal à la plus petite dimension du tableau initial moins une unité (ici, par exemple, il y a 6 colonnes et 10 lignes ; la décomposition se fera donc en cinq étapes).

La première étape met en évidence les proximités les plus frappantes à l'intérieur du tableau initial ; c'est donc une approximation grossière qui fait ressortir le principal facteur d'analyse des phénomènes observés : ce facteur est porté sur l'axe des abscisses du premier plan d'analyse. Ici, il s'agit de l'opposition très nette entre les personnes I, II et IV qui se trouvent du côté du présent / futur et la personne III qui se trouve du côté du passé ; c'est en quelque sorte une opposition benvenistienne. On remarque en revanche que les personnes V et VI perturbent quelque peu cette répartition ; cependant l'interprétation doit rester en suspend dans la mesure où les points qui les figurent se trouvent très proches de l'axe central.

On devine maintenant que l'axe des ordonnées figure le second facteur d'analyse, susceptible d'expliquer les proximités mises en évidence au niveau du deuxième tableau partiel dans la décomposition de la matrice initiale. Dans l'exemple choisi cet axe permet d'opposer très nettement le comportement des deux futurs latins et de mettre en évidence la proximité de la personne V avec le futur du *perfectum*. On observe aussi que les personnes I et IV se situent de part et d'autre de l'origine de cet axe, ce qui ne laisse pas de surprendre quand on pense que la plupart des premières personnes du pluriel semblent être des formes emphatiques (ou de pseudo-modestie) référant à l'*ego* du locuteur.

L'envie surgit alors de pousser l'analyse un peu plus loin. Cela est tout à fait possible : dans notre cas de figure, il y a cinq facteurs d'analyse ; on peut donc construire un deuxième plan à partir des facteurs 2 et 3. Les principes de lecture restent les mêmes ; on se souviendra seulement que le facteur 3 a une part explicative moindre dans l'analyse des phénomènes étudiés. Il permet aussi d'aller au-delà des évidences que les premiers facteurs se contentent parfois de relever. Voici le résultat de cette deuxième projection graphique :



L'opposition entre personnes I et IV est largement confirmée. On note qu'elle se superpose, au niveau des temps verbaux de l'indicatif, à une opposition bien connue, celle du parfait d'un côté, de l'imparfait et du plus-que-parfait de l'autre. Reste alors au linguiste à prendre le relais, à formuler les hypothèses explicatives et à reprendre l'étude du texte pour les confirmer[23]. La statistique ne lui en dira pas plus ; mais elle a déjà beaucoup suggéré, elle a surtout fourni les éléments objectifs et précisément quantifiés qui viendront étayer solidement son argumentation.

6. STATISTIQUE MÉTRIQUE (Étienne Évrard)

La métrique se prête bien à des recherches quantitatives, surtout quand il s'agit d'une métrique qui est elle-même quantitative, comme c'est le cas pour le latin et pour le grec. Il faut cependant noter que les possibilités varient fort selon les cas. Les mètres qui n'admettent ni substitution ni remplacement ne comportent qu'une modalité, ce qui interdit toute comparaison entre effectifs de modalités multiples d'un même mètre ; on pourrait tout au plus y étudier par des techniques quantitatives la place des intermots, ou encore la coïncidence ictus / accent verbal, en n'oubliant toutefois pas que la nature et même l'emplacement des ictus ne sont pas toujours connus avec certitude. Il en résulte que la plupart des recherches qui ont été faites concernent l'hexamètre dactylique et, dans une moindre mesure, le distique élégiaque.

Préalablement à tout calcul, il faut disposer de fichiers présentant les données métriques des œuvres à étudier. On sait que la scansion des vers anciens, sauf pour

23. Cf. Mellet 1994.

certaines catégories assez restreintes, ne présente pas de difficulté, pour peu qu'on en ait l'habitude. Mais une chose est de scander tout en lisant, c'en est une autre d'enregistrer la scansion dans tous ses détails. Des pionniers tels Drobisch et Hultgren, cités en introduction, ont eu le courage admirable de compter manuellement les effectifs des divers types d'hexamètre dans un grand nombre d'œuvres. Mais ils n'ont guère eu d'émules.

L'apparition de l'ordinateur a suscité l'idée d'informatiser la scansion, la mise en mémoire de ses résultats et leur élaboration sous forme de listes, d'index, de comptages et de tableaux. Les latinistes spécialistes de Virgile connaissent bien les index de W. Ott. Ce chercheur a mis au point, en 1966, un logiciel qui scande les hexamètres et fait toutes les sélections et tous les comptages utiles; on dispose ainsi des références relatives à chaque type d'hexamètre, de leur dénombrement, d'un index des formes rangées selon leur prosodie et leur emplacement métrique, etc. Il faut toutefois noter que, dans le système de W. Ott, le texte poétique, avant d'être saisi en mémoire, fait l'objet d'une préédition qui pourvoit chaque vers de six signes (0 ou 1) indiquant que le pied correspondant est un spondée ou un dactyle, et y ajoute des sigles relatifs, entre autres choses, aux synizèses. Au même moment, Nathan Greenberg avait conçu un logiciel différent, mais qui, lui aussi, grâce à une légère préédition, scande les hexamètres latins et en mémorise les caractères.

Pour ma part, j'ai fait un travail analogue pour l'hexamètre grec[24]. Le système que j'ai adopté ne demande aucune préédition; à partir des règles contenues dans le programme, l'ordinateur propose une scansion ou, éventuellement, deux, voire trois, le philologue étant alors en mesure de choisir celle qu'il tient pour correcte. Le logiciel que j'ai écrit, une fois en possession des scansions épurées, constitue un fichier des formes rangées selon leur prosodie, puis selon leur position métrique dans le vers; en outre, il fait toute une série de dénombrements. Mais ce n'est là que la phase préparatoire, et toutes les informations ainsi rassemblées n'ont de raison d'être que pour servir à des recherches interprétatives spécifiques.

C'est ce qu'avaient déjà compris Drobisch et Hultgren. Ils soumirent les dénombrements qu'ils avaient réalisés à des traitements méritoires, mais qui souffrent d'une double faiblesse: d'une part, l'outil statistique, à leur époque, était encore assez élémentaire: par exemple, le test de χ^2, dont l'utilisation est devenue toute banale, n'était pas encore connu; d'autre part, ils paraissent surtout intéressés par la recherche de lois générales (la loi de l'hexamètre ou la loi du distique) et ont tendance à ne pas suffisamment exploiter les différences entre auteurs, que leurs calculs mettent en évidence.

Il arrive aussi que, même avec un outil statistique tout à fait performant, certaines recherches s'égarent parce que les données à traiter sont, du point de vue de la langue, analysées insuffisamment ou de manière erronée. On en trouve un exemple dans les *Hexameter Studies* éditées par R. Grotjahn, qui contiennent six études de métrique statistique[25]. L'une d'entre elles va m'aider à mettre en garde contre certains dangers; je le fais, non par goût de la critique, mais parce que le risque est grand que le chercheur

24. Voir Évrard 1972.

25. Grotjahn R. 1981, avec la note de chronique que j'y ai consacrée dans *Revue, Informatique et Statistique dans les Sciences Humaines* 29, 1983: 226-243.

trop attentif à la technique statistique, néglige ou méconnaisse certains aspects littéraires ou linguistiques de l'objet de sa recherche.

C'est ce qui se passe dans l'étude de R. Schmiel 1981. Ce dernier reprend l'idée de W. F. J. Knight 1939 (et Knight 1944) sur la distinction entre vers homodynes et hétérodynes selon que l'ictus du 4e pied coïncide ou non avec un accent verbal. Il veut vérifier les effets de cette distinction dans l'épopée grecque et utilise comme données des relevés relatifs au nombre de syllabes accentuées d'un aigu ou d'un circonflexe et au nombre d'ictus accentués, le nombre d'ictus par vers étant invariablement 6. Ni Schmiel, ni Grotjahn dans sa révision de ce travail, n'ont tenu compte du fait qu'il y a des syllabes accentuées de l'aigu qui sont brèves et, de ce fait, ne peuvent être à l'ictus. Cette bévue est d'autant plus regrettable que la méthode statistique mise en œuvre est, quant à elle, particulièrement bien conçue. Au reste, je ne suis pas sûr que le problème des homodynes ait intérêt à être traité par la statistique. L'hypothèse de Schmiel est que les vers homodynes sont paisibles et relèvent plutôt du discours, tandis que les hétérodynes sont généralement animés et appartiennent aux passages narratifs. On voit tout de suite le caractère subjectif des critères et la possibilité d'une autocorrélation : le lecteur ayant en tête la théorie de Schmiel risque de sentir un vers comme animé pour la seule raison qu'il en constate l'hétérodynie.

Il convient, en revanche, de signaler que, si l'analyse quantitative ne s'impose pas dans tous les domaines, il en est où elle permet d'éviter des conclusions aventureuses fondées sur des impressions insuffisamment critiquées. J'en proposerai deux exemples. W. R. Nethercut 1979, étudiant l'épisode de la poursuite de Daphné par Apollon dans les *Métamorphoses* d'Ovide et y appliquant, lui aussi, la distinction de W. F. J. Knight entre hexamètres homodynes et hétérodynes, croit qu'elle aide le poète à accentuer l'expressivité de son récit. En fait, le calcul montre que la proportion des deux espèces d'hexamètres n'est pas significativement différente dans le récit considéré de ce qu'elle est dans la totalité du livre I des *Métamorphoses*, dans lequel se trouve ce récit. Quelle qu'ait pu être l'intention du poète, il faut donc avouer que l'opposition homodynes / hétérodynes n'a pas contribué à particulariser l'expressivité de ce passage. On trouve un cas du même genre dans l'étude du prologue de l'*Enéide* de W. D. Lebek 1982. L'auteur croit y déceler des recherches d'expressivité fondées sur les effectifs des divers types d'hexamètres et sur les fréquences des dactyles et des spondées. Ici encore, si l'on compare ces effectifs et ces fréquences à celles de l'intégralité du livre I de l'*Enéide*, le calcul ne met en évidence aucune différence significative. Dans les deux cas, il semble donc que les interprétations proposées n'ont pas de fondement dans les textes et qu'un recours à l'analyse quantitative aurait incité à une prudence bien nécessaire.

Dans les mètres qui pratiquent des remplacements et des substitutions, un vaste champ d'investigations se présente, puisqu'il est possible d'énumérer les diverses formes que ces mètres peuvent prendre et de dénombrer les effectifs de chacune d'elles. C'est surtout à l'hexamètre et au distique élégiaque que l'on pense alors, d'autant que tous deux sont parmi les plus employés tant en grec qu'en latin.

Je voudrais au préalable rappeler un avatar de l'hexamètre qui réduit à rien l'avantage que je viens d'indiquer. Le Moyen Âge, qui continuait à pratiquer la métrique quantitative à côté d'autres formes de son invention (métrique accentuelle et métrique syllabique), y a apporté certaines innovations. L'une d'entre elles est à l'origine de longs poèmes hexamétriques du XIIe siècle, tel le *De contemptu mundi* de Bernard Morlanensis. Les hexamètres y sont constitués uniquement de dactyles. Ces dactyles sont groupés par deux (partageant ainsi le vers en trois portions égales), de telle manière que les deux premiers groupes d'un vers riment entre eux et que le dernier rime avec le

dernier groupe du vers voisin (précédent ou suivant selon les cas). En voici un exemple (vers 951 et 952 ; en romain, les rimes) :

*Nunc ubi Reg*ulus *aut ubi Rom*ulus *aut ubi R*emus

*Stat rosa prist*ina *nomine, nom*ina *nuda ten*emus.

(On a reconnu dans le deuxième vers celui qu'Umberto Eco a mis à la fin de son roman *Le nom de la rose*.) La déviation que je viens de décrire est peut-être due à une interprétation trop littérale de la désignation *hexamètre dactylique* : la substitution d'un spondée à un dactyle y est sans doute apparue comme une licence qu'en puriste, on tenait à ignorer. De toute manière, elle interdit la possibilité de diversification qui est l'une des richesses de l'hexamètre.

Le jeu des remplacements et substitutions autorise 16 formes d'hexamètre (de 13 à 17 syllabes) et même 32 si l'on tient compte des hexamètres spondaïques. Comme le pentamètre, pour sa part, peut prendre 4 formes, le distique élégiaque se présente sous 64 formes (ou 128 dans les mêmes conditions). On sait qu'en latin, l'hexamètre spondaïque est rare (un peu plus de deux cents exemples d'après A. Viertel 1862 ; chiffre à peine modifié par J. Soubiran 1969) ; il n'en sera pratiquement plus question dans la suite de cet exposé.

L'une des idées les plus courantes sur les capacités expressives liées à cette multiplicité de formes est que l'accumulation de spondées donne une impression de solennité, de grandeur, mais aussi de lourdeur, tandis que l'accumulation de dactyles crée un sentiment de légèreté, de rapidité. On connaît certains exemples où ces effets soulignent de manière évidente le contenu du texte. Quand Virgile, dans la première *Bucolique*, écrit au vers 19 :

Urbem quam dicunt Romam Meliboee putaui

la succession des trois spondées au début du vers donne une impression de solennité, correspondant à la situation dans laquelle ces mots sont dits. Ce caractère expressif est encore souligné par le contraste que produit, trois vers plus loin, l'apparition d'une accumulation de dactyles dans un contexte auquel conviennent particulièrement bien les idées de légèreté et de mouvement :

Sic canibus catulos similes sic matribus haedos

Le fragment d'Ennius constitué des vers 138 et 139 de Vahlen est remarquable lui aussi de ce point de vue ; en outre, son premier vers exprime par des mots employés au sens propre ce que le second reprend en une métaphore, que l'on peut ainsi goûter tout à loisir puisque la clé en a été préalablement fournie :

Vulturus in spinis miserum mandebat homonem
Heu quam crudeli condebat membra sepulcro

Les quatre spondées du deuxième vers en expriment bien le caractère lugubre. Mais il faut avouer que les cas où une correspondance expressive d'une telle précision est possible ne sont pas nombreux et l'on peut craindre que des interprétations abusives n'en multiplient indûment le nombre. Il n'empêche cependant que, même sans parallélisme expressif aussi strict, l'accumulation soit de spondées soit de dactyles donne à un poème une coloration rythmique spécifique qui, pourvu que le sens du texte s'y prête, a chance d'être remarquée par un auditeur sensible et attentif et d'être

interprétée comme un effet de style ou d'expressivité. Une comparaison quantitative entre différents poètes, ou entre plusieurs œuvres d'un même auteur, ou entre les parties d'un même œuvre permet de déceler soit des parentés soit des diversités significatives.

Il me semble, par ailleurs, que la diversité rythmique peut avoir pour rôle de révéler ou souligner la structure d'une œuvre. La musique offre ici un point de comparaison utile. Soit la fugue en ut mineur, qui est la deuxième du second cahier du *Clavecin bien tempéré* de J. S. Bach. Le sujet en est exprimé en un rythme dont la valeur prédominante est la croche. Dans toute la première partie, qui est à deux puis, très vite, à trois voix, c'est cette valeur qui domine, avec un jeu sur le sujet, ses transpositions et son renversement. La deuxième partie, en revanche, expose d'abord le sujet en valeurs doublées (donc en noires), à la voix médiane avec, aux deux autres voix, en un foisonnant contrepoint, le sujet en croches, soit dans sa forme première, soit en renversement avec, de l'une à l'autre, des chevauchements. Ensuite, une quatrième voix, qui s'introduit à la basse, énonce à nouveau le sujet en valeurs doublées, avec, aux trois voix supérieures, un contrepoint qui ne doit à peu près rien au sujet mais utilise des rythmes en syncope. Ainsi se créent trois rythmes qui devraient être perçus à l'audition; ils marquent les articulations du discours musical.

Un poète écrivant en hexamètres peut, lui aussi (mais il ne le fait pas nécessairement) marquer les articulations de son texte par des épisodes rythmiques marquants. On notera toutefois qu'en musique, c'est la variation rythmique elle-même qui crée les articulations, tandis que dans un texte littéraire, la charpente est fournie par les articulations du récit ou par la succession des thèmes, etc., la variation rythmique pouvant seulement souligner éventuellement les passages d'une partie à l'autre. C'est ce qu'il me semble que l'analyse quantitative peut mettre en lumière: vérifier dans quelle mesure l'auteur emploie la variation des types hexamétriques pour structurer son texte, dans quelle mesure il paraît indifférent à cette ressource.

Mais, au préalable, il faut disposer d'une technique qui permette d'évaluer les effectifs résultant des dépouillements. Celle que je propose consiste d'abord à dresser (cela peut se faire automatiquement) un tableau des effectifs des divers types d'hexamètre. On y observe immédiatement une grande inégalité de type à type, et cette inégalité se manifeste - de manière éventuellement très différente - d'auteur à auteur. Je donnerai en exemple un tableau relatif à trois poètes médiévaux peu fréquentés: Vynfreth (680-755) alias saint Boniface, l'apôtre de la Germanie qui, à côté de son œuvre évangélisatrice, avait aussi enseigné les belles lettres, écrit un très bref traité de métrique et composé un recueil intitulé *Enigmata*, comprenant 388 hexamètres; Alcuin (725-804), qui composa une *Vie de saint Willibrord* en deux livres, dont le second est en vers et se compose de 353 hexamètres; enfin Abbon de Saint-Germain (850-923), auteur d'un récit en vers du siège de Paris par les Normands en 885-886, qu'on intitule *Bella Parisiacae Vrbis* et dont le premier livre compte 660 hexamètres, dont un spondaïque. La première colonne du tableau donne la composition des 16 types, rangés selon leur rime rythmique: d'abord les huit types qui ont un spondée au quatrième pied, puis les huit qui ont en cette position un dactyle; dans chacun de ces deux groupes, d'abord les quatre types qui ont un spondée au troisième pied, puis les quatre qui y ont un dactyle, et ainsi de suite. Les trois colonnes suivantes indiquent, pour chacun des trois auteurs, les effectifs réellement observés. Pour Abbon, j'ai seulement laissé tomber l'unique spondaïque, d'où le total de 659.

Schémas	Boniface	Alcuin	Abbon
ssss	48	11	7
dsss	92	55	58
sdss	37	44	54
ddss	62	68	65
ssds	31	11	37
dsds	23	28	114
sdds	11	17	47
ddds	18	29	54
sssd	9	6	14
dssd	19	24	41
sdsd	6	13	39
ddsd	12	14	52
ssdd	4	4	15
dsdd	8	15	21
sddd	4	9	20
dddd	4	5	21
Totaux	388	353	659

On voit que les effectifs observés vont, chez Boniface, de 92 à 4 ; chez Alcuin, de 68 à 4 ; chez Abbon, de 114 à 7 (à quoi il faut ajouter, chez ce dernier, l'unique spondaïque, de la forme *ddsss*). On observe en outre que les effectifs extrêmes ne concernent pas les mêmes types hexamétriques ; pour Boniface, ce sont *dsss*, d'une part, et, d'autre part, *ssdd*, *sddd*, *dddd* (tous trois avec une fréquence de 4) ; pour Alcuin, ce sont *ddss* et *ssdd* (ce dernier en commun avec Boniface) ; pour Abbon, *dsds* et *ssss*. On voit bien ainsi que les effectifs varient très fort de type à type, et, pour un même type, d'auteur à auteur.

A partir de tableaux de ce genre, on peut entamer des comparaisons fructueuses, entre auteurs, entre œuvres (éventuellement d'un même auteur), entre parties d'une même œuvre. La méthode sera celle du χ^2, déjà exposée. Pour le tableau tel que celui qu'on vient de voir, on additionne les effectifs d'un même type hexamétrique : soit le type *dsds* qui donne un total global de 165. Les trois textes totalisant 1 400 vers, le type considéré y a une fréquence globale de 165/1 400 = 0,1179. En multipliant cette fréquence par le nombre de vers de chaque texte, on obtient un effectif calculé qui répartit l'effectif global au prorata de la longueur des textes considérés : ce sera 45,75, 41,62, 77,70. On opère de la sorte pour tous les types, puis on calcule χ^2 : pour chaque donnée on porte au carré l'écart entre effectif observé et effectif calculé et on divise par l'effectif calculé ; la somme de tous ces termes est la valeur de χ^2, que l'on interprète au moyen d'une table. On voit tout le profit que l'on peut tirer d'une pareille méthode, qui fournit le moyen d'évaluer l'homogénéité ou l'hétérogénéité d'un auteur, d'un texte, des diverses parties d'un même texte.

Si les données concernent un nombre élevé de textes, on aura profit à les traiter par l'analyse des correspondances, dont le principe et une application ont été exposés dans le paragraphe précédent. C'est ce que j'ai fait pour un ensemble de 25 textes poétiques[26]. Dans la publication que j'en ai faite, on peut voir les graphiques pour les

26. Voir Évrard 1988. Ce travail a été réalisé avec l'assistance technique de J. P. Benzecri.

trois premiers facteurs. J'y ai aussi reproduit les résultats des calculs effectués et listés par l'ordinateur. C'est une partie de ces résultats qui sert à tracer (automatiquement) les graphiques.

Je voudrais, à ce sujet, introduire une remarque qui me paraît importante. Pour chaque élément étudié, l'ordinateur calcule non seulement la position sur chacun des axes (au nombre de 15, dans ce cas-ci, puisque le tableau est de 16 x 25), mais aussi la contribution relative ainsi mise en œuvre et la contribution à l'inertie. Un mot d'explication à ce sujet. La détermination d'un axe correspondant à un facteur se fait généralement par la méthode des moindres carrés (d'autres méthodes sont possibles, mais je ne m'y arrêterai pas ici) : la perpendiculaire que l'on abaisse sur l'axe à partir du point où se situe un objet a une mesure que l'on porte au carré ; on additionne tous les carrés correspondant aux divers objets ; une méthode de calcul bien connue permet de déterminer les conditions où cette somme est la moins élevée, ce qui donne la possibilité de définir l'axe qui est le plus proche de l'ensemble des données. Mais dans le calcul d'un axe, les diverses données n'interviennent pas de manière égale ; le calcul de la contribution relative indique dans quelle mesure telle donnée est responsable du choix de l'axe ; si on additionne toutes les contributions relatives (ici, 15) d'un objet, on arrive à l'unité (ou à 1 000, si, comme cela se pratique souvent, on calcule jusqu'à la troisième décimale et qu'on multiplie le résultat par 1 000, pour faciliter la lecture). Par ailleurs, relativement à un axe, les points qui représentent les objets forment un nuage qui a son centre de gravité, son inertie ; ici, à nouveau, la contribution des différents objets est plus ou moins forte et c'est ce que mesure le calcul de l'ordinateur. On voit donc que l'interprétation d'un graphique d'analyse des correspondances ne peut se faire avec sécurité que si on tient compte de ces deux éléments. Un exemple : la contribution relative au premier facteur du poème 64 de Catulle est de 599 ; celle du premier livre des *Epitres* d'Horace est de 7. En revanche, pour le deuxième facteur, elle est de 181 pour le premier texte et de 808 pour le second. On voit que l'interprétation ne peut ignorer de telles indications. On trouvera plus de précisions sur ce sujet dans J.-P. et Fr. Benzecri 1984.

Revenons au tableau de distribution des hexamètres qu'on a vu plus haut. Il permet de calculer les effectifs et les pourcentages du spondée et du dactyle à chacun des quatre premiers pieds chez le ou les auteurs considérés : il suffit, pour le spondée quatrième, d'additionner les effectifs des huit premiers types, les huit autres donnant l'effectif des dactyles ; pour le troisième pied, on groupe les types par quatre : les deux groupes impairs donnent l'effectif des spondées et les deux pairs, celui des dactyles ; et ainsi de suite. Pour Boniface et pour Abbon, on obtient ainsi le tableau suivant.

Spondées et dactyles aux quatre premiers pieds
Effectifs et pourcentages pour les *Enigmata*

	I	II	III	V	Total
D : eff.	238	154	103	66	561
%	0,61	0,40	0,27	0,17	0,36
S : eff.	150	234	285	322	991
%	0,39	0,60	0,73	0,83	0,64

Pour Abbon, les résultats sont les suivants.

	I	II	III	IV	Total
D : eff.	426	352	329	223	1330
%	0,65	0,53	0,50	0,34	0,50
S : eff.	233	307	330	436	1306
%	0,35	0,47	0,50	0,66	0,50

On observe que, chez les deux auteurs, le dactyle, majoritaire au premier pied, perd progressivement de son importance, le spondée ayant le mouvement inverse. C'est là une caractéristique qui se retrouve chez à peu près tous les utilisateurs de l'hexamètre, avec des variantes dont certaines sont révélatrices. On peut la désigner comme la spondaïsation croissante de l'hexamètre.

Il me semble qu'à partir de ces données, il est possible de calculer un autre type d'effectif théorique, avec d'autres implications pour l'interprétation. Le poète qui veut écrire en hexamètres dactyliques doit se soumettre aux exigences relatives à la succession des quantités. Mais il peut le faire avec le seul souci de respecter la règle et ses licences ou bien il peut se soucier, de manière plus ou moins consciente, de créer une atmosphère rythmique résultant de la disposition et du mode de succession des spondées et des dactyles. En d'autres termes, je voudrais vérifier dans quelle mesure le choix, pour telle position dans le vers, d'un dactyle ou d'un spondée est indépendant de la réalisation des pieds voisins ; dans quelle mesure, en revanche, on peut déceler une tendance à favoriser ou à écarter telle ou telle combinaison. Pour ce faire, il est utile de partir de données aussi proches que possible du texte qu'on soumet à examen. Par une application de la loi des probabilités composées (voir le paragraphe sur la métrique phonémique), il est possible, à partir des fréquences de chaque pied en chaque position, de calculer la probabilité des divers types d'hexamètre et, en multipliant par la longueur du texte en nombre de vers, de déterminer un effectif probable, d'une probabilité qui résulterait du fait que le choix du spondée ou du dactyle est aléatoire et ne dépend d'aucun autre facteur. Prenons par exemple le type *dsds*. Chez Boniface, on en obtient la probabilité et l'effectif théorique en pratiquant les opérations suivantes :

$$0{,}61 \times 0{,}60 \times 0{,}27 \times 0{,}83 = 0{,}082$$

$$0{,}082 \times 388 = 31{,}82$$

(on a reconnu dans les facteurs de la multiplication les fréquences du dactyle premier, du spondée deuxième, du dactyle troisième et du spondée quatrième ; quant au facteur 388, c'est le nombre de vers du poème de Boniface). Comme l'effectif observé de *dsds* chez Boniface est 23, on croit déceler une tendance, de sa part, à éviter ce rythme en alternance. Faisons les mêmes opérations pour Abbon :

$$0{,}65 \times 0{,}47 \times 0{,}50 \times 0{,}66 = 0{,}10$$

$$0{,}10 \times 659 = 65{,}9$$

Ici la situation est inverse et a un caractère beaucoup plus accusé : en face d'un effectif aléatoire de 66, l'effectif observé, qui est 114, marque une recherche presque indubitable, qu'elle soit consciente ou non, de ce rythme alternant.

Si l'on opère de la sorte pour tous les types dans un poème, on a une série d'effectifs calculés qui diffèrent de ceux qui résultent d'une répartition proportionnelle d'effectifs globaux (selon la méthode exposée précédemment). Cela n'a rien d'étonnant,

puisque non seulement le mode de calcul, mais surtout les données utilisées sont tout à fait différentes.

Pour apprécier les écarts entre les observations et les effectifs calculés de la sorte, on dispose de deux méthodes. La première consiste à vérifier dans quelle mesure la prévision a été réalisée: il suffit, pour ce faire, de diviser l'effectif observé par l'effectif calculé; si le premier est supérieur au second, le quotient est supérieur à l'unité et marque que la réalisation dépasse l'espérance (au sens qui a été défini plus haut); dans le cas contraire, un quotient inférieur à l'unité conduit à l'idée que la prévision n'a pas été complètement réalisée. L'autre méthode consiste à soumettre les deux séries (l'une observée et l'autre calculée) à un test de χ^2.

Il serait sans doute fastidieux d'exposer le détail de ces calculs. Au reste, le lecteur curieux de la chose dispose de toutes les données nécessaires et peut exécuter toutes les opérations. Si l'on examine dans quelle mesure la prévision a été réalisée, on consate que, chez Vynfreth, les résultats vont de 0,72 à 1,55 et que, quatre fois, la donnée observée est égale à la prévision (quotient 1), tandis que, pour Abbon, les quotients s'étagent de 0,19 à 1,86 et qu'aucun d'entre eux n'est égal à 1. Parallèlement, le *chi*^{2} calculé pour Vynfreth correspond à une probabilité de dépassement aléatoire d'approximativement 0,50, tandis que, pour Abbon, cette probabilité est inférieure à 0,00001. La conclusion s'impose: tandis que Vynfreth, volontairement ou non, ne s'écarte guère du tout-venant au point de vue métrique, Abbon se singularise à un point très élevé.

Les méthodes décrites ci-dessus permettent de déceler dans une même œuvre, ou dans un ensemble d'œuvres, des zones d'homogénéité ou d'hétérogénéité, dont seule l'analyse littéraire pourra tenter de définir la signification. Je donnerai un seul exemple: j'ai montré (Évrard 1978) que, dans la *5*^{e} *Bucolique*, la métrique varie significativement selon que c'est Ménalque ou Mopsus qui parle. Parallèlement à cette opposition, on observe, chez les deux bergers, des différences marquées dans l'utilisation de la syntaxe et dans les modes d'expression; on a donc là des indices convergents. Cela ne me porte nullement à affirmer que Virgile ait consciemment voulu cette différenciation, mais me semble suffisant pour justifier l'affirmation que, voulue ou non, cette différenciation contribue au caractère de la pièce où elle se manifeste. J'ajouterai que, si cette opinion ne se fondait sur une analyse quantitative, on pourrait craindre qu'elle ne soit que l'expression d'une impression subjective sans fondement, alors que l'assise qui lui est fournie par le calcul force à reconnaître qu'il y a là quelque chose à expliquer.

Sylvie Mellet
CNRS, 'Bases, Corpus et Langage'
98, bd Edouard-Herriot
F-06204 Nice Cedex 3.

Étienne Évrard
L.A.S.L.A.
Rue Sous-le-Bois, 12
B-4031 Angleur.

BIBLIOGRAPHIE

BENZECRI, Jean-Paul 1982. Histoire et préhistoire de l'analyse des données. Paris: Dunod.

— 1991a. "Typologie de textes grecs d'après les occurrences des formes des mots-outils". *Les Cahiers de l'Analyse des Données* XVI, 1, 61-86.

— 1991b. "Typologie de textes latins d'après les occurrences des formes des mots-outils ". *Les Cahiers de l'Analyse des Données* XVI, 4, 439-465.

BENZECRI, Jean-Paul & *al.* 1981. *Pratique de l'analyse des données.* Paris: Dunod.

BENZECRI, J.-P. & F. 1984[2]. *Pratique de l'analyse des données - 1 Analyse des correspondances: exposé élémentaire.* Paris: Dunod.

BRUNET, Étienne 1981. *Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française.* Genève-Paris: Slatkine-Champion

COHEN, Marcel 1967. "Sur l'histoire de la statistique en linguistique". *Études de linguistique appliquée* 5, 3-8.

COSSETTE, André 1994. *La richesse lexicale et sa mesure.* Genève-Paris: Slatkine-Champion.

DAVID, Jean & Robert MARTIN (éds.) 1974. *Statistique et linguistique.* Actes du colloque organisé par le Centre d'Analyse Syntaxique de l'Univ. de Metz (2-3 mars 1973). Paris: Klincksieck.

DROBISCH 1866. "Ein statistischer Versuch über die Formen des lateinischen Hexameters". *Ber. Verh. Königl. Sächs. Ges. der Wiss., philol.-hist. Cl.* 18, 75-139.

— 1868. "Weitere Untersuchungen ...". *Ibid.* 20, 16-65.

— 1871. "Ueber die Classification der Formen des Distichons". *Ibid.* 23, 1-33.

ESTOUP, Jean-Baptiste 1916[4]. *Gammes sténographiques.* Paris.

ÉVRARD, Étienne 1964. "Note sur le calcul d'une distribution de vocabulaire". *Bulletin des Amis de l'Université de Liège* 36, 3s.

— 1967. "Deux programmes d'ordinateur pour l'étude quantitative du vocabulaire". *Revue de l'org. intern. pour l'étude des langues anciennes par ordinateur,* 81-95 (remaniement d'une conférence faite en 1963).

— 1972. "Scansion automatique de l'hexamètre grec". Revue de l'org. intern. pour l'étude des langues anciennes par ordinateur 2, 1-29.

— 1978. "Quelques observations sur la 5[e] *Bucolique* de Virgile". *Les Études Classiques* XLVI, 327-338.

— 1982. "Quelques traits quantitatifs du vocabulaire du *Moretum*". *Latomus* 41, 550-565.

— 1983. "La fréquence des voyelles en latin". *Computers in literary and linguistic Research,* Pise, 157-166.

— 1984. "Réflexions sur les structures phoniques du vers latin". *Revue, Informatique et Statistique dans les Sciences humaines* XX, 121-134.

— 1985. "Richesse et mode d'enrichissement d'un vocabulaire". In: *Actes de la 11[e] Conférence internationale de l'A.L.L.C.,* Genève-Paris: Slatkine-Champion, 147-152.

— 1988. "Analyse factorielle de la composition métrique de l'hexamètre dactylique latin." *Les Cahiers de l'Analyse des Données* XIII, 9-17.

FÖRSTEMANN, E. 1852. "Numerische Lautverhältnisse im Griechischen, Lateinischen und Deutschen" (en fait, il s'agit du gotique). *Zeitschrift für vergleichende Sprachforschung* I, 161-179.

GRAMMONT, M. 1933 (1946[3]). *Traité de Phonétique.* Paris: Delagrave.

GROTJAHN, Rüdiger 1981. *Hexameter Studies*. Bochum: Brockmeyer (coll. Quantitative Linguistics 11).

GUIRAUD, Pierre 1954. *Les caractères statistiques du vocabulaire*. Paris: PUF.

— 1960. *Problèmes et méthodes de la statistique linguistique*. Paris: PUF.

HERDAN, G. 1956. *Language as Choice and Chance*. Groningen.

— 1964. *Quantitative Linguistics*. Londres: Butterworths.

HERESCU, N.I. 1960. *La poésie latine. Étude des structures phoniques*. Paris.

HULTGREN 1872. "Statistische Untersuchungen des Distichons". *Ber. Verh. Königl. Sächs. Ges. der Wiss., philol.-hist. Cl.* 24, 1-28.

KENDALL, M. G. 1948. *Rank Correlation Methods*. Londres.

KNIGHT W.F.J. 1939. *Accentual symmetry in Virgil*. Oxford.

— 1944. *Roman Virgil*. Londres, 239-242.

LEBART, Ludovic & André SALEM 1994. *Statistique textuelle*. Paris: Dunod.

LEBEK W.D. 1982. "Sinnbezug und Hexametergestalt im Aeneisproömium". *Hermes* 110, 195-211.

LOPES, L. & ODIN, G. 1987. "Distinguishing between random and non-random events". *Journal of Exp. Psychol.* 13, 392-400.

MARTIN, Robert & Charles MULLER 1964. "Syntaxe et analyse statistique". *Tra. Li. Li.* II, 1, 207-233.

MELLET, Sylvie 1988. *L'imparfait de l'indicatif en latin classique. Temps, aspect, modalité*. Paris - Louvain: Peeters (coll. B.I.G.).

— 1990. "Quelques réflexions sur l'exploitation statistique des données informatisées". *Les Études classiques* LVIII, 105-113.

— 1994. "Statistique, syntaxe latine, pragmatique". *Travaux du Cercle linguistique de Nice* 16, 131-140.

— 1996. "Les atouts de la lemmatisation". In G. Moracchini (éd.) Bases de données linguistiques conceptions, réalisations, exploitations, Actes du Colloque International de Corte (octobre 1995). Univ. de Corse - Univ. de Nice Sophia-Antipolis, 309-316.

— 1998. "Les tragédies de Sénèque vues à travers Hyperbase". In: S. Mellet & M. Vuillaume (éds), 255-271.

MELLET, Sylvie & Marcel VUILLAUME (éds) 1998. *Mots chiffrés et déchiffrés*, Mélanges offerts à Étienne Brunet. Genève - Paris: Slatkine - Champion (coll. Travaux de linguistique quantitative 64).

MULLER, Charles 1964. "Calcul des probabilités et calcul d'un vocabulaire". *Tra. Li. Li.* II 1, 1-7.

— 1967. *Étude de statistique lexicale: le vocabulaire du théâtre de Corneille*. Paris: Larousse.

— 1973. *Initiation aux méthodes de la statistique linguistique*. Paris: Hachette.

— 1992. Initiation aux méthodes de la statistique linguistique (réimpr.de l'éd. de 1973). Paris: Champion.

NETHERCUT. W.R. 1979. "Daphne and Apollo: a dynamic encounter". *The classical Journal* 74, 333-347.

NOVI, Michel 1998. *Pourcentages et tableaux statistiques.* Paris: PUF (coll. Que sais-je? n° 3337).

PEARSON K. 1900. "On the criterion that a given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling". *Philosophical Magazine* 5e sér., vol. 50, n° 302, 157-175.

PEARSON, E. S. & H. O. HARTLEY 1958. *Biometrika Tables for Statisticians* I. Cambridge.

PFISTER, Raimund 1952. "Zur Lautstruktur des Lateinischen". *Münchener Studien zur Sprachwissenschaft* 1, 5-19 (rééd. en 1956).

RIGO, Georges 1994. "L'entropie comme indice de diversification du vocabulaire dans les tragédies de Sophocle". *Revue, Informatique et Statistique dans les Sciences humaines* XXX, 109-125.

SCHMIEL R. 1981. "Rhythm and accent: Texture in Greek Epic poetry". In: R. Grotjahn, 1-32, avec une note complémentaire de R. Grotjahn, 33-74.

SOUBIRAN J. 1969. "Les hexamètres spondaïques à quadrisyllabe final". *Giornale italiano di filologia* 21, 329-349.

TROUBETZKOY, N. S. 1949 (1957^2). *Principes de phonologie*, tr. J. Cantineau. Paris.

SIEGEL, S. 1956. *Nonparametric Statistics.* New-York.

VIERTEL Antonius 1862. "De versibus poetarum latinorum spondiacis". *Jahrbücher für classische Philologie* 8, 801-811.

WAGENAAR, W.A. 1972. "Generation of random sequences by human subjects: a critical survey of literature". *Psychol. Bull.* 77, 65-72.

YULE, G. U. 1944. *The Statistical Study of Literary Vocabulary.* Cambridge Univ. Press (réimpr. Archon Books, Hamden, Connecticut, 1968).

ZIPF, G. K. 1935. *The Psychobiology of Language, an Introduction to Dynamic Philology.* Boston: Houghton - Mifflin.