CAHIER DE RECHERCHE

MAI 1994 ■ N°61

AVANCÉES EN ANALYSE LEXICALE

Structure lexicale, registres et thèmes

Valérie BEAUDOUIN

Les motivations des volontaires bénévoles à une grande étude épidémiologique

Pascale HEBEL

Crédoc - Cahier de recherche. N° 061.
Mai 1994.

Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie

CRÉDOC

AVANCÉES EN ANALYSE LEXICALE

Structure lexicale, registres et thèmes

Valérie BEAUDOUIN

Les motivations des volontaires bénévoles à une grande étude épidémiologique

Pascale HÉBEL

dans le cadre du programme général de recherches du CRÉDOC

Département Prospective de la Consommation

MAI 1994

142, rue du Chevaleret
75013 - PARIS

Sommaire

AVANT-PROPOS 1

STRUCTURE LEXICALE, REGISTRES ET THÈMES 2

LES MOTIVATIONS DES VOLONTAIRES BÉNÉVOLES À UNE GRANDE ÉTUDE ÉPIDÉMIOLOGIQUE 79

AVANT-PROPOS

Les travaux de recherche en analyse des données textuelles se poursuivent au CRÉDOC et s'orientent vers des voies nouvelles. Le rapport suivant est composé de deux parties qui reflètent les avancées du Département Prospective de la consommation dans deux directions différentes : en amont, les méthodes de traitement, en aval les méthodes d'interprétation.

La première partie porte sur l'analyse de grands corpus textuels. On cherche tout d'abord à étudier la pertinence d'indicateurs statistiques comme la richesse lexicale ou la distribution des fréquences pour caractériser la structure lexicale de grands corpus. D'autre part, sur deux de ces corpus, les oeuvres complètes de Corneille et Racine, différentes méthodes statistiques (analyse des données, méthodes probabilistes) sont utilisées de manière complémentaire pour analyser les textes. Les résultats des analyses sont ensuite soumis à l'interprétation. Étant donné que ces corpus ont déjà été largement étudiés par la critique littéraire, on pourra vérifier la cohérence des interprétations et comparer les performances de méthodes assistées par ordinateur par rapport à des méthodes plus traditionnelles. Ce sera également l'occasion de présenter une méthode en deux temps pour classer les textes.

La seconde partie est consacrée à l'analyse d'une question ouverte sur les motivations des bénévoles à participer à une grande enquête épidémiologique. On montre que les motivations peuvent être interprétées à la lumière de la théorie des représentations sociales. Une typologie des réponses segmente les individus selon leur degré d'implication vis-à-vis de leurs contraintes primaires (état de santé, alimentation, ...). Les niveaux de préoccupation de chacun sont plus au moins élevés et vont de l'utilitaire (suivi médical) à l'altruisme (survie de l'espèce) en passant par le stratégique (bonne alimentation donc meilleure forme). Les caractéristiques socio-démographiques des individus expliquent les comportements altruistes de 43 % des bénévoles. Nous voyons à nouveau ici le rôle que peut jouer l'analyse lexicale comme outil d'exploration des représentations sociales.

STRUCTURE LEXICALE, REGISTRES ET THÈMES

Valérie BEAUDOUIN

Avec la collaboration de Claire EVANS

Sommaire

INTRODUCTION 4

1 - RICHESSE LEXICALE ET DISTRIBUTION DES FRÉQUENCES 8

1.1. Comparaison de la richesse lexicale 9

1.2. La distribution fréquentielle des mots sur de larges corpus vérifie-t-elle la loi de Zipf ? 13

1.3. Incidence sur l'analyse lexicale 22

2 - STYLISTIQUE ET ANALYSE LEXICALE : CORNEILLE ET RACINE 26

2.1. Outils de statistique textuelle : Alceste et Hyperbase 26

2.2. Tragédies et comédies 32

2.3. Les pièces de Racine 37

2.4. Tragédies de Racine et Corneille 49

BIBLIOGRAPHIE 55

ANNEXES 59

Annexe 1 : Distributions de fréquences 60

Annexe 2 : Vocabulaire spécifique de chacune des pièces de Racine 63

Annexe 3 : Les tragédies de Corneille et Racine par rapport à la base Frantext (TLF) 69

SOMMAIRE DES FIGURES 77

INTRODUCTION

Les travaux sur les méthodes de traitement des données textuelles se dirigent au CRÉDOC vers l'analyse de textes de plus en plus longs. Du traitement de réponses à des questions ouvertes, nous cherchons à étendre les méthodes à l'exploitation de textes plus riches comme les entretiens semi-directifs. De plus en plus nombreux sont les textes saisis sur support informatique, qui sont disponibles pour des traitements à l'aide de méthodes de statistique textuelle —pour reprendre le nom générique de toutes ces méthodes proposé par L. Lebart et A. Salem— (1994).

Les corpus dont nous disposons se diversifient —comme nous le montrions dans le Cahier de recherche n° 48 (Beaudouin et Lahlou, 1993)— et s'allongent, dépassant parfois les capacités de calcul de certains logiciels. La variété des corpus nous pousse à enrichir les méthodes existantes et à en adopter de nouvelles. En effet, des textes longs qui ont chacun une structure formelle spécifique ne peuvent être analysés par des méthodes standardisées comme celles que l'on utilise pour les questions ouvertes. L'exploration de ces nouveaux corpus nous permet, en confrontant la technique à des matériaux différents, de développer de nouvelles méthodologies et de nouveaux concepts enrichissant ainsi progressivement notre "boîte à outils". Nous continuons à tester les nouveaux développements d'Alceste, le logiciel de Max Reinert, et nous commençons à utiliser systématiquement de manière complémentaire Hyperbase (logiciel conçu par Étienne Brunet (1978, 1992) - INALF - Nice).

Parallèlement, Max Reinert avec qui nous collaborons depuis quelques années est en train de mettre en place une nouvelle version de son logiciel Alceste qui permettra de résoudre les problèmes de taille de corpus et de vocabulaire. Pour le moment, nous ne pouvons traiter des textes dont la taille du vocabulaire excède 10 000 formes. Ses travaux visent également à améliorer le système de lemmatisation, qui est encore critiqué par une partie de la communauté scientifique, bien que ses adeptes gagnent du terrain. Nous avons pu montrer l'an dernier (Beaudouin et Lahlou, 1993) que l'incidence du mode de lemmatisation sur les résultats était minime, mais une lemmatisation de type linguistique aurait le mérite de rendre

encore plus crédible les analyses. L'intégration de dictionnaires linguistiques avec des entrées de dictionnaire et toutes les formes fléchies devrait améliorer la qualité des résultats et des présentations.

De plus, les nouvelles orientations de Max Reinert permettront de définir avec plus de liberté le découpage en énoncés. Dans la version antérieure du logiciel Alceste, le corpus était arbitrairement découpé en unités de 1 à 4 lignes en respectant les ponctuations. Dans la prochaine version, il sera possible d'imposer des types de découpages plus souples du texte. Par exemple, si l'on travaille sur des entretiens semi-directifs, il serait préférable que chaque réponse constitue une unité. De même quand on travaille sur des textes littéraires, l'unité peut être de manière préférentielle le paragraphe, la strophe ou des groupements de vers.

Cette nouvelle version sera également adaptée au système UNIX avec la collaboration du CRÉDOC ce qui permettra d'effectuer les analyses sur station de travail. Le traitement sur de longs corpus peut prendre deux heures sur un Mac FX , mais ce temps diminue de 90 % sur station de travail. Or, comme pour toute analyse, on est toujours amené à faire de nombreux traitements en variant les paramètres, le gain de temps est considérable et la qualité des analyses bien meilleure, car on hésite moins alors à tester différents paramètres pour mettre à jour les stabilités.

Alors que nous allons être de plus en plus souvent confrontés à l'analyse de corpus très longs, il nous paraissait indispensable d'avoir des indicateurs sur la richesse et la structure lexicale de chacun des corpus pour pouvoir les comparer mais aussi pour mieux comprendre les effets réducteurs de la statistique lexicale. Nous allons essayer de caractériser cinq corpus :

- les oeuvres théâtrales de Corneille[1] ;
- les oeuvres de Racine[2] ;

Ces deux corpus nous ont été aimablement fournis par Jacques Roubaud. Ils proviennent du fonds informatisé de la Bibliothèque de France.

- 580 portraits de jeunes en réinsertion rédigés par les membres des Missions Locales et des PAIO. Ces corpus nous ont été fournis par la DIJ (Délégation interministérielle à la jeunesse) dans le cadre d'une recherche sur les trajectoires d'insertion ;

---

[1] Oeuvres de P. Corneille, Nouvelle édition revue sur les plus anciennes impressions et les autographes et augmentée de morceaux inédits, des variantes, de notices, de notes, d'un lexique des mots et locutions remarquables, d'un portrait, d'un fac-similé, etc. Par M. CH. Marty-Laveaux, Paris, Hachette et Cie, 1862.

[2] Oeuvres de J. Racine, Nouvelle édition revue sur les plus anciennes impressions et les autographes et augmentée de morceaux inédits, de variantes, de notices, de notes, d'un lexique des mots et locutions remarquables, d'un portrait, d'un fac-simile, etc. Par M. Paul Mesnard, Paris, Librairie Hachette et Cie, 1885.

- un ensemble de définitions extraites du Robert électronique autour du concept d'alimentation. Ce corpus a été constitué et analysé par Saadi Lahlou (1994) dans le cadre de ses recherches sur les représentations sociales.
- 1 000 réponses à la question ouverte : "*Un Petit déjeuner idéal, à quoi ça vous fait penser ?*". Cette question a été posée par téléphone dans le cadre de l'enquête sur la consommation du CRÉDOC en novembre 1992 (Beaudouin, Lahlou et Yvon 1993).

Les résultats permettent de faire des analyses statistiques en connaissance de cause : grâce à la distribution des fréquences on sait exactement sur quelle partie du vocabulaire s'effectuent les analyses.

Une des préoccupations majeures était de mesurer l'efficacité de méthodes de traitement automatique par rapport à des méthodes traditionnelles comme l'analyse de contenu, l'étude stylistique.... Dès que l'on travaille sur de gros corpus, il est indispensable de savoir quelle confiance on peut accorder à des analyses automatiques.
Tant que l'on travaille sur des textes courts, une analyse manuelle paraît toujours plus fine qu'une analyse automatique. Pour de longs corpus, peut-on s'épargner la tâche délicate, fastidieuse et coûteuse du post codage ? Peut-on s'appuyer sur des résultats d'analyse statistique pour se lancer dans des opérations d'interprétation de type sociologique ou autre ? Ou est-ce un outrage au texte que de l'examiner au travers de la loupe grossissante des statistiques ? C'est pour mieux comprendre les aspects réducteurs de la statistique lexicale que nous avons d'ailleurs fait une analyse précise des distributions de fréquence.

La seconde des préoccupations à l'origine de ce travail est le problème du classement des objets, problème qui avait déjà été longuement abordé dans un précédent Cahier de recherche[1]. Toujours dans cette optique, nous cherchons à proposer des systèmes de description de type combinatoire pour classer les textes. Après avoir classé les énoncés, chaque texte est constitué d'un certain pourcentage d'énoncés provenant de la première classe, d'un autre pourcentage provenant de la seconde... On obtient ainsi un profil lexico-sémantique de chaque texte. Classer les objets revient à regrouper les textes qui ont un profil similaire (nous entendons par profil une combinaison de classes d'unités de contexte).

[1] LAHLOU, S., MAFFRE, J., BEAUDOUIN, V., (1991).- *La codification des objets complexes : réflexions théoriques et application à un corpus de 8000 produits alimentaires*, CRÉDOC, Cahier de recherche n°23, Paris.

Nous n'avons choisi comme terrain d'analyse que le corpus constitué par les oeuvres complètes de Corneille et Racine. Cela peut paraître à première vue assez éloigné des préoccupations du CRÉDOC et mérite quelque justification.

C'est tout d'abord le plus grand corpus que nous ayons jamais eu l'occasion d'analyser. Il est en effet constitué de plus de 700 000 occurrences.

Le travail sur des textes en vers est particulièrement intéressant dans la mesure où il s'agit de corpus contraints à une forme d'expression très régulière. Les corpus en alexandrins de la période "classique" respectent avec une cohérence extrême des contraintes très rigides (Roubaud, 1978, 1986). Ils constituent donc une sorte de cas idéal, bien contrôlé, qui est un excellent matériau de test technique un peu comme le rat de laboratoire de souche génétique bien définie est un bon matériel d'expérience biologique.

D'autre part, nous n'avons guère eu jusqu'à présent l'occasion de comparer les performances d'une analyse automatique par rapport à une analyse traditionnelle. C'est pourquoi nous avons choisi un corpus de textes qui ont été explorés par la critique à de très nombreuses reprises depuis leur création y compris à l'aide de méthodes de statistique lexicale. En effet, les travaux liminaires dans ce domaine ont été réalisés par Charles Muller (1967) sur les oeuvres complètes de Corneille et remontent à 1967 ; en 1983, dans la même lignée, Charles Bernet (1983) consacrait un ouvrage aux tragédies de Racine. On cherchera donc à évaluer la qualité des analyses assistées par les logiciels Alceste et Hyperbase sur nos corpus.

1 - RICHESSE LEXICALE ET DISTRIBUTION DES FRÉQUENCES

Depuis qu'au CRÉDOC nous travaillons dans le domaine de la statistique textuelle, nous avons accumulé des corpus de textes nombreux et variés tant par leur taille que par leur contenu. Nous avons naturellement ressenti la nécessité d'indicateurs pour caractériser et comparer ces corpus.
La distribution (ou la gamme) des fréquences de mots et la richesse du vocabulaire (qui en découle) sont depuis longtemps considérées comme des indicateurs pertinents pour la caractérisation des textes. La distribution de fréquence est un tableau qui associe à chaque classe de fréquence un effectif qui correspond au nombre de vocables qui ont cette fréquence. Elle fait abstraction du **contenu** lexical du texte, mais elle donne une image de sa **structure** lexicale. En ce qui concerne l'évaluation de la richesse lexicale, la question qui n'est que partiellement résolue.

L'utilisation de ces outils sur nos corpus nous montrera que la richesse lexicale varie sensiblement d'un texte à l'autre et que dans la *plupart des cas,* on est en mesure de comparer la richesse de nos corpus deux à deux. En revanche, l'étude des courbes de distribution conduit à des résultats assez décevants. En effet, même si certaines constantes peuvent être mises en évidence pour chacune des courbes de distribution, ces dernières sont difficiles à modéliser, contrairement à ce que prétendait Zipf (1936). Il semble en effet qu'il y ait un nombre élevé de paramètres entrant en jeu, qui ne sont pas toujours simples à évaluer et encore moins à interpréter.

L'étude et la comparaison de gammes de fréquence permettent également de mieux mesurer sur quelle partie du vocabulaire porte l'analyse statistique textuelle. En effet, quelles que soient les techniques d'analyse, on est amené à fixer un seuil de fréquence en dessous duquel les mots ne sont pas analysés. L'incidence de ce seuil variant en fonction de la gamme de fréquence, il est nécessaire de bien connaître celle-ci, afin de faire des choix pertinents, ou au moins en connaissance de cause.

1.1. COMPARAISON DE LA RICHESSE LEXICALE

Avant d'étudier la distribution fréquentielle dans sa globalité, nous allons utiliser certains indicateurs de richesse lexicale qui nous permettront de comparer différents corpus.
La richesse lexicale a été un des grands thèmes de recherche dans la statistique lexicale (Charles Muller (1967), Étienne Brunet (1978), Charles Bernet (1983), Philippe Thoiron, Dominique Labbé, Pierre Hubert (1988)...), mais l'intérêt pour ce thème s'est un peu émoussé en raison des difficultés qu'il y a à trouver de bons indicateurs. Rappelons une fois encore que l'on emploie le terme de "richesse lexicale" - et ce n'est là qu'une convention - comme un terme technique pour décrire la structure du vocabulaire indépendamment de son contenu et qu'aucune "valeur" ne lui est attachée. En général, on compare la taille du vocabulaire (**V: nombre de vocables**) à la taille du corpus (**N : nombre de mots**)[1]. Contrairement à ce qui pourrait paraître, il est extrêmement difficile de comparer la richesse lexicale de différents corpus.

Deux cas ne posent cependant pas de problèmes. Quand les deux textes sont de taille similaire, plus le rapport N / V est petit plus le texte est riche. Un autre cas se résout simplement : soit le texte T (N, V) et le texte T' (N', V'), si N > N' et V< V', de toute évidence le texte T' est plus riche que T.

Dans tous les autres cas, pour comparer des textes de longueur différentes, il faut réduire la taille du plus long. Pour cela, Charles Muller (1977) proposait d'avoir recours à un modèle théorique d'accroissement du vocabulaire (issu du modèle de l'urne). Cette courbe d'accroissement du vocabulaire permet de prédire la taille du vocabulaire quand on atteint N mots. Ainsi, quand on compare deux textes, on "rétrécit" (Labbé, 1994) le plus long jusqu'à ce qu'il atteigne la taille du plus court, puis on estime la taille du vocabulaire à partir de la courbe théorique. Ce système de réduction n'est acceptable que si la courbe théorique correspond à la courbe réelle. Or ce n'est pas toujours le cas. Par exemple pour les oeuvres de Racine, la courbe de l'accroissement réel du vocabulaire passe en dessous de la courbe théorique (Bernet, 1983, p. 116). Dominique Labbé (Labbé et Hubert, 1994) aboutit aux mêmes résultats sur les discours politiques de de Gaulle et de Mitterrand. Cet écart entre

[1] Nous essaierons autant que possible de respecter les conventions d'appellation et de notation proposées par Ch. Muller (1977). On emploiera le terme de *mot* ou *occurrence* pour désigner chaque unité présente dans un texte, et le terme de *vocable* pour désigner les unités distinctes.

accroissement réel et accroissement théorique introduit donc un biais dans le calcul de la richesse du vocabulaire.

Nous allons sur nos corpus essayer de caractériser autant que possible la richesse du vocabulaire. Nous avons donc sélectionné cinq corpus :

- les oeuvres théâtrales de Corneille qui représentent un volume d'un demi million d'occurrences (corpus "Corneille") ;

- les oeuvres de Racine constituées de 165 mille occurrences. Comme le corpus précédent, il s'agit de textes représentatifs d'une langue soutenue (corpus "Racine") ;

- 580 portraits de jeunes en réinsertion qui ont un volume de 168 000 occurrences (corpus "Les portraits") ;

- un ensemble de définitions extraites du Robert électronique autour du concept d'alimentation. Ce corpus a été constitué et analysé par Saadi Lahlou (1994) dans le cadre de ses recherches sur les représentations sociales. Il est constitué de 140 000 occurrences (corpus "Manger").

- 1000 réponses à la question ouverte : "*Un Petit déjeuner idéal, à quoi ça vous fait penser ?*". Cette question a été posée par téléphone dans le cadre de l'enquête sur la consommation du CRÉDOC en novembre 1992 (Beaudouin, Lahlou et Yvon 1993). Il s'agit donc de réponses orales retranscrites qui représentent un volume d'un peu plus de 10 000 occurrences (corpus "Petit déjeuner").

On a réalisé le décompte de toutes les formes présentes y compris les noms propres qui sont particulièrement nombreux dans les trois premiers corpus. Les signes de ponctuation n'ont pas été comptabilisés. Notons que les choix du décompte ne sont pas parfaits puisque par exemple les noms composés, en dehors de ceux qui sont reliés par un signe typographique comme le tiret, ne sont pas reconnus comme des entités autonomes. Tout choix de dépouillement est arbitraire et peut prêter à critique. Ce qui compte c'est la constance dans le dépouillement. On a donc choisi pour les quatre corpus le module Hyperbase (logiciel conçu par Etienne Brunet (1992) - INALF) qui fabrique le dictionnaire des mots du corpus.

Le tableau ci-dessous propose pour les cinq corpus quelques indicateurs statistiques pour caractériser les textes.

Figure 1 : Indicateurs statistiques de caractérisation des corpus

Nom du corpus	V	N	N/V	Fréquence médiane	V_1 Hapax	V_1 / V	% V (freq <45)	% N (freq <45)	% N 10 freq max
Corneille	14 055	547 297	39	249	5060	36	91,7	13,8	20,0
Portraits	8 188	167 937	20,5	139	3297	40	95	23,6	24,3
Racine	9 288	164 845	17,7	151	3805	41	95,5	26,4	19,6
Manger	16 896	137 576	8,1	118	8734	52	98,1	40,8	19,9
Petit déjeuner	879	11 292	12,8	49	440	50	94,5	29	34,3

Légende :

V	: Nombre de vocables (étendue du vocabulaire)
N	: Nombre de mots (étendue du texte)
N / V	: Fréquence moyenne (Nombre moyen d'occurrences)
V_1	: Nombre d'hapax : nombre de mots n'apparaissant qu'une seule fois
V_1 / V	: proportion en vocables de fréquence 1
% V (freq<45)	: Pourcentage de vocables représentés par les fréquences de 1 à 45
% N (freq<45)	: Pourcentage d'occurrences représentées par les fréquences de 1 à 45
% N 10 freq max	: proportion des mots du corpus représenté par les dix fréquences les plus élevées

Il est assez remarquable, même si cela a été déjà dit à de très nombreuses reprises, qu'en ne conservant que les fréquences de 1 à 45, on arrive à couvrir plus de 90 % du vocabulaire (les formes distinctes employées) mais moins de 40 % des occurrences globales. Plus précisément, pour le corpus "Manger", les fréquences inférieures à 45 représentent 98 % du vocabulaire et 41 % des occurrences. En revanche, pour Corneille, les pourcentages respectifs sont de 92 % et 14 %. Il y a donc d'importantes variations selon les corpus.

Le pourcentage d'hapax dans le vocabulaire ainsi que la proportion du texte représentée par les dix fréquences les plus élevées sont également de bons indicateurs pour caractériser les textes. Ils varient selon les corpus, il nous reste à savoir si ces variations sont dues à la taille du corpus ou à des spécificités des textes. Il se peut par exemple, que le corpus du dictionnaire soit très différent des autres parce qu'il ne correspond pas à ce que Muller appelait "l'exercice normal du langage". La même question peut être posée pour le corpus constitué par des réponses à des questions ouvertes.

Le corpus "petit déjeuner" est nettement plus petit que tous les autres au niveau de la taille comme du vocabulaire, il sera donc difficile de le comparer avec les autres.

En ce qui concerne la richesse lexicale, on peut d'ores et déjà tirer certaines conclusions. Le corpus "Manger" est de loin le corpus le plus riche. Plus petit en taille que "Corneille", "Racine" et "Les portraits", il a le vocabulaire le plus étendu. Ceci est confirmé par la fréquence moyenne d'occurrences très inférieure à celle des autres corpus et une médiane assez basse (50 % des mots ou occurrences ont une fréquence inférieure à 118). Dans le dictionnaire, les mots sont moins souvent répétés que dans le discours et comme celui-ci a pour objectif de recenser tout le vocabulaire existant, son vocabulaire est beaucoup plus large : la moitié du vocabulaire est constituée d'hapax. D'autre part, le dictionnaire a moins recours à la syntaxe et aux mots-outils que le discours courant, or comme ce sont les mots-outils qui ont les fréquences les plus élevées, il est assez logique que ce soit le corpus le plus "riche".
Le corpus des "Portraits" est légèrement plus grand que celui de Racine, mais a un vocabulaire moins large, on peut en déduire la plus grande richesse de Racine. En revanche, entre "Corneille" et "Racine", on ne peut trancher. Charles Bernet (1983, p. 104-106) a montré, en utilisant le modèle de l'accroissement théorique, que le vocabulaire des tragédies de Corneille était nettement plus riche que celui des tragédies de Racine. Mais étant donné les incertitudes sur les biais liés à ce modèle, nous ne nous aventurerons pas à conclure pour le moment.

Pour des corpus de taille égale, on peut utiliser d'autres indicateurs comme la fréquence moyenne d'occurrence ou la médiane. Si l'on considère les corpus "Les portraits" et "Racine" (qui sont de taille à peu près similaire), on est surpris de constater la divergence des résultats. La moyenne est plus élevée pour "Les portraits" que pour "Racine", alors que la médiane l'est plus pour Racine. Ces indicateurs ne suffisent plus pour comparer les textes, il faut étudier les distributions de fréquence dans leur totalité et non à travers des indicateurs synthétiques.

1.2. LA DISTRIBUTION FRÉQUENTIELLE DES MOTS SUR DE LARGES CORPUS VÉRIFIE-T-ELLE LA LOI DE ZIPF ?[1]

Suite à de nombreuses "expériences de laboratoire" réalisées depuis le début du siècle sur des corpus divers, dans différentes langues, Ch. Muller (1977) montre que dans "l'exercice normal du langage", les distributions de fréquence obéissent à certaines constantes :

- Quand la fréquence croît, les effectifs décroissent. Ce fut la grande découverte de Zipf en 1936 qui peut nous paraître aujourd'hui évidente :

  > Nous nous trouvons devant un phénomène d'une évidence frappante, à savoir que si le nombre d'occurrences augmente, le nombre des différents mots présentant ce nombre d'occurrence décroît.
  >
  > (Zipf, 1974, p. 42)

- Les fréquences les plus faibles sont toutes représentées avec des effectifs élevés, alors que quand les effectifs deviennent plus faibles, certaines fréquences ne sont pas représentées.
- Quand N (la taille du texte) augmente, V (la taille du vocabulaire) croît, mais moins vite que N et de moins en moins vite.

Les régularités observées dans les distributions de fréquence laissent penser qu'elles obéissent à une loi valable quelle que soit la taille du texte et du vocabulaire et la langue utilisée. Le modèle doit pour Muller (1977, p. 95-96) comprendre au moins trois variables :

- une variable de discours liée à l'étendue du texte ;
- une variable de langue liée à l'étendue du vocabulaire ;
- une variable idiomatique.

Dès le début du siècle, en étudiant des distributions fréquentielles, G. K. Zipf (1936, 1974) observait certaines des régularités propres aux gammes de fréquence et proposait deux types de modélisation de ce phénomène.
Il s'appuyait pour cela sur les recherches de Kaedings tirées du chinois et du latin de Plaute, et sur celle d'Elridge sur des articles de presse américains qui présentaient la distribution fréquentielle des mots dans leur flexion complète sur des corpus provenant de différentes

[1] Je tiens à remercier Patrick Babayou, Aude Collerie de Borely et Pascale Hébel du département Prospective de la Consommation du CRÉDOC pour leurs précieux conseils.

langues (chinois, latin, anglais) et de tailles différentes. Sur ces trois corpus différents, en représentant, sur du papier logarithmique double, le nombre de mots en abscisse et la fréquence en ordonnée, il montre qu'en dehors des mots ayant une fréquence très élevée, la courbe de régression a pour équation $ab^2 = k$, où a représente le nombre de mots d'une fréquence donnée (ou l'effectif de la classe de fréquence) et b la classe de fréquence. Cette relation n'est valable que pour les faibles fréquences qui représentent cependant l'essentiel des vocables utilisés.
Dans la suite de son article, Zipf proposait une seconde manière d'aborder le problème, qui lui avait, dit-il, été suggérée par un ami et qui avait pour extrême avantage apparent d'éliminer un des biais pour la comparaison des corpus : la taille. C'est cette seconde version de la loi de Zipf, qui revient à dire que le produit du rang (r) (mots classés par fréquence décroissante) par la classe de fréquence (f) est à peu près constant, qui a été le plus largement commentée par la suite (Guilbaud, 1980). Cette loi qui s'exprime sous la forme :

$$f \times r = cste$$

a le désavantage important de mal s'adapter aux fréquences les plus basses comme aux fréquences les plus élevées.

De nombreuses recherches ont été menées sur la loi de Zipf. Les normes de dépouillements n'ont pas toujours été les mêmes, tantôt les études s'appuient sur une formulation (classe de fréquence), tantôt sur l'autre (rang de fréquence), tantôt on présente des fréquences cumulées, tantôt non... La simple description statistique du phénomène —ne parlons pas de son sens—, sur des échantillons nombreux et varié est une tâche plus délicate qu'il n'y paraît.—

Ici nous allons nous pencher modestement sur Zipf première manière, c'est-à-dire sur la loi qui lie classe de fréquence et nombre d'occurrences.

Cette loi, $ab^2 = k$, mérite d'être examinée de plus près car elle est sans doute trop générale et mal adaptée à tous les corpus. Connaître ses variations et ses domaines d'application permettrait de construire des indices synthétiques utiles, par exemple pour la comparaison des corpus, au repérage de corpus atypiques sur le plan de la distribution lexicale...

Nous cherchons à voir si sur nos corpus cette même loi s'applique, quelles que soient la taille et la nature du corpus. Bien sûr, il ne s'agit que d'un début de vérification, il faudrait soumettre une très grande quantité de textes de longueur et de contenu variables pour mener à bien cette vérification.

Déjà dans son texte, Zipf avait l'intuition que l'exposant 2 ne pouvait sans doute pas convenir à tous les corpus :

> On peut penser que l'exposant de b peut varier selon les différences de grandeur du volume examiné. Il semble incroyable qu'une fréquence de distribution présente invariablement l'exposant 2 pour b ; que l'on ait affaire à un volume de mille mots ou à un volume d'un million de mots.
>
> (Zipf, 1974, 1936, p.43)

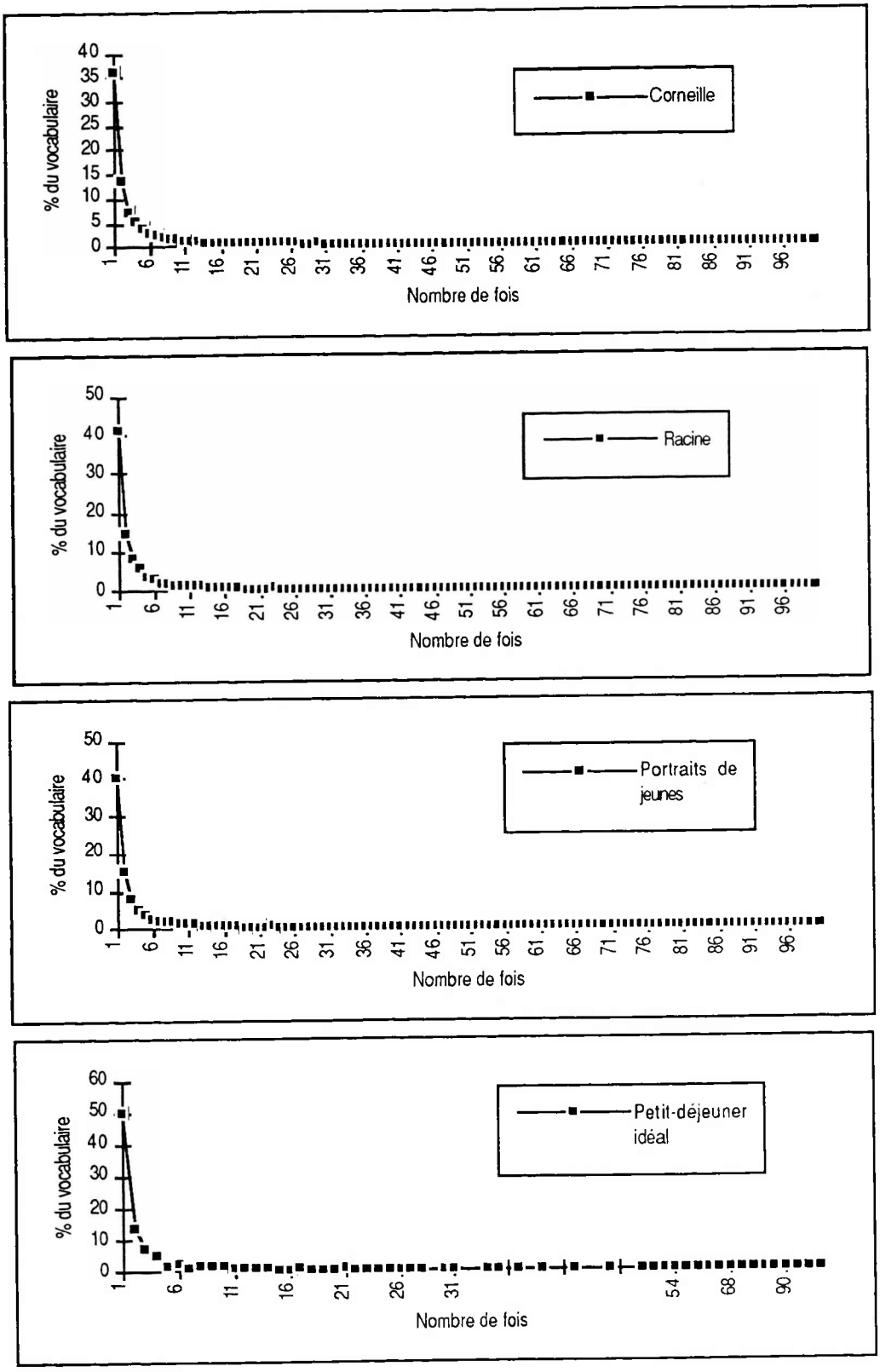
Nous voudrions voir comment varie l'exposant de b, quels sont les critères qui font varier la valeur de k, et surtout si cette relation s'applique sur n'importe quel type de corpus.

Nous avons donc calculé la distribution des fréquences sur les cinq corpus, toujours en utilisant le dépouillement d'Hyperbase. Les tableaux complets figurent dans l'annexe 1.

En représentant les courbes de répartition fréquentielle sur un graphique standard (fréquence 1 à 50), on ne peut que s'étonner de voir la similitude des courbes.

On n'a pas représenté les fréquences supérieures à 100 (et à 50 pour le corpus "petit déjeuner") car à partir du moment où les effectifs deviennent faibles, certaines fréquences manquent et la distribution des fréquence est "trouée".

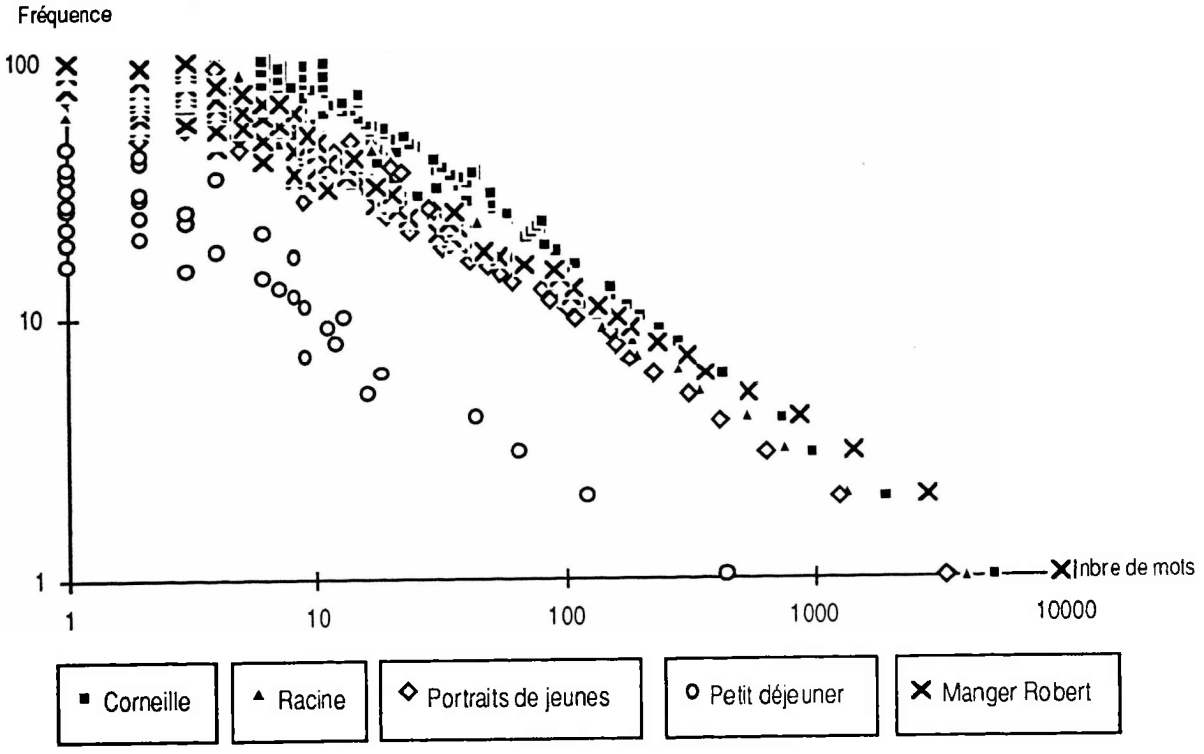
Figure 2 : Distribution des fréquences sur les quatre corpus



Ces courbes méritent une analyse plus précise. Dans son ouvrage, Zipf analysait les fréquences de 1 à 45, nous reprendrons donc ces mêmes classes de fréquence sur nos corpus.

On représente la répartition fréquentielle des cinq corpus sur une double échelle logarithmique qui laisse clairement deviner les droites de régression qui s'ajusteraient le mieux à ces différents nuages de points. Nous n'avons retenu que les fréquences de 1 à 100 pour les trois premiers corpus, et de 1 à 50 pour le Petit déjeuner. Au-delà, on s'écarte de plus en plus du modèle linéaire (on voit déjà que plus on s'éloigne de la partie droite du graphique, moins les points sont ordonnés).

Figure 3 : Distribution des fréquences sur les cinq corpus (double logarithme)



Clef de lecture : pour le corpus Petit déjeuner, on a 440 mots qui n'apparaissent qu'une fois.

Notons tout d'abord que plus le corpus est gros, plus les points sont proches de la droite et ce même pour des fréquences assez élevées. Alors que les points pour les réponses à une question ouverte s'écartent très tôt de la droite, ils sont beaucoup plus régulièrement répartis pour "Corneille" et ce presque jusqu'à la fréquence 100. On a aussi le sentiment que, à taille de corpus similaire, plus la taille du vocabulaire est élevée, plus les points sont alignés.

Dans "Racine" et "Les portraits", le nombre d'occurrences est sensiblement le même, mais "Racine" a un vocabulaire plus élevé et semble effectivement avoir une distribution plus régulière que "Les portraits".

D'après Zipf, la distribution fréquentielle vérifie la relation $ab^2 = k$ (où a représente le nombre de mots d'une fréquence donnée et b cette fréquence). Nous allons voir si la relation $\log b = -1/2 \log a + 1/2 \log k$ (équivalente à : $ab^2 = k$ par transformation logarithmique) s'applique à nos corpus.
On effectue sur les quatre corpus une régression pour évaluer l'équation de la droite et sa validité, en prenant comme données $\log a$ et $\log b$. Si la relation mise en évidence par Zipf est juste, on devrait trouver 0,5 comme coefficient de la droite.
On a réalisé pour chaque corpus plusieurs tests en faisant varier le seuil de fréquence.
Moins on conserve d'observations en diminuant le seuil de fréquence, plus la qualité du coefficient de corrélation au carré (R^2) s'améliore par un effet purement mécanique. Ce qui est plus intéressant, c'est que le modèle est mieux ajusté quand le corpus est plus grand et cela à nombre d'observations égal.
La variation du seuil de fréquence a de plus une incidence sur la pente de la droite.

Figure 4 : Régressions sur la distribution des fréquences : variation de R^2 en fonction des seuils

R^2	Toutes les fréquences	Fréq<100	Fréq<70	Fréq<50	Fréq<40	Fréq<30	Fréq<20
Corneille	0,6701	0,9475	0,9772	0,9834	0,9832	0,9848	0,9929
Portraits de jeunes	0,6846	0,9429	0,9477	0,9459	0,9688	0,9778	0,9943
Racine	0,6717	0,9253	0,9475	0,9679	0,9869	0,9886	0,9928
Manger (Le Robert)	0,7092	0,9613	0,9649	0,9802	0,9849	0,9924	0,9944
Petit déjeuner	0,6694	0,7661	0,8365	0,8375	0,8680	0,8824	0,8892

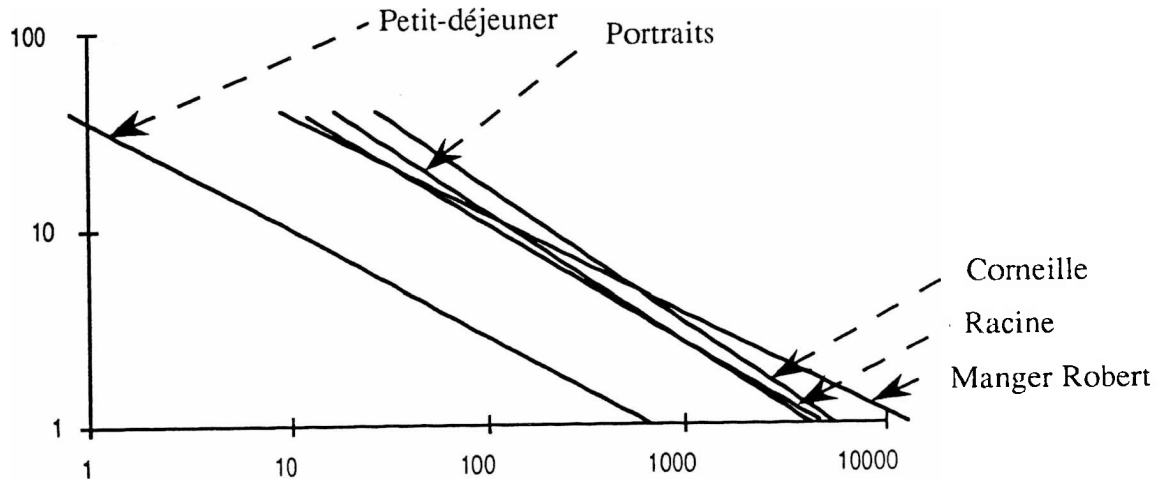
Figure 5 : Régressions sur la distribution des fréquences : variation de la pente en fonction des seuils

Coefficient de la droite	Toutes les fréquences	Fréq<100	Fréq<70	Fréq<50	Fréq<40	Fréq<30	Fréq<20
Corneille	-0,854152	-0,598690	-0,653149	-0,680775	-0,692335	-0,696473	-0,698520
Portraits de jeunes	-0,811001	-0,545886	-0,571357	-0,591513	-0,628250	-0,626531	-0,650253
Racine	-0,766460	-0,517465	-0,556526	-0,587458	-0,617662	-0,621419	-0,643377
Manger (Le Robert)	-0,714566	-0,493304	-0,493598	-0,508279	-0,507876	-0,533388	-0,545647
Petit déjeuner	-0,906966	-0,670500	-0,616788	-0,576891	-0,546375	-0,529412	-0,511166

Clef de lecture
Pour le corpus "Corneille", en faisant une régression sur l'ensemble des fréquences, la pente de la droite est de -0,854152 et le R^2 vaut 0,6701. Sur ce même corpus si on se limite aux fréquences inférieures à 20, la pente prend la valeur -0,698520.

On peut donc tracer sur un même graphique en double logarithme, pour les fréquences inférieures à 40, les différentes droites de régression correspondant aux cinq corpus.

Figure 6 : Droites de régression correspondant aux cinq corpus



Les tests associés à la régression ne sont pas interprétables parce que les résidus ne suivent pas une loi normale d'après le test de Shapiro-Wilk[1]. Cependant, en diminuant le seuil de fréquence pour la régression, les résidus s'alignent de mieux en mieux sur la loi normale.

[1] On teste l'hypothèse Ho : les résidus suivent une loi normale.
On est amené à rejeter l'hypothèse nulle dans tous les cas de régression, car la probabilité de rejet du test est toujours inférieure au seuil de fréquence que l'on s'est fixé (5 %).

La régression avec les moindres carrés ordinaires n'est licite que si la distribution est homoscédastique, or la distribution fréquentielle ne l'est pas de toute évidence. En effet, la variance est inversement proportionnelle à la fréquence : elle est beaucoup plus élevée pour les basses fréquences où le nombre d'occurrences est élevé que pour les fréquences les plus élevées (faible nombre d'occurrences).

Un résultat notable, si on se limite aux fréquences inférieures à 50, c'est que la pente de la droite varie en fonction de la taille du corpus. Alors que pour le corpus Petit déjeuner, la pente est proche de -0,54 (coefficient proposé par Zipf), elle est de -0,7 pour Corneille (le corpus le plus gros). De plus, la pente est quasiment identique pour les Portraits comme pour Racine qui sont des corpus de même taille. Le corpus tiré du dictionnaire "Manger - Le Robert" contredit cette affirmation : puisque la pente est égale à -0,5 alors que le corpus est nettement plus gros que celui du "petit déjeuner". La taille du corpus fait donc varier le paramètre, mais on ne sait pas très bien dans quel sens.
On peut émettre l'hypothèse que la distribution fréquentielle des mots dans un corpus vérifie une relation de type : $ab^p=k$, où p varie en fonction de la taille du corpus et peut-être en fonction de la taille du vocabulaire.

Nous avons l'impression tout de même que cette relation entre la fréquence et le nombre de mots se vérifie quel que soit le contenu du corpus. Ainsi, la relation semble s'appliquer aussi bien à des textes littéraires, des articles de dictionnaires qu'à un corpus de réponses à une question ouverte. En revanche, il est difficile de savoir quels sont les éléments (taille du corpus, du vocabulaire, contenu...) qui influent sur la forme de la courbe.
Tous ces résultats restent encore à être confirmés ou infirmés sur des corpus de taille plus conséquente.

Pour voir si la loi de Zipf peut permettre de caractériser la nature de certains corpus (d'un point de vue stylistique ou thématique), il faudrait reprendre les mêmes textes, constituer cinq corpus de taille similaire (en éliminant une partie des corpus les plus longs) et recalculer les gammes de fréquence. Ainsi, les paramètres des courbes de régressions obtenus pourront éventuellement jouer le rôle d'indicateur de la richesse lexicale d'un corpus.

Redisons-le encore, *la loi de Zipf première manière ne s'applique qu'aux basses fréquences* ; dès que la fréquence augmente, la distribution devient plus irrégulière et ne peut être estimée par un modèle log-linéaire. Donc cette loi n'approxime qu'une faible partie de la distribution. Bien d'autres modèles ont été proposés pour décrire la distribution fréquentielle (Waring-

Herdan (1964), Carroll (1967), Sichel (1975), Muller (1979), Orlov (1983)...). Nous sommes en ce moment à la recherche d'informations bibliographiques. Certains de ces modèles font intervenir un nombre redoutable de paramètres qui ne sont pas toujours faciles à estimer et qui ne s'adaptent pas très bien à la distribution fréquentielle. Harald Baayen, de l'Institut de Psycholinguistique Max Planck, à Nymegen aux Pays-Bas, considère que le modèle de Sichel est celui qui s'adapte le mieux aux données, mais que ses paramètres sont difficiles à interpréter (communication personnelle).

Suite aux travaux de Zipf, de très nombreux modèles ont été et continuent d'être proposés, mais comme le font remarquer Lebart et Salem (1994, p. 48) :

> La présentation de ces "lois théoriques" dont la formule analytique est souvent complexe, apporte peu de renseignements sur les raisons profondes qui sont à l'origine de ce phénomène.

La conclusion de Zipf reste toujours d'actualité :

> Le degré d'ordre dans la distribution des mots dans le flux du discours indique sans aucun doute une tendance à maintenir un équilibre entre la fréquence, d'une part, et ce que l'on pourrait appeler la variété, d'autre part.

Il semble donc que cette loi intervienne dans la production du discours, mais quel est le sens de ce processus de sélection des mots qui mène inconsciemment à cette distribution ordonnée ?

En tout cas, connaître la répartition fréquentielle des mots dans un corpus permet de mieux maîtriser le fonctionnement de l'analyse lexicale.

1.3. INCIDENCE SUR L'ANALYSE LEXICALE

Les basses fréquences représentent une proportion élevée du vocabulaire d'un corpus. Ainsi, les fréquences de 1 à 3 représentent pour Racine 63,7 % du vocabulaire et 5,3 % des occurrences, pour "Les portraits" 63,3 % du vocabulaire et 4,5 % des occurrences. Or pour l'analyse statistique qui compare les profils lexicaux d'unités textuelles, ces basses fréquences sont insignifiantes voire aberrantes (c'est le cas pour les hapax : n'ayant qu'une occurrence, ils n'entrent dans aucun système de comparaison). Elles sont d'ailleurs éliminées d'emblée. Ainsi, une bonne partie du vocabulaire utilisé passe aux oubliettes. Sachant que le seuil minimal au-delà duquel on conserve les vocables est de quatre dans Alceste, on évalue bien l'appauvrissement que l'on fait subir au texte.

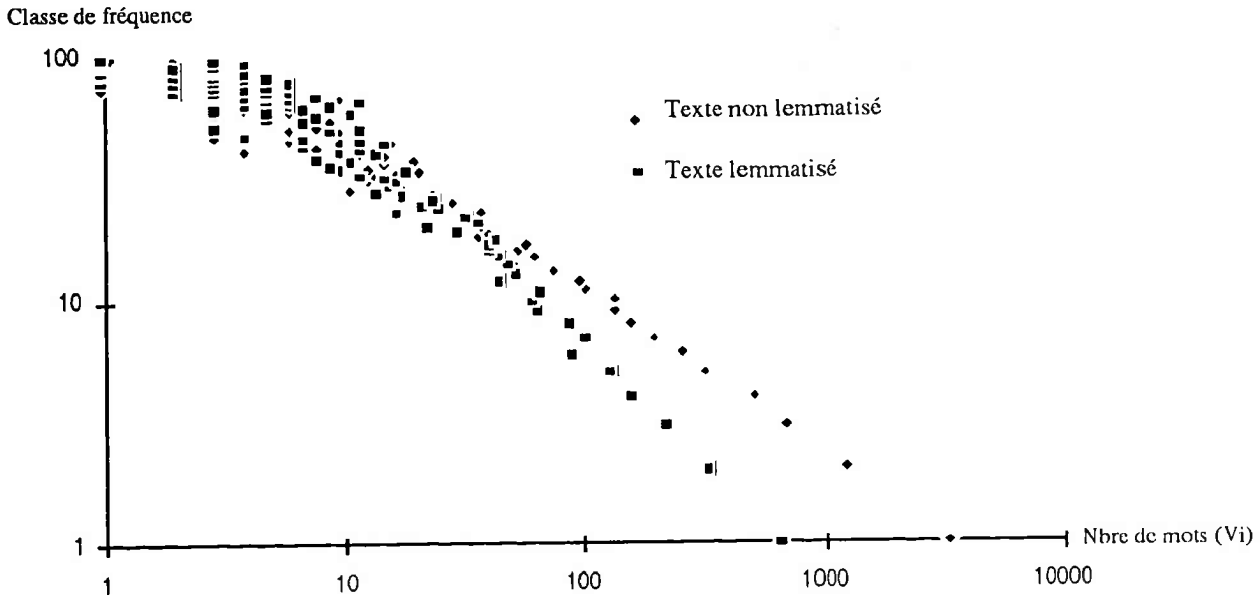
C'est entre autre pour éviter ces troncations drastiques, que les méthodes de lemmatisation ont été utilisées par certains courants dans la statistique textuelle. Mais la lemmatisation est-elle en mesure de réduire de manière significative les effectifs des basses fréquences ? Nous nous sommes livrés à une petite expérience sur les tragédies de Racine en comparant la gamme de fréquence avant et après lemmatisation. Nous avons constitué la gamme de fréquence sur le corpus non lemmatisé des onze tragédies avec le module d'Hyperbase précédemment utilisé et repris la gamme de fréquence constituée par Charles Bernet (1983) à partir du texte lemmatisé. Les normes de dépouillement et de lemmatisation sont clairement explicitées dans son ouvrage.

Tragédies de Racine	N	V
Dépouillement non lemmatisé	156 951	8 620
Dépouillement lemmatisé (Ch. Bernet)	158 899	3 262

Première observation, on a un écart de deux mille mots entre les deux dépouillements, qui est sans doute dû au fait que pour le dépouillement avec Hyperbase, nous somme partis d'un corpus de tragédies où tout le paratexte (numéros de scènes et d'actes, noms de personnages dans les tours de parole, didascalies...) a été éliminé.

Le graphique ci-dessous représente la distribution des fréquences de 1 à 100 avec et sans lemmatisation.

Figure 7 : Distribution des fréquences de 1 à 100 avec et sans lemmatisation



On voit clairement qu'une lemmatisation morpho-syntaxique (qui ramène les flexions à la forme canonique : verbes à l'infinitif, adjectifs au masculin singulier, noms au singulier...), a comme effet de réduire de 40 % la taille du vocabulaire et de réduire principalement les effectifs des faibles fréquences (approximativement jusqu'à la fréquence 10). On passe par exemple de 3 369 hapax à 680.

Fréquence	Dépouillement non lemmatisé (Hyperbase)	Dépouillement lemmatisé (Ch. Bernet)
f_i	V_i	V_i
1	3 369	680
2	1 269	339
3	721	230
4	528	164
5	337	134
6	266	92
7	203	105
8	163	90
9	142	66
10	140	62
11	104	67
12	100	45
13	79	54
14	53	50

Grâce à la lemmatisation, la taille du vocabulaire a été réduite de 40 %. Les fréquences de 1 à 3 ne représentent plus que 38 % du vocabulaire contre 63,7 % avant lemmatisation.

Dans les méthodes d'analyse statistique, en fixant un seuil de fréquence en deçà duquel les mots ne sont pas analysés, on élimine une partie importante du vocabulaire, mais nettement moins importante quand on procède à la lemmatisation.
En travaillant sur un corpus de moins de 5000 vocables, un seuil de fréquence de 4 est suffisant pour que puisse être effectuée une analyse, c'est-à-dire qu'en moyenne on supprime 40 % du vocabulaire sur un corpus, après lemmatisation. Mais quand on passe à des corpus de taille supérieure, des impératifs techniques de la classification obligent à augmenter le seuil de fréquence et donc à réduire de manière très conséquente le pourcentage du vocabulaire analysé.

En effet, la classification descendante d'Alceste travaille sur des tableaux de données qui croisent les unités de contexte avec les mots du vocabulaire. A l'intersection d'une ligne et d'une colonne, on a un "zéro" si le mot n'apparaît pas dans l'unité de contexte et un "un" s'il y apparaît une fois ou plus. Or pour la classification, ce tableau en question ne peut contenir plus de 50000 "un" dans la version actuelle de l'algorithme. Quand le nombre de "un" est supérieur, le logiciel augmente automatiquement le seuil de fréquence en deçà duquel les formes ne sont pas analysées.

Quand on travaille avec le logiciel Alceste, sur des corpus comme "Racine" ou "les portraits" (corpus de plus d'un million d'octets), pour arriver à un nombre de "un" acceptable, le logiciel augmente le seuil de fréquence jusqu'à 50, ce qui veut dire que plus de 90 % du vocabulaire du corpus n'est pas analysé ; ceci a une répercussion directe sur le pourcentage d'unités classées en fin d'analyse.

La version actuelle d'Alceste n'est pas tout à fait adaptée à l'analyse de très gros corpus. Nous attendons avec impatience la version 3. Pour le moment, plusieurs solutions se présentent :

- augmenter la taille des unités de contexte analysées ;
- éliminer d'emblée des catégories de mots de l'analyse : les modalisateurs, les mots-outils, les noms propres...
- faire de l'échantillonnage sur les corpus, même si cette procédure ne s'avère pas toujours très satisfaisante.

En veillant à tenir compte des informations sur la structure lexicale, nous avons cherché à étudier, avec les méthodes de statistique textuelle, la richesse lexicale des pièces de Corneille et Racine tant au niveau de la structure que du contenu.

2 - STYLISTIQUE ET ANALYSE LEXICALE : CORNEILLE ET RACINE

On se propose de montrer, à travers l'analyse d'oeuvres théâtrales classiques, la complémentarité des deux branches de la statistique textuelle (probabilités et analyse des données) pour l'analyse de corpus textuels. Cette étude aura pour second objet de montrer quelques apports de l'analyse lexicale à la stylistique.
Le corpus analysé est constitué des oeuvres théâtrales complètes de Corneille et Racine (Corneille : 33 pièces dont 21 tragédies, 9 comédies, 3 comédies héroïques ; Racine : 11 tragédies, 1 comédie). Ce corpus nous a été confié par Jacques Roubaud dans le cadre de recherches sur la métrique et le rythme.

2.1. OUTILS DE STATISTIQUE TEXTUELLE : ALCESTE ET HYPERBASE

Dans l'histoire de la statistique textuelle, deux branches peuvent être distinguées : l'une est née dans les années 50 dans l'Est de la France à Strasbourg, au moment où la grande entreprise du Trésor de la Langue Française démarrait. L'INALF, dans le cadre de la rédaction du TLF, eut comme projet de constituer une base de textes informatisés, Frantext, où pourraient être puisées des citations pour illustrer les articles du dictionnaire. Parallèlement, à cette époque, Charles Muller (1967) se lança dans l'analyse lexicométrique des oeuvres de Corneille. Pour simplifier, la démarche statistique adoptée revient à comparer les données observées aux données calculées à partir d'un modèle théorique.
Implicitement, il y a l'idée que le texte analysé est un échantillon représentatif de la langue comme le rappelle Maurice Tournier (1980), et que par l'étude de ce corpus, on pourra inférer des informations sur la langue. Ainsi dans ce cadre compare-t-on la sous-fréquence observée dans une sous-partie d'un corpus à la sous-fréquence théorique (calculée à partir de la fréquence du mot dans l'ensemble du corpus) et on parvient à la notion d'écart par rapport à une norme. Plutôt qu'une norme interne (une sous-partie par rapport à l'ensemble), on peut

choisir une norme externe comme celle que représente le TLF (fréquence de tous les mots présents dans la base textuelle Frantext) ou une norme du français courant... Dans cette approche probabiliste de la lexicométrie où la constitution d'un texte est assimilée à des tirages dans une urne, le modèle est construit de façon empirique à partir de données, provenant soit du corpus dans son ensemble, soit d'un corpus externe. L'introduction des probabilités est presque artefactuelle et rend finalement assez fragile la distinction entre les deux branches.
Dans cette branche, les recherches ont principalement porté sur la richesse, la spécificité, l'accroissement et l'évolution chronologique du vocabulaire... en ayant toujours comme point de comparaison un modèle théorique.

L'autre branche a été animée par Jean-Paul Benzécri (1981, 1982), père de l'analyse des données à la française. Ayant mis en place tout un arsenal d'outils d'analyse des données multidimensionnelle, il les applique aux données particulière que sont les textes. L'objectif de Benzécri était d'ailleurs l'analyse des textes et des données linguistiques.

> L'analyse des correspondances a été initialement proposée comme une méthode inductive d'analyse des données linguistiques
>
> (Benzécri, 1982 cité par (Reinert, 1993)).

Partant de la constatation que, dans les modèles probabilistes, les hypothèses sont rarement satisfaites, Benzécri propose donc des méthodes inductives qui font émerger un modèle à partir des données.

Ces deux branches se sont développées parallèlement, chacune avec ses propres publications, ses propres collections[1], des échanges nombreux puisque certains chercheurs sont à la frontière de ces deux disciplines, que des logiciels d'une branche intègrent des modules de la seconde (Hyperbase (première branche) a un module d'analyse factorielle, Spad.T et Alceste (seconde branche) utilisent le calcul des spécificités pour caractériser les classes de discours...) et que les colloques actuels réunissent les deux communautés.

Pour analyser nos corpus, on utilisera ici deux méthodologies statistiques complémentaires : le logiciel Alceste conçu par Max Reinert (CNRS, Toulouse) (Reinert, 1983, 1987, 1992) qui

[1] Travaux de linguistique quantitative, publiées sous la direction de Charles Muller, Slatkine-Champion ; Les cahiers de l'analyse des données, publiées sous la direction de Jean-Paul Benzécri, Dunod.

est issu de la branche "Analyse des données" et Hyperbase conçu par Etienne Brunet (INALF, Nice) (Brunet, 1992, 1978) qui relève plutôt de la branche "probabiliste"[1].
D'un côté, on a une approche descriptive qui utilise les outils de l'analyse des données multidimensionnelles reposant sur les associations de mots dans des fragments de textes et de l'autre une approche probabiliste qui explicitement utilise le modèle de l'urne pour décrire les phénomènes lexicaux.
Les deux approches diffèrent par leurs présupposés (identification des écarts par rapport à un modèle théorique pour Hyperbase, recherche des univers lexicaux pour Alceste) mais aussi par la définition des unités d'analyse.

Hyperbase est plutôt conçu pour analyser des unités "naturelles" ou macrostructurales : une pièce, un ensemble de pièces (les comédies de Corneille, toutes les pièces de Racine...). Il permet de mettre en évidence les termes qui différencient une oeuvre ou un ensemble d'oeuvres d'un cadre de référence donné. Ce cadre de référence peut être interne ; on compare alors une sous-partie du corpus au corpus dans son ensemble (par exemple les comédies de Corneille par rapport à l'ensembles de ses oeuvres). Mais il peut aussi être externe, on compare alors le corpus à un ensemble de textes censé jouer le rôle de norme. Cette approche se situe dans la tradition de la stylistique de l'écart qui considère que le style apparaît par différence avec la norme, ou d'une manière moins normative, avec le cadre de référence.
L'intérêt d'Hyperbase est de pouvoir avoir recours à une norme interne (représentée par l'ensemble du corpus analysé), mais aussi à une norme externe. On peut en effet comparer les textes dont on dispose à un extrait de la base de textes littéraires Frantext. Cette partie de Frantext est constituée des oeuvres "majeures" du XIXe et XXe siècles et a servi de base d'exemples pour la réalisation du Trésor de la Langue Française. Bien sûr, dans notre cas, le TLF n'est pas la meilleure référence possible puisque l'on compare l'écriture du XVIIe à celle du XIXe. On obtiendra autant des différences de langue que des différences liées à des auteurs.

Alceste découpe les textes en segments homogènes (unités de contexte élémentaires ou UCE) de longueur assez faible (quatre alexandrins en moyenne) qui sont tous indicés par la pièce d'origine, la date de publication, l'auteur et le genre. Chaque texte (dans notre cas, chaque pièce) est considéré comme un ensemble de segments de texte ou d'unités de contexte (UCE). Celles-ci sont elles-mêmes des combinaisons d'unités plus petites : les mots, qui jouent le rôle de variables dans l'analyse statistique.

---

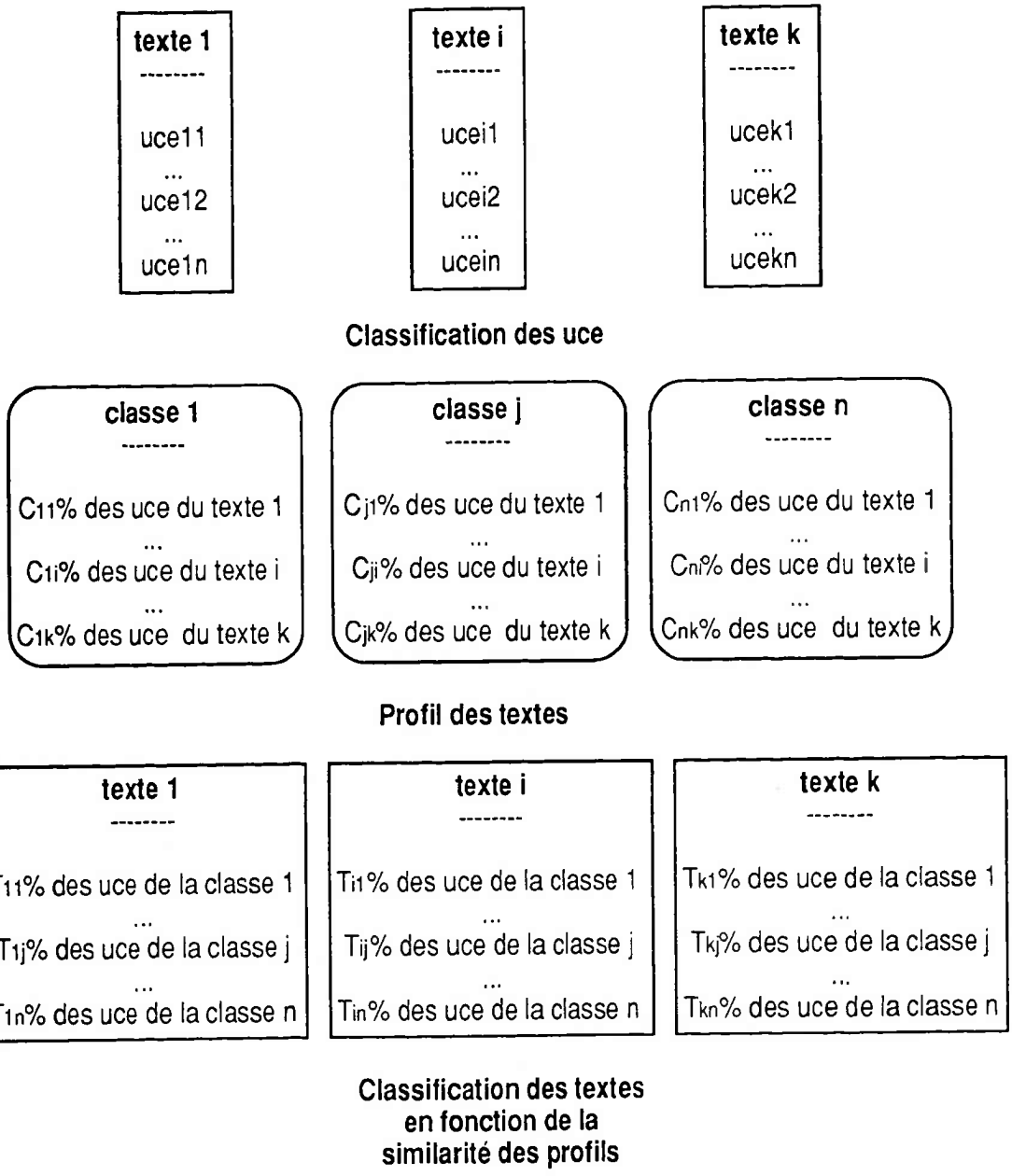
[1] Nous ne prétendons pas explorer sur nos corpus tous les modules d'analyse des deux logiciels, mais présenter les fonctions centrales de chacun d'eux.

L'unité des pièces est rompue au cours de l'analyse, puisque ce sont les profils lexicaux des UCE (et non pas des pièces) qui sont examinés. Les mots, par la manière dont ils apparaissent ou non dans les UCE, permettent de classer les unités de contexte. On a au départ un ensemble d'UCE qui, grâce à une méthode de classification descendante, est segmenté en deux classes, de telle sorte que chaque classe soit aussi homogène que possible d'un point de vue lexical, et le plus distincte de l'autre. Le processus de classification est itératif et conduit à une série de classes constituées. En fin de compte, chaque unité de contexte appartient à une classe[1].

Cette approche permet de reconstituer des *visions du monde* (Reinert, 1987) ou des *noyaux de sens* (Lahlou, 1994). Un noyau peut parfois coïncider avec un texte, mais généralement, un texte parcourt différents noyaux lexico-sémantiques. Il est alors constitué d'un certain pourcentage d'unités de contexte provenant de la première classe, d'un autre provenant de la seconde... Classer les textes (ici les pièces) revient à regrouper les objets qui ont un profil similaire (nous entendons par profil un vecteur de classes d'unités de contexte). Le graphe suivant résume la démarche.

[1] sauf les unités qui ne sont pas classées parce qu'elles emploient un vocabulaire trop marginal ou qu'elles sont "à cheval" entre deux classes.

Figure 8 : Méthode de classification en fonction des profils lexico-sémantiques



Ce processus en deux temps permet d'envisager une classification des textes en fonction de la similarité des profils lexico-sémantiques.

Avec d'autres outils d'analyse des données textuelles, comme SPAD.T, les textes peuvent être directement classés à partir de leurs profils lexicaux. Pour caractériser une classe, on aura une liste des termes spécifiques, classés par significativité décroissante. Il sera impossible d'inférer à partir de cette liste les différents modes d'organisation lexicale internes qui ont contribué à la constitution des classes. La démarche est certes plus économique que celle que

nous avons adoptée avec Alceste, mais elle est moins fine puisqu'elle ne permet pas l'analyse des microstructures ou énoncés à l'intérieur des textes.

Avec Hyperbase, on considère chaque pièce ou des regroupements de pièces, avec Alceste on travaille sur des unités plus petites censées ressembler aux énoncés. Ces deux approches sont donc complémentaires.

Sur nos corpus, nous utiliserons Alceste et Hyperbase pour répondre à plusieurs questions :

a) en quoi les registres de langue de la tragédie et de la comédie se distinguent-ils ? Ces différences prévalent-elles sur les différences d'auteurs ?
b) quelles sont les caractéristiques de l'oeuvre théâtrale de Racine, comment son écriture a-t-elle évolué au cours du temps ?
c) en quoi la tragédie classique se distingue-t-elle de la littérature prise dans son ensemble ? Quelles sont les différences ou ressemblances que l'on peut identifier entre les tragédies de Corneille et celles de Racine ?

L'intérêt de ces questions dépasse évidemment la stylistique cornélo-racinienne. Les textes classiques sont ici utiles comme exemples, comme tests :

a) pour mettre en évidence des constantes par-delà les idiosyncrasies ;
b) pour analyser la cohérence d'une source dans son expression.
c) pour comprendre les contraintes qu'impose un genre dans la production du discours.

2.2. TRAGÉDIES ET COMÉDIES

A l'époque classique, les différences de genres étaient extrêmement codifiées, en particulier l'opposition entre tragédie et comédie :

> (...) j'ai découvert cette vérité que je crois capitale : que la tragédie est le développement d'une action et la comédie d'un caractère.
>
> STENDHAL, Journal, p. 50.

La tragédie devait représenter une action tragique, fondée sur un grand sujet, la comédie présenter les travers et ridicules des caractères et moeurs des bourgeois pour divertir le public. Comment ces différences thématiques se traduisent-elles dans l'organisation lexicale ?

2.2.1. Deux univers impénétrables

En effectuant une classification descendante avec Alceste, sur l'ensemble des pièces de Racine ou sur des pièces de Corneille (l'intégralité des oeuvres dépasse les capacités de calcul d'Alceste), dès la première itération, les UCE sont réparties en deux classes, selon qu'elles proviennent des tragédies ou des comédies. Ce regroupement dans les catégories d'origine des fragments traduit une remarquable cohérence de genre.

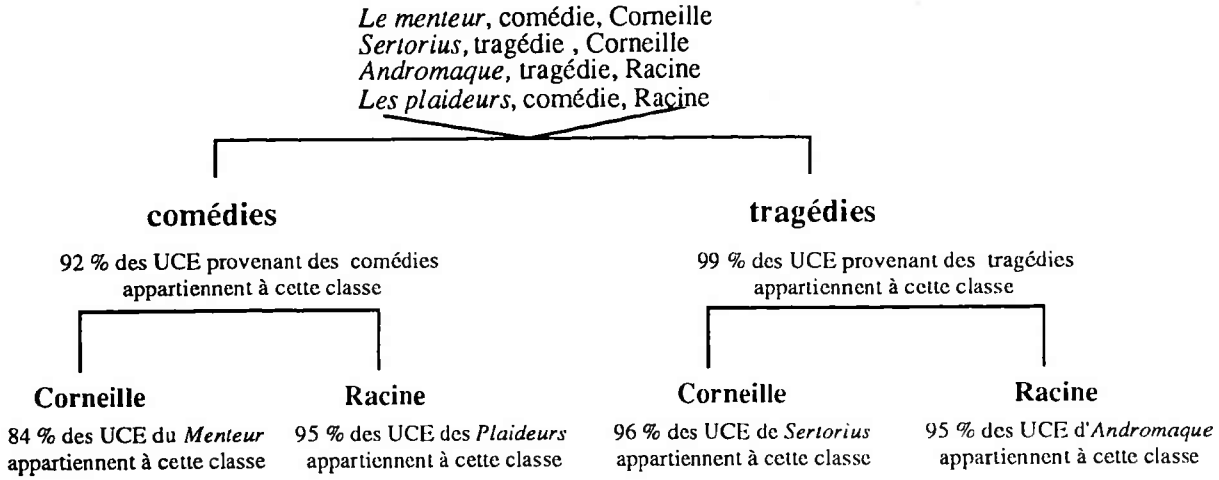
L'analyse effectuée avec le logiciel Alceste sur le corpus "Racine" isole l'unique comédie, *Les plaideurs*, de toutes les tragédies dans une classe au trois quart composée d'UCE provenant de cette comédie. On observe le même phénomène pour Corneille.

Les différences de genre sont même plus fortes que les différences d'auteurs. On a effectué un test en sélectionnant pour Corneille et Racine, une tragédie et une comédie. Le premier noeud de la classification descendante sépare les comédies des tragédies : la première classe contient en effet 99 % des UCE provenant des tragédies, la seconde 92 % des UCE des comédies. Cela signifie qu'à l'intérieur d'une pièce, il n'y a pas de variations de ton. Ceci confirme les différences de registre lexical qui distinguent ces deux genres :

> "La comédie doit prendre un ton moins haut que la tragédie : le socque est inférieur au cothurne."
>
> FÉNELON, OE., t. XXI, p. 221.

Le genre implique que tous les personnages quel que soit leur niveau social : princes, confidents ou domestiques, emploient le même registre de vocabulaire. On comprend mieux alors le caractère innovateur du drame bourgeois puis du drame romantique qui dans la même oeuvre associent le tragique au comique, le beau au laid, la noblesse au grotesque...

Figure 9 : Classification descendante sur quatre pièces



Les différences de genre, comme on l'a annoncé, prévalent sur les différences d'auteurs : c'est au second noeud de la classification seulement que les pièces de Corneille et de Racine se distinguent comme on peut le voir sur le graphique ci-dessus.

2.2.2. Caractéristiques lexicales des deux genres

Le logiciel Hyperbase a permis d'identifier chez Corneille et Racine les termes les plus caractéristiques de chacun de ces genres théâtraux. On a choisi une norme interne constituée par l'ensemble de la production théâtrale de ces deux auteurs. On compare donc les tragédies de Corneille, puis ses comédies, à l'ensemble de son oeuvre et enfin l'unique comédie de Racine à l'ensemble de sa production.

Le tableau suivant présente, pour chacun des cas, la liste des termes les plus caractéristiques, par spécificité décroissante. Dans la première colonne, on lit la fréquence du mot dans le sous-corpus (pour le premier cas, il s'agit des tragédies de Corneille), dans la seconde, sa fréquence dans l'ensemble du corpus et dans la troisième un indice de spécificité ou d'écart. Volontairement, nous avons supprimé les noms propres qui apparaissent dans les listes de spécificité.

Tragédies / comédies de Corneille

Fréq. trag.	Fréq. com. + trag.	Ecart	Mot
651	726	15	roi
847	1027	13	seigneur
582	674	12	sang
547	670	10	mort
484	603	9	gloire
298	348	9	prince
260	298	9	rois
166	172	9	tyran
339	408	8	haine
117	120	8	romains
274	319	8	vertu
269	328	7	bras
129	143	7	couronne
337	416	7	crime
400	508	7	dieux
314	395	7	fils
136	151	7	héros
875	1184	7	leur
455	590	7	main
164	183	7	peuple
230	273	7	trône
157	179	7	victoire
738	1019	6	aux
209	257	6	craindre

Comédies / tragédies de Corneille

Fréq. com.	Fréq. com. + trag	Ecart	Mot
137	154	13	monsieur
966	2011	11	tu
115	151	10	humeur
145	207	10	maîtresse
543	1058	10	te
63	73	9	amours
203	345	9	fort
1981	4704	8	?
127	196	8	ami
89	121	8	beauté
162	264	8	belle
136	224	8	discours
92	134	8	nuit
50	64	7	affaire
859	1947	7	bien
30	30	7	blanche
81	119	7	cela
237	441	7	esprit
28	28	7	marquis
1791	4262	7	me
27	27	7	paris
38	43	7	portrait
28	29	7	rire
1325	3170	6	!

Comédie Les plaideurs / tragédies de Racine

Fréq. plaid	Fréq. com. + trag.	Ecart	Mot
115	115	44	monsieur
32	38	21	bon
18	18	17	affaire
18	18	17	messieurs
18	18	17	procès
13	13	15	sergent
13	13	15	souffleur
12	13	13	exploit
23	45	13	là
207	1623	12	!
92	580	11	à
73	403	11	bien
13	21	11	fort
25	79	10	j
17	40	10	juge
11	19	10	juger
12	25	9	cela
9	14	9	chose
11	22	9	homme
98	742	9	on
15	35	9	voilà
37	186	8	hé
8	14	8	partie
7	13	7	maison
48	321	7	père

L'examen de ces listes montre que tragédies et comédies se distinguent par les modalités de l'interlocution et par les univers de référence.

2.2.1. Fonction conative du langage : l'interlocution

Les noms des personnages, qui n'apparaissent pas ci-dessus, sont les éléments qui distinguent le plus une pièce d'une autre. Ceux-ci sont de bons marqueurs lexicaux, ils apparaissent souvent : sans doute pour faciliter la compréhension des spectateurs, l'auteur se soucie-t-il de rappeler régulièrement le nom de ses héros.

A ce niveau apparaît crûment une des différences majeures entre comédie et tragédie : les écarts de registre. Pour Racine, le simple examen des noms de personnages permet d'identifier le genre de la pièce. Dans la tragédie racinienne, les noms de personnages, venus de la mythologie ou de l'histoire ancienne, sont savants et rares (*Néron*, *Porus*, *Achille*, *Titus*,

Pyrrhus, *Roxane*...). Même les confidents ont des noms aux sonorités gréco-latines (*Phénice*, *Phoenix*, *Oenone*...).
Dans l'unique comédie, *Les plaideurs*, les noms de personnages ont des connotations toutes différentes, bien plus communes (*Chicanneau*, *L'intimé*, *Léandre*, *Dandin*, *Petit-Jean*, *Isabelle*).

Pour Corneille, en revanche, les noms des personnages ne semblent pas être autant marqueurs de genre :

- Tragédies : Jason, Camille, Marcelle, Pauline, Placide, Attila, Cinna, Cléopâtre, Eurydice, Médée, Oedipe, Othon, Phocas, Polyeucte, Pompée, Rodrigue, Sertorius, Spitridate.

- Comédies : Dorante, Cliton, Clarice, Lyse, Philiste, Isabelle, Daphnis, Florame, Hippolyte, Lysandre, Tircis, Alcidon, Cloris, Dorimant...

Par ailleurs, dans l'interlocution, dans tous les éléments destinés à produire un certain effet sur le récepteur (fonction conative du langage, (Jakobson, 1963)), apparaissent d'autres marqueurs de l'orientation vers le destinataire qui distinguent les deux genres : les titres utilisés pour s'interpeller, et le vouvoiement/tutoiement. Chez Racine, les termes *Seigneur* et *Madame* sont presque exclusivement présents dans la tragédie. Dans la comédie les personnages s'interpellent plutôt par le prénom ou par *monsieur*. Chez Corneille, la ligne de partage entre comédie et tragédie se fait autour de *Seigneur*, *Prince / Monsieur* ou patronyme. Enfin, dans la comédie, on se tutoie beaucoup plus facilement que dans la tragédie.

Les types d'appellation propres aux deux genres traduisent la différence entre les univers sociaux : d'un côté les sommets du pouvoir royal ou princier (registre élevé), de l'autre le monde des bourgeois (registre commun). Et la situation sociale globale, caractéristique du genre, a un effet sur l'ensemble du personnel de la pièce. Ainsi, le registre de langage de la comtesse dans la comédie ne sera-t-il pas sensiblement différent de celui de *Chicanneau* ou de *Petit-Jean* et inversement dans la tragédie, les héros comme les suivants emploieront le même registre.
Tout se passe comme si le genre, parce qu'il impose de se situer à un certain niveau de l'échelle sociale, uniformisait les manières de parler.

D'autre part, le lieu et l'époque où se déroule l'action sont caractéristiques des genres. L'action dans les tragédies est lointaine aussi bien dans le temps que dans l'espace, comme si la distance spatio-temporelle était garante de la noblesse du genre. Les comédies se déroulent

plutôt à l'époque et souvent dans la ville. Ainsi, le nom de ville le plus spécifique des tragédies de Corneille est-il *Rome* (351 des 398 occurrences de *Rome* apparaissent dans les tragédies), tandis que celui des comédies est *Paris* (les 27 occurrences de *Paris* appartiennent aux comédies).
Il est à noter que ces spécificités ne sont pas a priori nécessaires à l'intrigue. Ce sont des traits qui activent un ensemble de connotations qui renforcent la tonalité de genre : grandeur pour la tragédie, familiarité pour la comédie.

2.2.2. Les univers sociaux et mentaux

L'examen avec Hyperbase des mots caractéristiques montre chez Corneille comme chez Racine l'opposition nette (et somme toute prévisible) des univers référentiels entre les univers tragiques et comiques.
Alors que dans la tragédie, il est question de pouvoir (*roi*, *seigneur*, *prince*, etc.), de *gloire*, de *haine*, de *vengeance* et de *mort*, dans la comédie il est question d'*humeur*, d'*amours* (pluriel qui rend trivial le sentiment), de *discours* et d'*affaire*. D'un côté le sérieux, la noblesse des sentiments et la constance, de l'autre le jeu, le rire et la légèreté. Tels sont les éléments qui sont *exclusivement* spécifiques de chacun des genres. Les marques exclamatives, ponctuations (!, ?) et interjections (*hé*), accentuent les différences de ton. Ce point est d'autant plus notable que les marques interrogatoires sont aussi très fréquentes dans la tragédie (cf. infra).

La tragédie est plutôt du côté du sentiment, de la délibération intérieure où s'affrontent sentiments contradictoires (amour et honneur,...). La comédie ancrée dans les préoccupations matérielles tourne autour d'affaires, d'argent... : l'ancrage dans la réalité sociale y est beaucoup plus marqué.
C'est aussi le registre du vocabulaire qui distingue les genres lorsque l'on se réfère à un objet identique : *hymen* dans la tragédie, *mariage* dans la comédie.

2.3. LES PIÈCES DE RACINE

Considérons le corpus constitué des oeuvres complètes de Racine, soit onze tragédies et une comédie. Il est constitué des douze pièces suivantes :

La Thébaïde ou les frères ennemis	tragédie	1664
Alexandre le Grand	tragédie	1665
Andromaque	tragédie	1667
Les plaideurs	comédie	1668
Britannicus	tragédie	1669
Bérénice	tragédie	1670
Bajazet	tragédie	1672
Mithridate	tragédie	1673
Iphigénie	tragédie	1674
Phèdre	tragédie	1677
Esther	tragédie tirée de l'écriture sainte	1689
Athalie	tragédie tirée de l'écriture sainte	1691

Nous proposons ici une série de manipulations qui peuvent s'appliquer à d'autres types de corpus et qui permettent d'avoir une vision globale et synthétique sur le contenu du corpus et de ses sous-éléments et sur l'évolution temporelle. Nous utilisons ici encore les méthodologies Hyperbase et Alceste.

2.3.1. Les spécificités lexicales : vers un résumé automatique des textes

Pour calculer les spécificités lexicales de chaque pièce, le corpus a subi quelques traitements préalables. On ne conserve à l'intérieur du corpus que le titre et les alexandrins qui sont prononcés au cours des représentations. Les informations sur les numéros d'actes, de scènes et sur les noms de personnages qui prennent la parole ainsi que les didascalies n'interviennent pas dans les analyses.

La méthode des spécificités mise au point par Charles Muller (1967, 1977) et reprise dans le logiciel Hyperbase d'Étienne Brunet (INALF, Nice) permet d'obtenir rapidement les termes

spécifiques de chaque pièce par rapport à l'ensemble du corpus. Les calculs de spécificités sont établis sur l'ensemble du vocabulaire non lemmatisé, y compris les noms propres. Dans ces listes de mots caractéristiques, les mots sont classés par spécificité décroissante. Pour chaque pièce les mots les plus spécifiques sont les noms de personnages ce qui montre un renouvellement complet des personnages. Les termes spécifiques constituent une forme de résumé de la thématique de chaque pièce, non pas dans son déroulement mais à travers les thèmes abordés comme on peut le voir dans l'annexe 2.

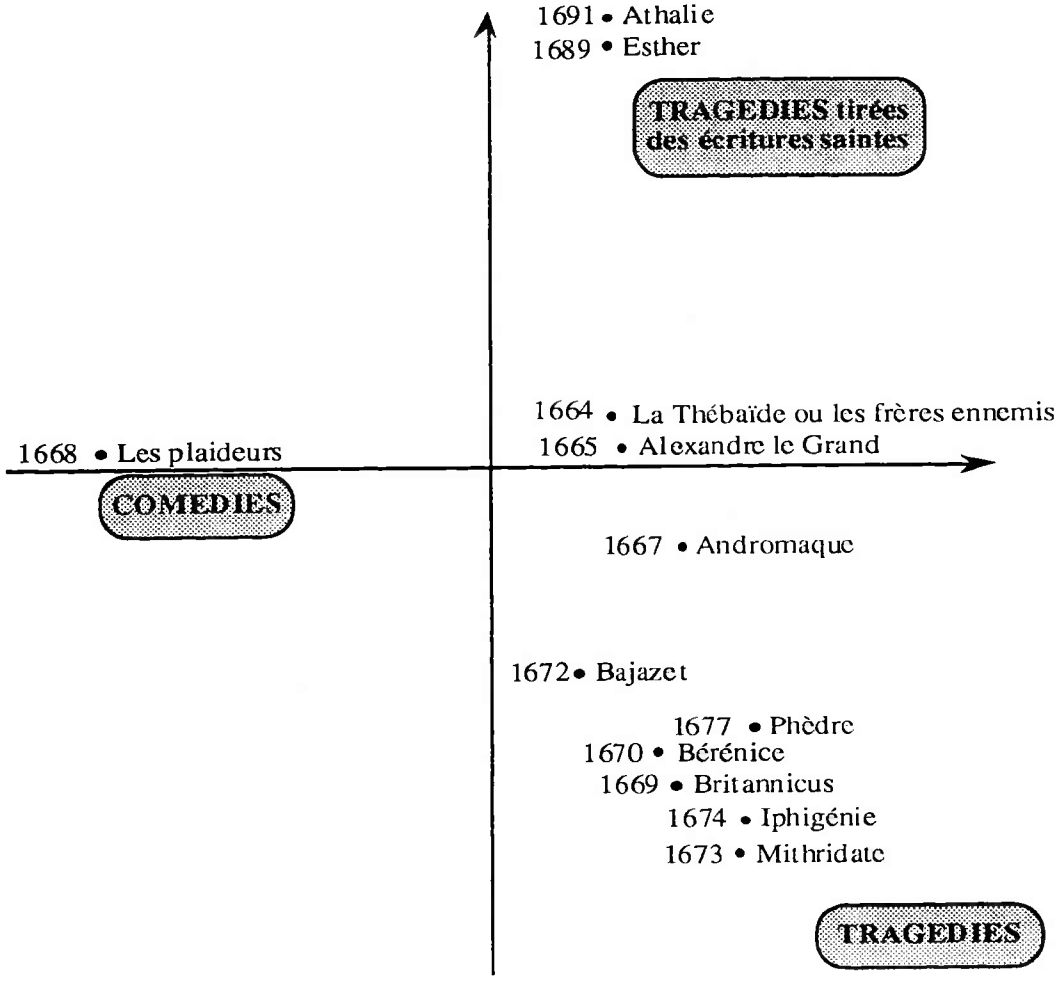
Charles Bernet (1983, p. 315-359), pour mettre en évidence le vocabulaire caractéristique de chaque pièce, s'était appuyé sur le vocabulaire lemmatisé sans tenir compte des noms propres. Outre le fait que les noms propres n'apparaissent pas dans les vocables spécifiques qu'il propose, on retrouve grosso modo les mêmes termes spécifiques, mais ils ne sont pas toujours classés dans le même ordre. Ainsi, *monter*, qui apparaît en 7ème position dans la liste de Bernet, n'apparaît qu'en 72ème (non compris les noms propres) dans celle que fournit Hyperbase. Ce glissement est dû au fait que toutes les autres formes conjuguées de *monter* ont été comptabilisées séparément.

Le calcul des spécificités sur le texte lemmatisé permet de recenser plus rapidement les thématiques abordées. Les spécificités sur les textes non lemmatisés donnent des informations complémentaires sur les modes de l'énonciation : temps et modes employés, pronoms personnels, genres grammaticaux... Ainsi voit-on pour *La Thébaïde*, le verbe *être* apparaît comme un verbe très spécifique de cette pièce. L'examen de la liste fournie par Hyperbase, montre que ce sont principalement les temps du futur et du conditionnel qui sont spécifiques (*serait*, *seront*).

2.3.2. Evolution des univers sémantiques

Sur l'ensemble des oeuvres de Racine, nous avons effectué une première analyse en conservant le texte intégral des versions éditées (avec toutes les informations scénographiques : noms de personnages, didascalies...). Le graphe suivant qui résulte d'une analyse factorielle établie sur la base de la classification, présente la projection des pièces dans l'espace factoriel.

Figure 10 : Projection sur le premier plan factoriel des oeuvres de Racine (analyse 1)



Cette première analyse à l'aide d'Alceste, en considérant tous les mots pleins comme des variables actives, conduit à des regroupements de pièces qui font sens d'une manière remarquable et recoupent la segmentation historique et critique.

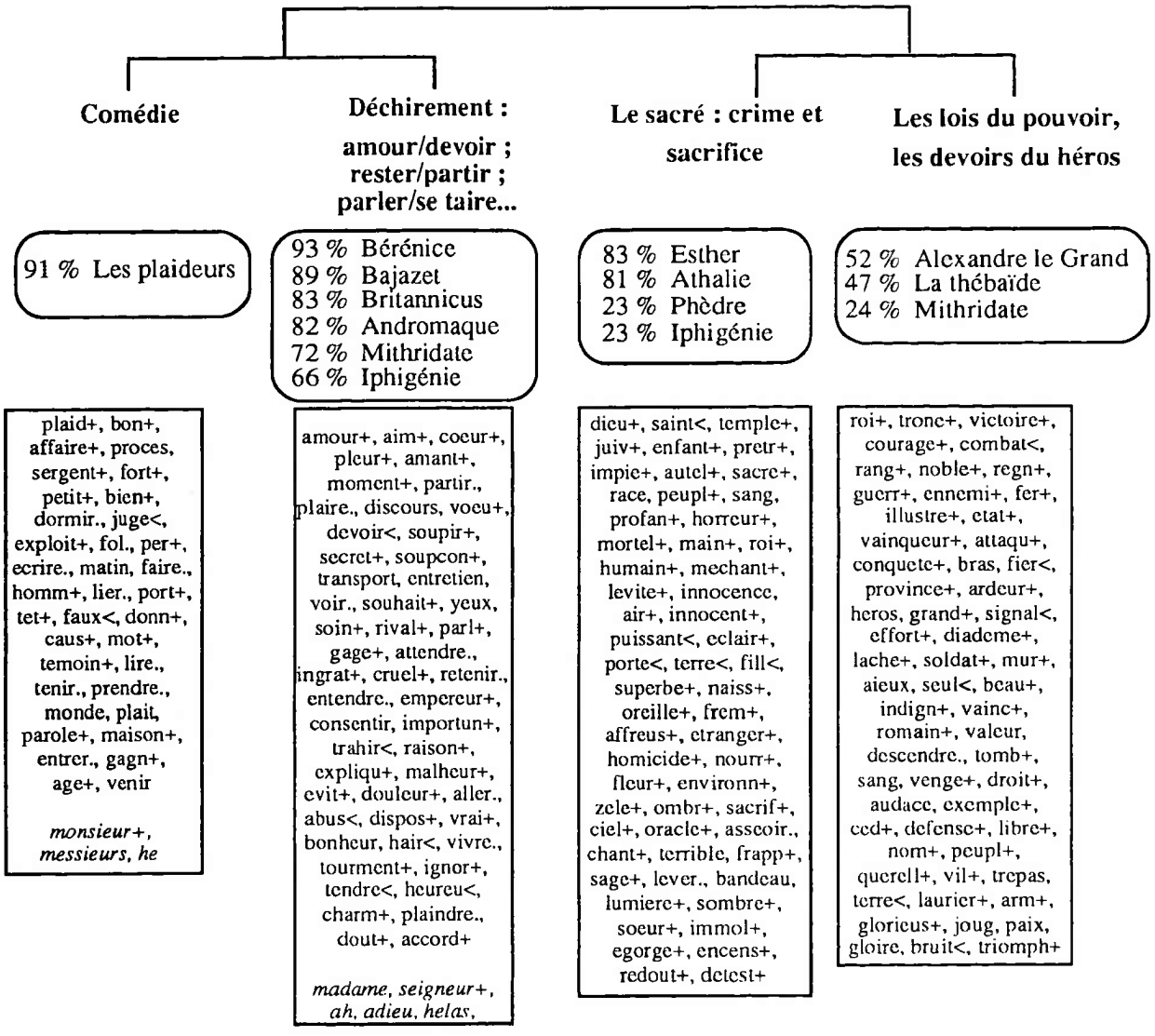
La classification distingue tout d'abord l'unique comédie *Les plaideurs* des tragédies. Ces dernières se répartissent en trois grands ensembles de pièces qui correspondent à trois périodes de l'écriture de Racine. L'un regroupe *La Thébaïde* et *Alexandre le Grand*, un autre toutes les tragédies d'*Andromaque* à *Phèdre* (*Bajazet* occupe une place spéciale dans cet ensemble) et enfin un troisième ensemble regroupe *Esther* et *Athalie*. Cette classification effectuée à partir des données lexicométriques met en évidence l'évolution du style de Racine et de ses sources d'inspiration. Les deux premières tragédies sont encore très marquées par l'influence de Corneille (Bénichou, 1948 ; Benzécri 1981) qui jouissait alors d'une grande

audience. Ce n'est qu'à partir d'*Andromaque* que s'affirme avec le plus de vigueur une vision du monde et du tragique très personnelle. *Andromaque* semble jouer un rôle de transition dans l'écriture de Racine, puisqu'elle se situe entre *La Thébaïde* et *Alexandre* et les pièces centrales (de *Britannicus* à *Phèdre*). Enfin, après un silence de douze années, sur commande, Racine écrit ses deux dernières tragédies où la source d'inspiration n'est plus l'histoire antique grecque ou romaine (voire turque) mais l'écriture sainte.

Le fait que les noms de personnages soient traités comme des variables actives a accru fortement la cohérence de chacune des classes. Nous avons donc effectué une seconde analyse sur les oeuvres de Racine dans laquelle toutes les informations scéniques ont été éliminées : l'analyse porte exclusivement sur les vers, sur les paroles effectivement prononcées sur scène. D'autre part, nous avons constitué la liste des noms propres (noms de personnages et noms de lieux) employés dans les vers, et nous les avons exclus de l'analyse. En effet, les thématiques théâtrales devaient être identifiées par delà les phénomènes d'actualisation que représentent l'ancrage en un lieu spécifique et les noms donnés aux personnages. Le lieu et le nom des personnes sont assez anecdotiques, ils apportent une forme de couleur locale aux pièces, et permettent surtout de les distinguer les unes des autres comme nous avons pu le voir à travers le calcul des spécificités.

On effectue une classification descendante sur un tableau qui croise les mots lexicaux lemmatisés et des fragments des pièces qui comprennent au moins 20 mots actifs. Nous avons vu dans la première partie que quand la taille du corpus est grande, il fallait constituer des unités textuelles assez longues pour que la plus grande partie du vocabulaire soit analysée. En choisissant des unités comprenant plus de 20 mots, on retient tous les mots ayant une fréquence supérieure à 20, ce qui est moyennement satisfaisant en soi. Lors d'une première analyse, en prenant des unités de taille plus petite, nous ne pouvions retenir que les fréquences supérieures à 45, ce qui était encore moins satisfaisant.
Cette analyse permet de classer 80 % des unités textuelles dans 4 classes. Le graphique ci-dessous présente pour chaque classe le vocabulaire spécifique, et les pièces qui y sont le plus représentées. Les noms qui ont été donnés aux classes relèvent d'un travail d'interprétation par sélection des traits sémantiques qui paraissent le mieux représenter la classe.

Figure 11 : Champs lexicaux dans les oeuvres de Racine



Ci-dessous, on lit pour chaque classe, quelques extraits significatifs.

Comédie

Ma foi, juge et plaideurs, il faudrait tout lier. Monsieur, encore un coup, je ne puis pas tout faire: Puisque je fais l'huissier, faites le commissaire.	Et vous, venez au fait. Un mot du fait. Hé! Faut il tant tourner autour du pot? Ils me font dire aussi des mots longs d'une toise, De grands mots qui tiendraient d'ici jusqu'à Pontoise.
Mon père, éveillez vous. Monsieur, êtes vous mort? Mon père! Hé bien? Hé bien? Quoi? Qu'est-ce? Ah! Ah! Quel homme! Certes, je n'ai jamais dormi d'un si bon somme.	Avez vous déchiré ce papier sans le lire? Monsieur, je l'ai lu. Bon. Continuez d'écrire. Et pourquoi l'avez vous déchiré?

Déchirement

J'aurais fini cent fois ma triste destinée, Si je n'eusse songé jusques à mon retour Que mon éloignement vous prouvait mon amour, Et que le souvenir de mon obéissance Pourrait en ma faveur parler en mon absence,	Surtout si de Junie évitant la présence, Vous condamniez vos yeux à quelques jours d'absence: Croyez moi, quelque amour qui semble vous charmer, On n'aime point, seigneur, si l'on ne veut aimer.
Je réponds, en partant, de son obéissance; Et même elle m'a dit que prêt a l'épouser, Vous ne la verrez plus que pour l'y disposer. Ah! Qu'un aveu si doux aurait lieu de me plaire!	Mais de mon amitié mon silence est un gage: J'oublie en sa faveur un discours qui m'outrage. Je n'en ai point troublé le cours injurieux. Je fais plus: à regret je reçois vos adieux.

Le sacré

Environné d'enfants, soutiens de ma puissance, Il ne manque à mon front que le bandeau royal. Cependant, des mortels aveuglement fatal!	Faut-il que sur le front d'un profane adultère Brille de la vertu le sacré caractère? Et ne devrait-on pas à des signes certains Reconnaître le coeur des perfides humains?
Relevez, relevez les superbes portiques Du temple ou notre dieu se plaît d'être adoré. Que de l'or le plus pur son autel soit paré, Et que du sein des monts le marbre soit tiré.	Ciel! Dans un des parvis aux hommes réservé Cette femme superbe entre, le front levé, Et se préparait même à passer les limites De l'enceinte sacrée ouverte aux seuls lévites.

Pouvoir, devoir

Étrange ambition qui n'aspire qu'au crime, Ou le plus furieux passe pour magnanime! Le vainqueur doit rougir en ce combat honteux; Et les premiers vaincus sont les plus généreux.	En vain quelques guerriers, qu'anime son grand coeur, Ont ramené l'effroi dans le camp du vainqueur: Il faut bien qu'il succombe, et qu'enfin son courage Tombe sur tant de morts qui ferment son passage.
Quels courages Venus n'a-t-elle pas domptés? Vous même, où seriez vous, vous qui la combattez, Si toujours Antiope à ses lois opposée, D'une pudique ardeur n'eut brûle pour Thésée?	Mais quelque noble ardeur dont ils puissent brûler, Peuvent-ils de leur roi venger seuls la querelle? Pour un si grand ouvrage est-ce assez de leur zèle?

Les deux graphiques suivants présentent les résultats de l'analyse factorielle (effectuée sur le tableau croisant les classes d'alexandrins et les mots), sur les plans factoriels définis par les axes 1 et 2, puis par les axes 3 et 2. Ceci permet d'identifier les "proximités" entre pièces.

Figure 12 : Projection sur le premier plan factoriel des oeuvres de Racine (analyse 2)

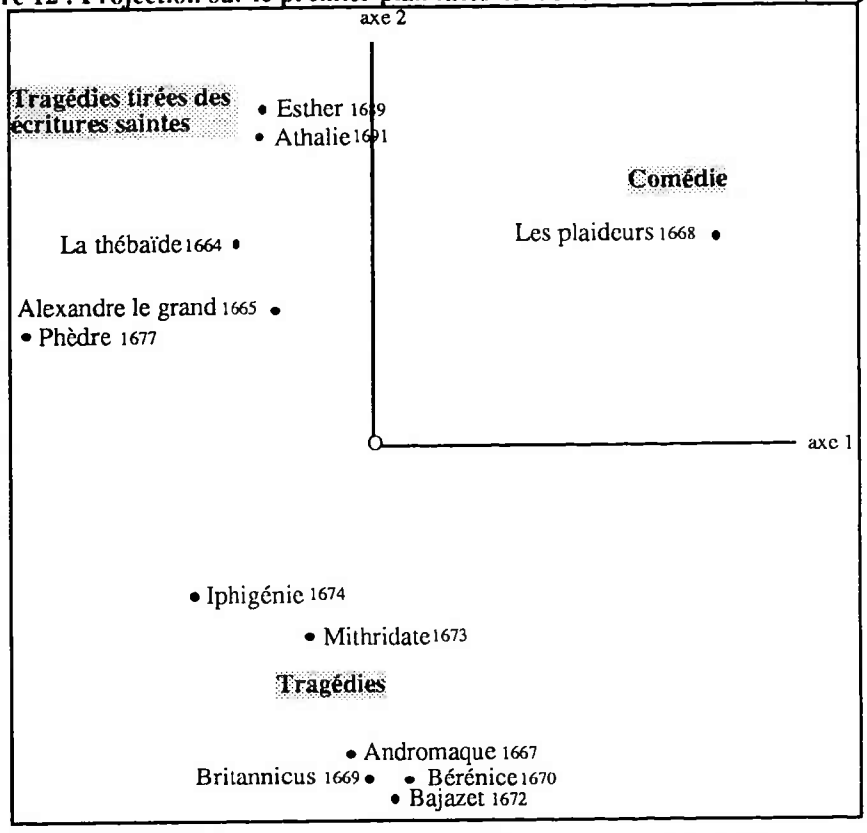
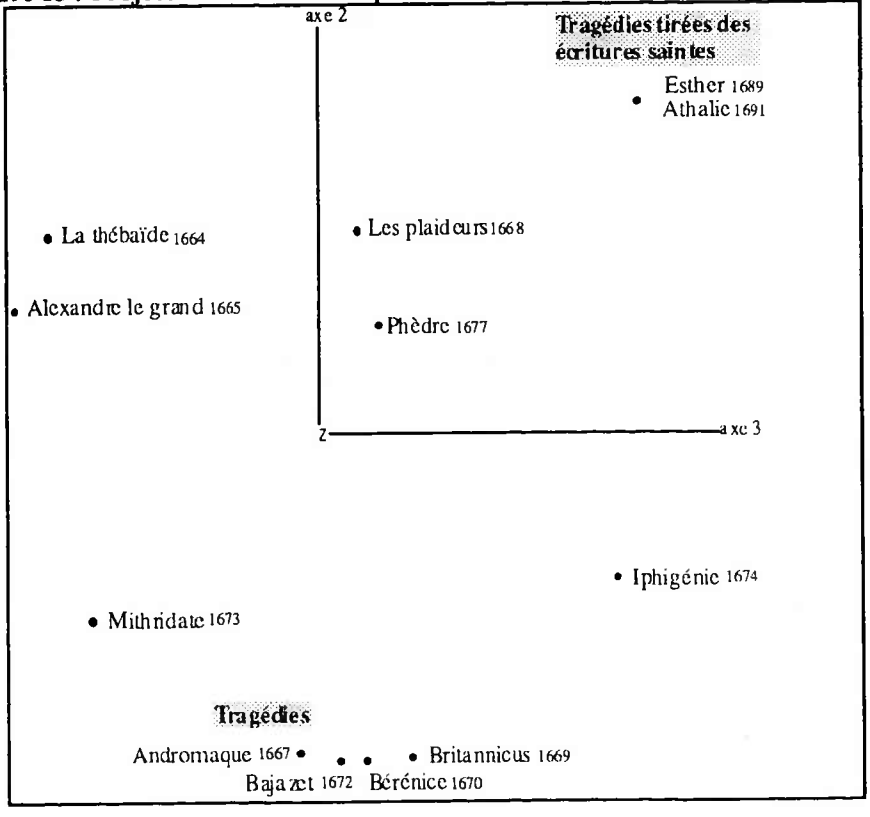


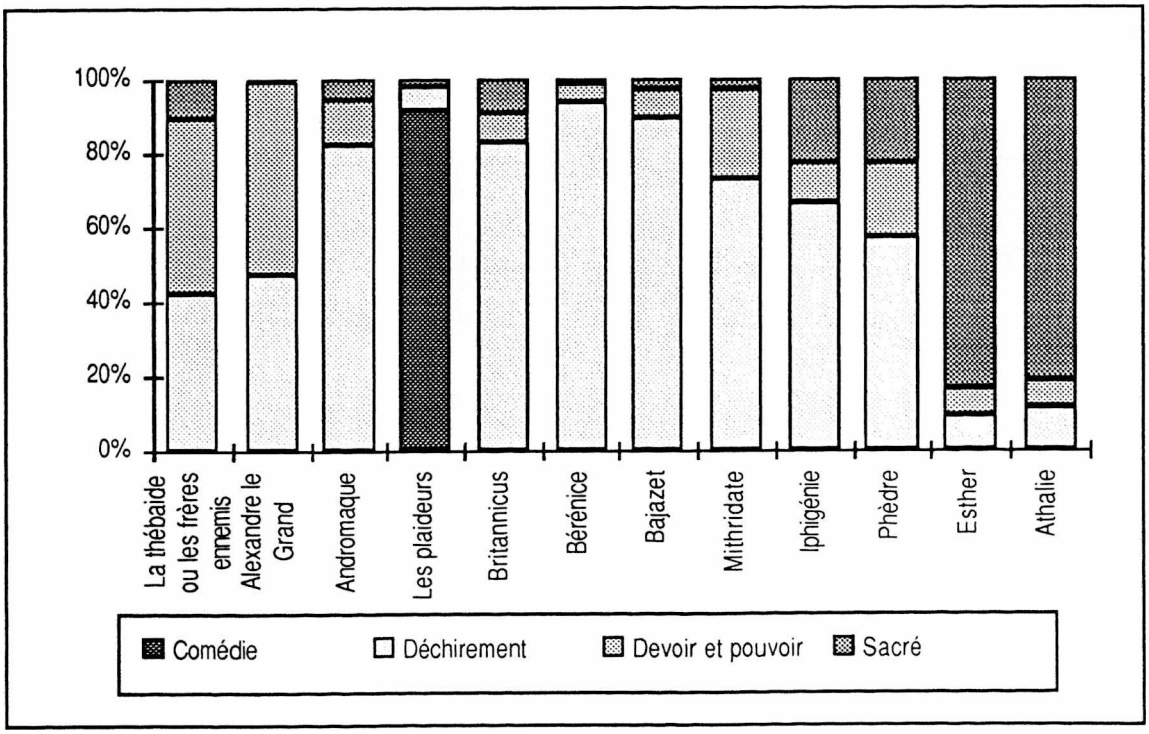
Figure 13 : Projection sur le second plan factoriel des oeuvres de Racine (analyse2)



Cette analyse fait apparaître de manière très claire l'opposition entre comédie et tragédies sur le premier axe ; l'opposition entre les tragédies de la période centrale et celle du début et de la fin sur le second axe ; et enfin sur le troisième, les différences entre les pièces sacrées et les autres. On voit ici qu'*Iphigénie* est assez proche par certains aspects des deux dernières tragédies. En revanche, *Andromaque* n'occupe pas ici une place intermédiaire.
La représentation sur le premier plan factoriel est assez trompeuse dans la mesure où elle fait apparaître une fausse proximité entre les deux tragédies du début et celles de la fin, proximité qui disparaît sur le plan factoriel suivant.
On note aussi la place remarquable de *Phèdre*, au centre du graphique. C'est une pièce centrale qui entretient des relations de proximités avec toutes les autres tragédies. Le profil lexico-sémantique de la pièce confirmera cette position centrale.

Alors que dans la première analyse (avec tout le paratexte), les classes regroupaient les pièces dans leur quasi intégralité, ici ces dernières se trouvent plutôt éclatées entre les différentes classes thématiques, montrant ainsi la permanence de certains thèmes dans l'univers de Racine.
Le graphique suivant présente les profils lexico-sémantiques des pièces : 42 % des fragments de *La thébaïde* ont été classés dans la classe "Déchirement", 48 % dans la classe "Devoir et pouvoir" et 10 % dans la classe "Sacré".

Figure 14 : Profils lexico-sémantiques des oeuvres de Racine



Les profils des pièces montrent clairement que la comédie se distingue très nettement des tragédies : la comédie et la tragédie ne partagent absolument pas les mêmes "mondes lexicaux" pour reprendre l'expression de Max Reinert : rien de commun dans l'organisation du vocabulaire.
En revanche, les tragédies partagent beaucoup d'éléments, nous voyons que les trois classes d'énoncés qui sont le propre de la tragédie sont représentées dans chacune d'entre elles. Ce qui varie de manière considérable d'une pièce à l'autre, c'est l'importance relative de chacun de ces univers lexicaux.

Les profils lexico-sémantiques permettent d'identifier les ressemblances et différences entre pièces et mettent en évidence deux résultats :

- il existe trois périodes dans l'écriture tragique de Racine ;
- le passage d'une période à une autre est amorcé progressivement.

La comparaison des histogrammes (fig. 14) montre que l'on a trois groupes de tragédies, l'un regroupant les deux premières, le second regroupant les tragédies centrales, le dernier les deux dernières. Cette classification recouvre d'une manière parfaite l'évolution chronologique et recouvre entre autre la distinction entre les tragédies profanes et les tragédies sacrées.
Il est intéressant de constater que les tragédies d'*Andromaque* à *Phèdre* ont globalement un profil très proche ce qui marque une grande homogénéité dans l'écriture et dans la trame tragique, en dépit du fait que les actions se déroulent en des lieux très distincts (Turquie, Rome, ...) et ne racontent pas les mêmes histoires.
Nous avons fait de nombreuses tentatives pour essayer de diviser la grande classe "Déchirement" et faire apparaître des différences plus significatives entre les pièces, mais l'organisation lexicale résiste à cette tentative[1].

Il est aussi surprenant de noter que les transitions d'une période à l'autre se font sous le signe de la continuité plus que de la rupture. Par exemple, le passage aux tragédies sacrées est amorcé dès *Iphigénie*. En effet, on voit apparaître dès cette pièce une thématique nouvelle qui prend de l'expansion dans *Phèdre* avant de s'affirmer pleinement dans les deux dernières tragédies, toutes deux d'inspiration chrétienne. Notons enfin que Phèdre (nous avions déjà vu sa place centrale sur le graphique factoriel) a le profil "le plus équilibré" : chacune des thématiques tragiques y est assez bien représentée.

---

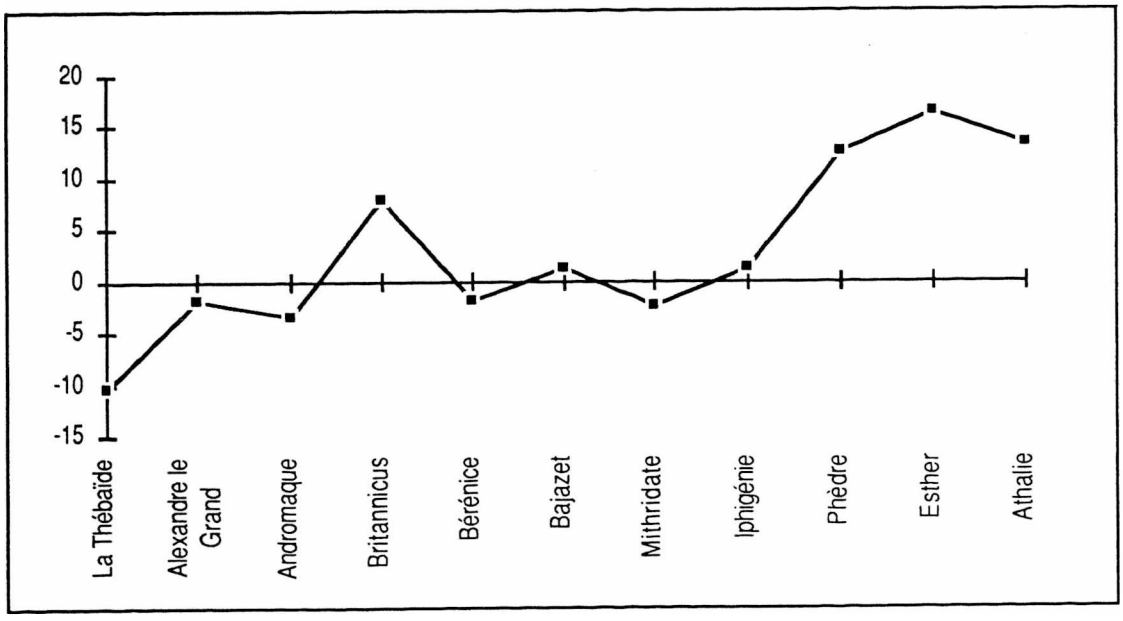
[1] En revanche, en gardant les noms propres, nous y parvenons aisément : *Bajazet*, la seule tragédie orientale se distingue des autres, par sa "couleur locale" très spécifique.

2.3.3. Identifier des périodes

Racine a écrit ses pièces sur une période de 27 ans. Ses sources d'inspirations et ses modèles ont évolué au cours du temps. Comme nous venons de le voir, on peut identifier trois grands types de tragédies qui correspondent à trois périodes. Nous voudrions voir si une approche différente qui repose sur l'étude de l'accroissement du vocabulaire conduit à des résultats similaires. Dans ce cas, on ne s'intéresse plus au **contenu** lexical mais à la **structure** lexicale.

Charles Bernet (1983, p.120) proposait de comparer la courbe de l'accroissement théorique à la courbe de l'accroissement réel en mesurant les écarts réduits. Dans le graphique suivant, l'accroissement théorique est figuré par l'axe des abscisses et en ordonnée, on lit l'écart réduit entre les valeurs observées et les valeurs théoriques. Le calcul des effectifs théoriques repose sur le modèle binomial. Quand le point a une valeur positive cela signifie qu'il y a dans le texte un apport de formes nouvelles plus important que la prédiction du modèle.

Figure 15 : Accroissement du vocabulaire : écarts réduits (Ch. Bernet)



Pour Bernet, la formule binomiale n'est pas valable pour $p<0,1$, or la première pièce représente moins de 1 % du corpus donc la première valeur n'est pas valide. La valeur de l'écart réduit pour *Alexandre*, *Bérénice*, *Bajazet* et *Britannicus* n'est pas significative. *Britannicus*, *Phèdre*, *Esther* et *Athalie* ont un vocabulaire plus original et plus riche. On voit à

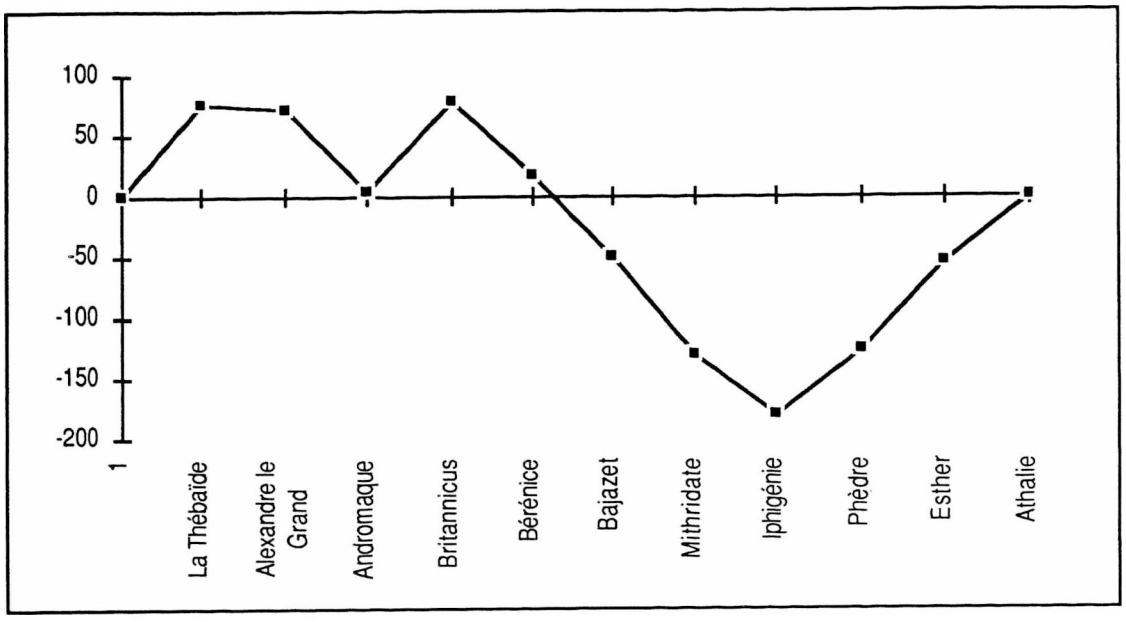
partir d'*Iphigénie* un accroissement du vocabulaire important qui semble coïncider avec l'émergence de la thématique sacrée que nous avions vue apparaître précédemment.
En revanche, l'apport d'*Andromaque* en vocabulaire neuf est assez décevant alors que l'analyse des profils montrait précédemment une nette modification du profil entre *Alexandre* et *Andromaque*. Il n'y a pas de coïncidence parfaite entre structure lexicale et contenu lexical. On peut peut-être dire qu'en réalité, toutes les thématiques étaient déjà apparues précédemment, et qu'*Andromaque* n'apporte qu'une modification de la répartition des thématiques.

Dominique Labbé (1993), partant du constat que le modèle théorique de l'accroissement du vocabulaire s'ajuste mal aux données réelles, propose un modèle d'urne plus sophistiqué qui comprend plusieurs urnes, l'une d'où est tiré le vocabulaire général (mots-outils, verbes, noms usuels...) et qui obéit à la loi de Muller, les autres d'où sont extraits des vocabulaires spécifiques. Un coefficient p mesure la part du vocabulaire spécialisé au vocabulaire total. En s'appuyant sur ce modèle, il propose une autre manière de mesurer l'accroissement marginal du vocabulaire.
A partir de cette courbe de l'accroissement marginal, Labbé propose d'identifier des périodes.

> Une période s'ouvre par un afflux de formes nouvelles, signalant la mise en place d'un ou de plusieurs grands thèmes, et se prolonge par une raréfaction progressive du vocabulaire nouveau au fur et à mesure que cette thématique s'épuise.
>
> (Labbé, 1993, p. 106)

On a donc appliqué aux pièces de Racine ce modèle (Dominique Labbé m'a aimablement confié les données et l'application informatique ; les données d'origine proviennent de l'ouvrage de Charles Bernet (1983, p.115 et pp.233-236)). La courbe ci-dessous présente l'accroissement marginal calculé à partir du modèle multi-urnes.

Figure 16 : Accroissement du vocabulaire : données observées par rapport aux données théoriques (D. Labbé)



Si l'on s'en tient au modèle d'interprétation de D. Labbé, deux résultats apparaissent ici. Jusqu'à *Bérénice*, le renouvellement du vocabulaire d'une pièce à l'autre, est supérieur à l'accroissement théorique. A partir de 1970, l'accroissement du vocabulaire est inférieur à la performance moyenne.

D'autre part, on voit apparaître de manière assez nette trois périodes : la première comprend les trois première pièces, la seconde les tragédies de *Britannicus* à *Iphigénie* et la troisième *Phèdre* et les tragédies chrétiennes. Ce découpage ne correspond pas à celui que nous avions obtenu précédemment. Il faudra examiner plus précisément le modèle sous-jacent.

Ces calculs reposent exclusivement sur l'accroissement du vocabulaire en ne tenant aucun compte des signifiants eux-mêmes. Il paraît assez évident que l'apparition d'une nouvelle thématique fait apparaître de nouveaux termes qui entraînent un accroissement du vocabulaire. Mais nous ne pouvons guère dire plus sur les liens entre ces deux approches.

2.4. TRAGÉDIES DE RACINE ET CORNEILLE

Voyons maintenant quelques caractéristiques de la tragédie. Ici encore deux approches complémentaires ont été retenues : tout d'abord mettre en évidence à partir des données lexicométriques les composantes caractéristiques du genre de la tragédie, ensuite montrer les particularités des univers tragiques de Corneille et Racine.

2.4.1. La tragédie comme genre littéraire

On a préalablement constitué un corpus contenant toutes les tragédies de Corneille et Racine, soit trente-deux textes. En utilisant Hyperbase, on obtient le vocabulaire spécifique des tragédies par rapport au TLF. Un biais dans les résultats provient de la différence d'époque entre les deux corpus comparés. Ce que nous observons est donc, d'une part, lié au genre de la tragédie, de l'autre à l'évolution de la langue. Comme il est impossible de démêler les deux effets, nous ferons "comme si" l'effet "langue" n'intervenait pas...
Voici donc un extrait de la liste des mots spécifiques des tragédies de Corneille et Racine par rapport au TLF.

Fréquence	Mots spécifiques	Écart/TLF
1338	seigneur	184.2
743	encor	171.4
284	courroux	160.9
11089	vous	152.0
220	trépas	122.0
1658	vos	111.3
1056	madame	99.3
367	voeux	98.9
628	dieux	90.5
2285	votre	89.9
152	souffrez	86.2
670	gloire	78.1

Fréquence	Mots spécifiques	Écart/TLF
177	régner	74.5
255	o	73.6
5941	a	72.9
313	trône	72.6
189	tyran	71.1
878	sang	71.0
1464	amour	70.8
5482	?	70.7
456	haine	68.3
208	venger	68.3
2778	ma	67.8

La liste intégrale figure dans l'annexe 3. On a donc repéré quelques champs sémantiques caractéristiques de la tragédie en regroupant autour de concepts les termes spécifiques.

2.4.1.1. Le dialogue

Dans la mesure où le dialogue est au coeur même de la tragédie, la fonction conative est très représentée. L'interpellation est un moyen efficace pour s'adresser à l'autre et pour motiver son attention : les termes *Seigneur* et *Madame* sont les plus fréquemment utilisés comme nous l'avions vu précédemment.
L'univers tragique est aussi caractérisé par des éléments du discours qui font avancer le dialogue et par conséquent l'action. L'interrogation (on voit dans la liste de termes que le point d'interrogation est caractéristique de la tragédie : en moyenne, un vers sur dix est sous forme interrogative) est un des moyens les plus naturels pour accélérer le déroulement de l'action. Les formes à l'impératif qui suscitent une implication de l'interlocuteur apparaissent aussi comme spécifiques de la tragédie (*souffrez*, *craignez*, *songez*, *jugez*...). Ce sont des éléments qui sont peut-être davantage du ressort du théâtre que de la tragédie. La particularité de la tragédie est la nature même des verbes qui portent la marque de l'impératif, qui sont paradoxalement rarement des verbes d'action. Spitzer (1970), parle à ce propos de verbes phraséologiques qui impliquent un état psychique plus qu'une action.

2.4.1.2. Les composantes de l'univers tragique

Quelques champs sémantiques caractéristiques de la tragédie peuvent être reconstitués "à la main" à partir de la liste des mots spécifiques :

- **La puissance et la gloire** sont au coeur de cet univers. Les termes qui s'y rapportent sont les suivants :
  Seigneur, régner, trône, tyran, sceptre, rois, roi, reine, empire, rang, prince, ordonne, couronne, tyrans, princesse.
  gloire, vainqueur, digne, ardeur, flamme, illustre, exploits, victoire, vertu, feux, devoir, honneur, généreux, zèle, mériter, juste, héros, glorieux, courage, éclat, vaincre, aspire, conquête, audace.
- **La passion amoureuse** est inséparable dans l'univers tragique de la souffrance :
  amour, pleurs, coeur, soupirs, ingrat, amant, indigne, flamme, malheurs, jaloux, cruel, faveur, feux, aime, charmes, mériter, maux, plaire, yeux, cruels, aimer, trouble, transports.

- **La colère et le désir de vengeance** sont souvent des moteurs de l'action :
  courroux, sang, haine, venger, rival, punir, perfide, fureur, fers, vengeance, fureurs, hait, traître, outrage, offense, péril, périls, haïr, rivale, trahir, ennemis, lâche, barbare, infamie, colère, mépris.
- **Le destin et la mort** créent l'enfermement tragique et piègent l'âme :
  dieux, choix, dessein, craindre, destins, sort, destin, ciel, fatal, auguste, crains, hélas, desseins, dieu, victime.
  trépas, sang, crime, funeste, périr, fers, mort.

L'analyse d'un échantillon de tragédies sous Alceste permet de mettre en évidence une typologie des mondes lexicaux les plus fréquents dans la tragédie et les attentions différentes que Corneille et Racine ont porté à chacune de ces représentations.

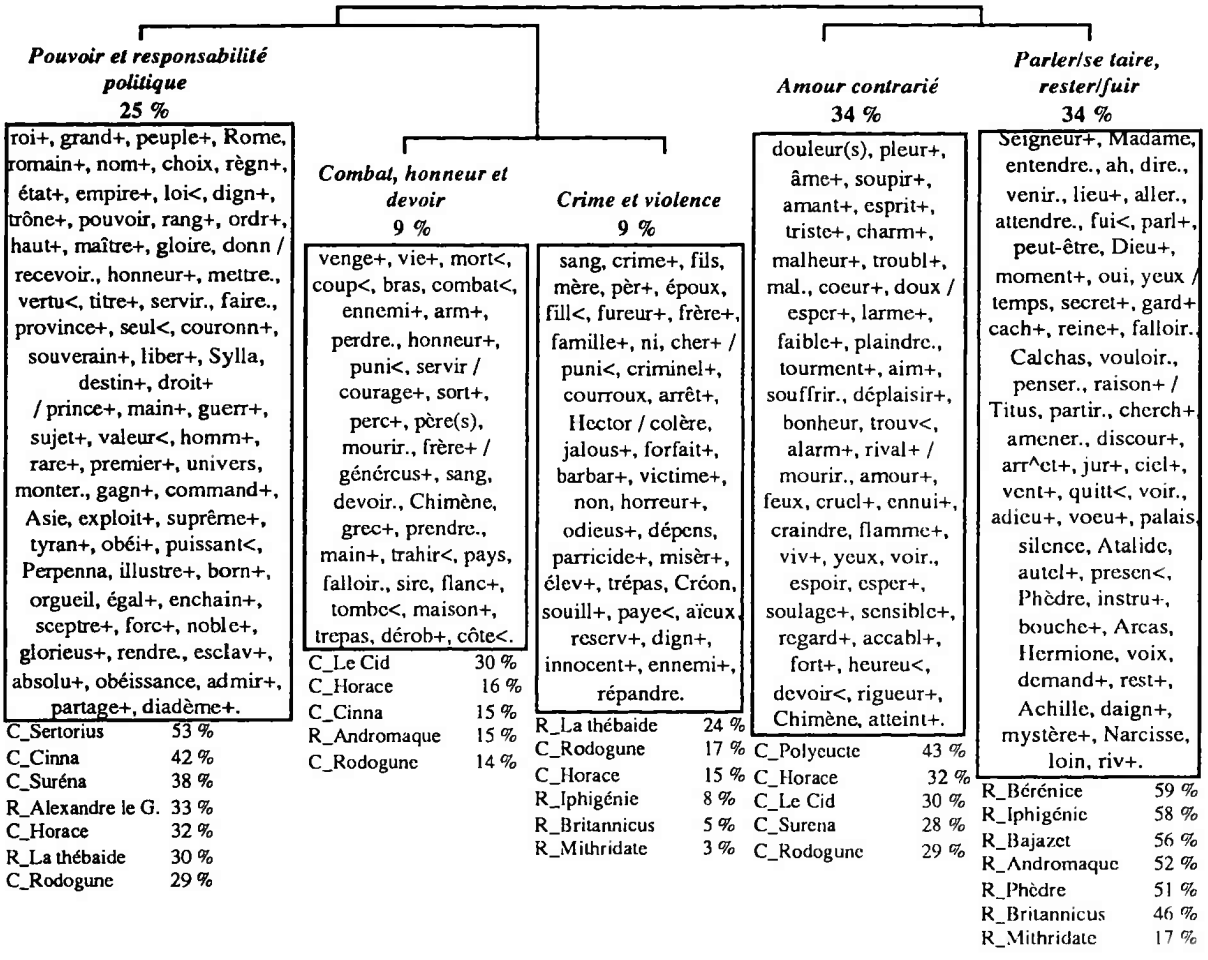
On a retenu pour l'analyse les pièces suivantes :

Corneille	
Le Cid	1636
Cinna	1640
Horace	1640
Polyeucte	1640
Rodogune	1644
Sertorius	1662
Surena	1674

Racine	
La thébaïde	1664
Alexandre le Grand	1665
Andromaque	1667
Britannicus	1669
Bérénice	1670
Bajazet	1672
Mithridate	1673
Iphigénie	1674
Phèdre	1677

Le choix est discutable et permet juste d'entrevoir ce que pourrait donner une analyse sur les oeuvres complètes. L'analyse donne cinq classes qui sont, pour une part, thématiques et recouvrent les thèmes précédents (amour contrarié, pouvoir, crime et violence) mais font aussi apparaître une classe qui concentre l'essence même du tragique.

Figure 17 : Les thématiques de la tragédie : Corneille et Racine



La classe *Pouvoir et responsabilité politique* traduit la situation de pouvoir des personnages et la puissance qu'ils exercent sur leurs sujets. Elle rappelle la grandeur des enjeux qui occupent les héros.

La classe *Combat, honneur et devoir* exalte les vertus du héros tragique qui, dans la vengeance et le combat, défend sa gloire et son honneur. On retrouve ici tous les attributs du héros cornélien tel que le définit Bénichou (1948).

La classe *Crime et violence* est celle qui se rattache le mieux à la tragédie grecque et latine. Elle est fortement influencée par le thème du sacrifice humain si fortement attaché à la famille des Atrides et par l'horreur qu'il suscite.

Dans la classe *Amour contrarié*, apparaît la conjonction du désir et de sa répression, puisque le destin s'oppose à ce que le sentiment soit partagé ou puisse avoir une existence sociale. Douleur, pleurs et soupirs sont les fidèles compagnons de l'amour. On notera aussi que le trouble amoureux s'exprime de manière privilégiée par le regard : *yeux, voir, vue, regard ...* ce qui a déjà longuement été commenté par Starobinski.

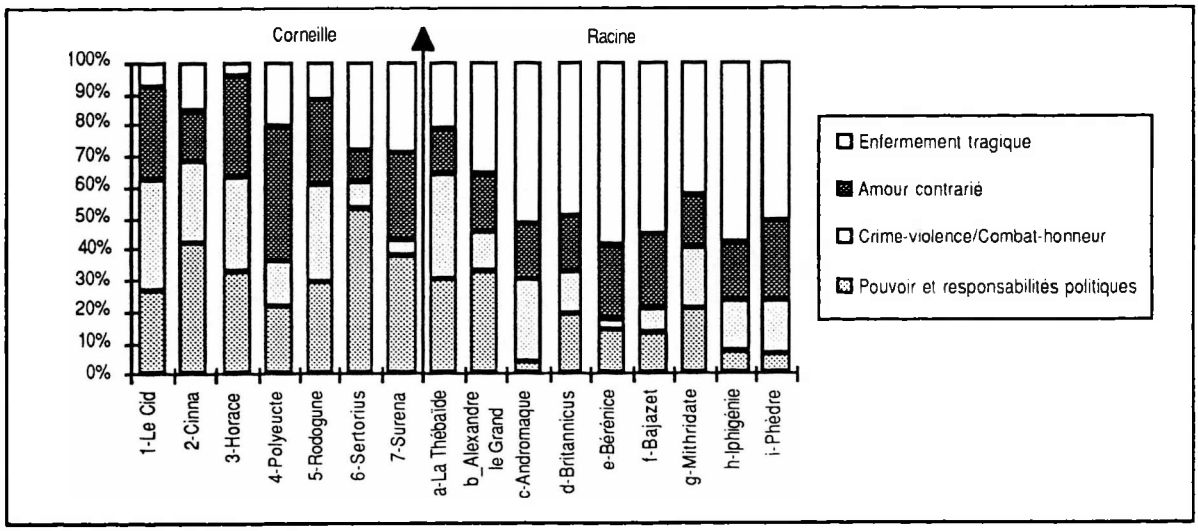
La dernière classe représente l'essence même du tragique en tant que moteur du déroulement de l'action mais aussi en tant que sentiment. C'est le lieu où s'exprime le mieux l'enfermement tragique. Cette classe regroupe des alexandrins qui annoncent l'imminence de l'aveu d'un sentiment ou de l'annonce d'une décision. C'est le dilemme, puissant moteur de la dialectique scénique, qui sous-tend le discours. Ces alexandrins jouent un rôle décisif dans le jeu des entrées-sorties puisqu'ils annoncent des changements de personnages. Le secret, la décision jusque-là retenus, en devenant paroles cristallisent le sentiment tragique. L'aveu piège définitivement l'âme et l'exil ou la mort (thèmes très représentés dans cette classe : *fuir*, *partir*, *quitter*, *adieu*) deviennent les seules issues possibles.

La "coexistence dans le même lieu du discours de catégories contradictoires" (Ubersfeld, 1982) qui apparaît dans la plupart des classes (bonheur/douleur, vie/mort, parole/silence...) semble bien être une des clefs du tragique.

2.4.2. Corneille et Racine : ressemblances et différences

Ces grands thèmes n'apparaissent pas de manière identique dans toutes les pièces.

Figure 18 : Répartition des thèmes selon les pièces : Corneille et Racine



Le théâtre de Racine traite moins de pouvoir, de vengeance et de sacrifice, en revanche le tragique s'exprime davantage dans les situations de dialogue et dans cette tension entre le désir d'échapper à la fatalité et le constat de l'enfermement. En ce sens le tragique racinien est plus "scénique". On voit ici aussi que le profil des deux premières pièces de Racine est assez proche de celui de Corneille, ce qui confirme une fois de plus le statut particulier qu'elles occupent dans la production de Racine.

Ces corpus pourraient également être soumis à d'autres types d'analyse : ils peuvent être considérés comme des séquences de textes et fort bien se prêter à des analyses chronologiques (Salem, 1994).

L'oeuvre de Corneille et celle de Racine ont déjà été largement étudiées et commentées. Il est remarquable de retrouver à l'aide des méthodes lexicométriques des résultats concordants. Ce sont donc des outils qui peuvent être utilisés de manière fiable sur des textes non encore exploités.
Mais même sur des textes visités et revisités, la statistique textuelle n'apporte pas que des évidences, elle donne aussi un point de vue synthétique sur des corpus de taille considérable, qu'il n'est pas toujours aisé d'analyser manuellement. D'autre part, de manière rétroactive, les résultats obtenus à l'aide de méthodes automatiques valident les approches stylistiques traditionnelles qui consistent à atteindre par cercles concentriques «l'étymon spirituel de l'oeuvre»[1].

Il nous reste à voir si, sur des corpus non littéraires et donc moins structurés comme les entretiens semi-directifs, ces outils de statistique textuelle peuvent être aussi fiables.

[1] SPITZER Léo, (1970).- *Etudes de style*, Paris, Tel, Gallimard.

BIBLIOGRAPHIE

BEAUDOUIN Valérie, (1993).- *Analyse lexicale et stylistique : Gravitations de Jules Supervielle*, CRÉDOC, Cahier de recherche n°49, Paris.

BEAUDOUIN Valérie, (1993).- "Stylistique et analyse lexicale : Corneille et Racine", *Secondes journées internationales d'analyse statistique de données textuelles*, Montpellier, ENST-TELECOM.

BEAUDOUIN Valérie, BOISBOUVIER Nathalie, HÉBEL Pascale, LITMAN Sonia, RACAUD Thierry, (1993).- *L'analyse lexicale : outil d'exploration des représentations. Résultats illustratifs*, CRÉDOC, Cahier de recherche n°48 bis, Paris.

BEAUDOUIN Valérie, LAHLOU Saadi (1993).- *L'analyse lexicale : outil d'exploration des représentations*, CRÉDOC, Cahier de recherche n°48, Paris.

BEAUDOUIN Valérie, LAHLOU Saadi, YVON François, «Réponse à une question ouverte : incidence du mode de questionnement», *Secondes journées internationales d'analyse statistique de données textuelles*, Montpellier, ENST-TELECOM, 1993.

BÉNICHOU M., (1948).- *Morales du grand siècle*, Gallimard, Paris.

BENZÉCRI Jean-Paul, LEBART Ludovic, REINERT Max (1981).- *Pratique de l'analyse des données, Linguistique et lexicologie*, Dunod, Paris.

BENZÉCRI Jean-Paul, (1982).- *Histoire et Préhistoire de l'Analyse des Données*, Dunod, Paris, 1982.

BERNET Charles, (1983).- *Le vocabulaire des tragédies de Racine*, Slatkine-Champion, Paris-Genève.

BRIAN Eric, (1984).- *Analyse des données lexicométriques. Elaboration des programmes.* Rapport CRÉDOC, n°4909, Paris.

BRIAN Eric, (1986).- *Techniques d'estimation et méthodes factorielles, exposé formel et application aux traitements de données lexicométriques*. Thèse de Docteur Ingénieur, Orsay.

BRUNET Etienne, (1978).- *Le vocabulaire de Jean Giraudoux, Structure et évolution,* Slatkine-Champion, Paris-Genève.

BRUNET Etienne, (1992).- *Hyperbase,* CUMFID n°17 (version 1.5).- 84 p.

CLERC Laurent, DUFOUR Ariane, (1992).- *Deux analyses lexicales : les améliorations à apporter au fonctionnement de la société et l'image du milieu professionnel.* CRÉDOC, Cahier de recherche, n°22, Paris.

DUFLOS Catherine, VOLATIER Jean-Luc, (1991).- *Les Français et la justice, un dialogue à renouer*, CRÉDOC, Collection des rapports, n°109, Paris.

GUILBAUD G. -Th., (1980).- "Zipf et les fréquences", *Mots*, Paris, Presses de la Fondation Nationale des Sciences Politiques, n° 1, , pp.97-125.

HAEUSLER Laurence, (1987).- *Analyse lexicale de réponses libres : le coût de l'électricité*, CRÉDOC, Collection des rapports, n°14, Paris.

HUBERT Pierre, LABBÉ Dominique, (1988).- «Un modèle de partition du vocabulaire», in LABBÉ, Dominique, THOIRON, Philippe, SERANT, Daniel, *Etudes sur la richesse et la structure lexicales*, Slatkine-Champion, Paris-Genève.

JAKOBSON R., (1963).- «Poétique», *Essais de linguistique générale,* Ed. de Minuit, Paris.

LABBÉ Dominique, (1993).- «Un modèle d'analyse du vocabulaire», *Secondes journées internationales d'analyse statistique de données textuelles*, Montpellier, ENST-TELECOM, pp.103-114.

LABBÉ Dominique, HUBERT Pierre, (1994).- «La richesse du vocabulaire», Colloque International *Consensus ex machina*, Paris, ALLC-ACH, 19-23 avril.

LAHLOU Saadi, (1992).- *SI/ALORS : «BIEN MANGER» ? - Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus constitué des associations libres de 2000 individus*, CRÉDOC, Cahier de recherche, n°34, Paris.

LAHLOU Saadi, COLLERIE DE BORELY Aude, BEAUDOUIN Valérie, (1993).- *Où en est la consommation aujourd'hui ? : une enquête sur le consommateur français des années 90*, CRÉDOC, Cahier de recherche, n°46, Paris.

LAHLOU Saadi, (1994).- "Modélisation des représentations sociales par l'analyse lexicale des énoncés de dictionnaires : une nouvelle approche pour la psychologie sociale", Colloque International *Consensus ex machina*, Paris, ALLC-ACH, 19-23 avril.

LEBART Ludovic, SALEM André, (1988).- *Analyse statistique des données textuelles.* Préface de Christian BAUDELOT.- Paris, Dunod.

LEBART Ludovic, SALEM André, (1994).- *Statistique textuelle*, Paris, Dunod, 342 p.

LION Sébastien, (1991).- *Constitution d'un corpus et perte d'information en analyse lexicale : méthodes et pratiques*, CRÉDOC, Cahier de recherche, n°13, Paris.

MULLER Charles, (1967 réed 1979).- *Le vocabulaire du théâtre de Pierre Corneille. Etude de statistique lexicale*, Larousse réimp. Slatkine-Champion, Genève-Paris.- 382 p.

MULLER Charles, (1977 réed 1992).- *Principes et méthodes de statistique lexicale*, Larousse, réimpression Slatkine-Champion, Genève-Paris.- 211p.

MULLER Charles, (1993).- *Langue française : débats et bilans. Recueil d'articles 1986-1993*, Champion-Slatkine, Genève-Paris.- 247 p.

REINERT Max, (1983 b).- "Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte", *Les cahiers de l'analyse des données*, Vol. VIII, n° 2, Dunod.- p 187-198.

REINERT Max (1987).- "Classification descendante hiérarchique et analyse lexicale par contexte : application au corpus des poésies d'Arthur Rimbaud", *Bulletin de Méthodologie Sociologique*, n°13.

REINERT Max, (1990).- "ALCESTE, une méthode d'analyse des données textuelles. Application au texte «Aurélia» de Gérard de Nerval" *Bulletin de Méthodologie Sociologique,* 26.- pp. 25-54.

REINERT Max, (1992).- *La méthodologie d'analyse des données textuelles ALCESTE ; Application à l'analyse des poésies d'Arthur Rimbaud*, Conférence d'Albi.

REINERT Max, (1992).- "La méthodologie ALCESTE et l'analyse d'un corpus de 304 récits de cauchemars d'enfants", Convention internationale *Ricerca qualitativa e computer nelle scienze sociali*, Rome, 30 novembre-2 décembre.

REINERT Max, (1993).- «Les "mondes lexicaux" et leur logique», *Langage et société*, Paris, Maison des Sciences de l'Homme, n°66.- pp. 5-39.

ROUBAUD Jacques, (1978).- *La vieillesse d'Alexandre*, Paris, Éditions Ramsay, 1988, (ED François Maspero 1978).

ROUBAUD Jacques, (1986).- "DYNASTIE : études sur le vers français, sur l'alexandrin classique", Première partie, *Cahiers de poétique comparée*, n° 13, Publications Langues' O, Paris.

SALEM André, (1994).- "La lexicométrie chronologique. L'exemple du Père Duchesne d'Hébert", 4e Colloque de *Lexicologie politique - Langages de la révolution 1770-1815.*

SICHEL H. S., (1975).- "On a Distribution Law for Word Frequencies", *Journal of the American Statistical Association*, 70, p. 542-547.

SPITZER L., (1982).- "L'effet de sourdine", *Études de style*, Tel, Gallimard, Paris, 1970.

THOIRON Philippe, LABBÉ Dominique, SERANT Daniel, (1988).- *Etudes sur la richesse et la structure lexicales*, Champion-Slatkine, Genève-Paris.- 174 p.

TOURNIER Maurice, (1980).- "D'où viennent les fréquences de vocabulaire ?", *Mots*, Presses de la Fondation Nationale des Sciences Politiques, n° 1, Paris.- pp. 189-209.

UBERSFELD Anne, (1982).- *Lire le théâtre*, Éditions Sociales, Paris.

YVON François, (1990).- *L'analyse lexicale appliquée à des données d'enquête : état des lieux.*, CRÉDOC, Cahier de recherche, n°5, Paris.

ZIPF G. K., (1974 (1ère édition 1936)).- *La psychobiologie du langage : une introduction à la philologie dynamique*, RETZ-CEPL, Paris.

ANNEXES

ANNEXE 1 : DISTRIBUTIONS DE FRÉQUENCES

Le tableau ci-dessous présente les distributions de fréquence calculées sur nos cinq corpus. Les distributions ont été calculées sur les formes non lemmatisées avec une norme de dépouillement identique (celle définie par Etienne Brunet dans Hyperbase). Nous avons retenu les fréquences de 1 à 100 pour les quatre premiers corpus. Pour le dernier, nous nous sommes arrêtés à 50 car le corpus est beaucoup plus petit que les autres. La gamme de fréquence est présentée selon les notations proposées par Charles Muller (1977).

Fré-quence	Corneille		Racine		Portraits		Manger		Petit déjeuner	
f_i	V_i	iV_i	V_i	iV_i	V_i	iV_i	V_i	iV_i	V_i	iV_i
1	5060	5060	3805	3805	3297	3297	8734	8734	440	440
2	1908	3816	1346	2692	1250	2500	2735	5470	120	240
3	997	2991	769	2307	633	1899	1363	4089	65	195
4	754	3016	540	2160	419	1676	850	3400	43	172
5	567	2835	348	1740	307	1535	530	2650	16	80
6	440	2640	282	1692	220	1320	355	2130	18	108
7	325	2275	194	1358	177	1239	292	2044	9	63
8	281	2248	184	1472	157	1256	224	1792	12	96
9	238	2142	142	1278	151	1359	177	1593	11	99
10	201	2010	133	1330	108	1080	155	1550	13	130
11	176	1936	120	1320	102	1122	130	1430	9	99
12	152	1824	95	1140	88	1056	102	1224	8	96
13	151	1963	89	1157	82	1066	104	1352	7	91
14	96	1344	59	826	62	868	92	1288	6	84
15	86	1290	67	1005	55	825	88	1320	3	45
16	109	1744	53	848	49	784	67	1072	1	16
17	90	1530	63	1071	41	697	53	901	8	136
18	94	1692	42	756	42	756	46	828	4	72
19	85	1615	39	741	32	608	36	684	1	19
20	71	1420	36	720	35	700	34	680	2	40
21	74	1554	35	735	36	756	30	630	6	126
22	78	1716	33	726	24	528	36	792	1	22
23	82	1886	44	1012	38	874	33	759	3	69
24	58	1392	23	552	24	576	23	552	2	48
25	59	1475	25	625	19	475	35	875	3	75
26	52	1352	30	780	30	780	21	546	1	26
27	52	1404	22	594	28	756	17	459	1	27

...

(Suite)

Fré-quence	Corneille		Racine		Portraits		Manger		Petit déjeuner	
f_i	V_i	iV_i	V_i	iV_i	V_i	iV_i	V_i	iV_i	V_i	iV_i
28	40	1120	17	476	21	588	19	532	2	56
29	26	754	20	580	9	261	19	551		0
30	50	1500	16	480	21	630	20	600	2	60
31	31	961	16	496	21	651	11	341	1	31
32	37	1184	14	448	20	640	17	544		0
33	40	1320	16	528	16	528	9	297		0
34	37	1258	16	544	17	578	8	272	4	136
35	35	1225	15	525	10	350	13	455	1	35
36	42	1512	13	468	21	756	8	288		0
37	31	1147	15	555	22	814	10	370	1	37
38	32	1216	11	418	9	342	13	494		0
39	18	702	15	585	20	780	11	429	2	78
40	30	1200	20	800	8	320	6	240		0
41	20	820	6	246	14	574	14	574		0
42	21	882	6	252	11	462	9	378	2	84
43	21	903	9	387	12	516	10	430		0
44	18	792	17	748	10	440	8	352		0
45	25	1125	14	630	12	540	4	180	1	45
46	24	1104	4	184	5	230	9	414		0
47	17	799	7	329	8	376	10	470		0
48	17	816	5	240	4	192	6	288	1	48
49	21	1029	10	490	14	686	4	196	3	147
50	23	1150	11	550	2	100	9	450	2	100
51	19	969	7	357	5	255	4	204		
52	17	884	9	468	6	312	4	208		
53	19	1007	6	318	5	265	4	212		
54	17	918	8	432	8	432	5	270		
55	17	935	7	385	3	165	2	110		
56	16	896	3	168	7	392	7	392		
57	10	570	6	342	8	456	3	171		
58	15	870	5	290	3	174	2	116		
59	11	649	9	531	3	177	6	354		
60	9	540	1	60	8	480	2	120		
61	11	671	3	183	6	366		0		
62	14	868	3	186	5	310	8	496		
63	13	819	5	315	6	378	4	252		
64	6	384	5	320	5	320	5	320		
65	9	585	5	325	3	195	4	260		
66	7	462	8	528	5	330	2	132		
67	12	804	4	268	5	335	4	268		
68	13	884	7	476	3	204	6	408		
69	15	1035	6	414	2	138	7	483		

...

(Fin)

Fré-quence	Corneille		Racine		Portraits		Manger		Petit déjeuner	
f_i	V_i	iV_i	V_i	iV_i	V_i	iV_i	V_i	iV_i	V_i	iV_i
70	9	630	2	140	5	350	2	140		
71	15	1065	4	284	3	213	3	213		
72	9	648	2	144	3	216	4	288		
73	15	1095	1	73		0	2	146		
74	7	518	2	148	4	296	3	222		
75	11	825	1	75	2	150	5	375		
76	10	760	1	76	4	304	3	228		
77	7	539	3	231	1	77	2	154		
78	11	858	3	234	3	234	2	156		
79	8	632	2	158	5	395	1	79		
80	6	480	5	400	4	320	3	240		
81	9	729	2	162	2	162	4	324		
82	7	574	4	328	5	410	2	164		
83	3	249	2	166	2	166	2	166		
84	7	588	5	420	2	168		0		
85	3	255	4	340	1	85	2	170		
86	9	774	1	86	3	258	3	258		
87	11	957	5	435	3	261		0		
88	3	264	1	88	1	88	1	88		
89	6	534	3	267	2	178	3	267		
90	6	540	1	90	3	270	1	90		
91	3	273		0	2	182	3	273		
92	7	644	1	92	2	184	2	184		
93	2	186	1	93	3	279		0		
94	9	846	2	188	2	188	2	188		
95	3	285	4	380	2	190	3	285		
96	4	384	4	384	4	384		0		
97	11	1067	3	291	1	97	1	97		
98	2	196	3	294	2	196	3	294		
99	6	594		0		0	1	99		
100	2	200	1	100	1	100		0		
>50									49	7721
>100	594	433669	198	107011	203	114114	145	69623		
Σ	14055	547297	9288	164845	8188	167937	16896	137576	879	11292
f_i	V_i	iV_i	V_i	iV_i	V_i	iV_i	V_i	iV_i	V_i	iV_i

ANNEXE 2 : VOCABULAIRE SPÉCIFIQUE DE CHACUNE DES PIÈCES DE RACINE

Pour chaque pièce de Racine, sont présentés par significativité décroissante (mesuré par une écart réduit), les termes les plus caractéristiques de la pièce par rapport à l'ensemble de l'oeuvre de Racine. On lit dans la première colonne la fréquence du mot dans la pièce, dans la seconde sa fréquence dans l'ensemble du corpus et dans la troisième la mesure de l'écart entre les deux valeurs.

La thébaïde ou les frères ennemis

34	34	20	créon
28	28	18	hémon
30	30	18	polynice
57	172	12	deux
13	13	12	thébains
13	13	12	thèbes
39	97	12	trône
44	125	11	frère
12	13	11	olympe
31	85	10	paix
69	320	9	sang
17	37	8	régner
248	2029	7	;
9	15	7	ambition
16	47	6	combat
11	31	6	dernier
413	3770	6	et
56	334	6	fils
10	25	6	frères
16	49	6	guerre
58	327	6	ils
18	67	6	rang
148	1266	5	:
6	12	5	anime
184	1577	5	ce
23	102	5	crime
11	33	5	diadème
6	12	5	enfers
197	1700	5	est
10	31	5	haïr
9	23	5	haut
309	2931	5	le
60	398	5	leur
11	37	5	princes
17	70	5	trépas
6	15	4	afin
5	15	4	argos
16	81	4	aussi
27	159	4	âme
6	15	4	beaucoup
53	403	4	bien
7	21	4	chacun
5	12	4	commun
7	21	4	couronne
21	119	4	courroux
8	30	4	crimes
18	98	4	ennemis
6	20	4	eut
6	15	4	gagner
22	128	4	haine
249	2365	4	il
5	12	4	inhumain
243	2305	4	la
6	16	4	magnanime
11	44	4	mourir
6	20	4	obstacle
80	668	4	pas
12	56	4	perdre
24	128	4	peuple
10	40	4	puisque
9	37	4	rage
8	30	4	règne
36	238	4	roi
10	39	4	serait
5	12	4	seront
5	13	4	sitôt
7	23	4	tyran
7	26	3	absence
10	46	3	aimer
12	60	3	armes
9	39	3	beau
5	17	3	descendre
34	239	3	dieux
5	16	3	effort
9	39	3	font
20	124	3	grand
6	23	3	hait
4	12	3	monter
5	16	3	noire
7	26	3	plaît
19	118	3	prince
13	65	3	princesse
5	18	3	régnez
7	29	3	répandre
4	12	3	répandu
104	940	3	si
10	49	3	su

Alexandre le grand

71	71	28	porus
56	56	25	alexandre
37	37	20	taxile
13	13	12	axiane
34	71	12	vainqueur
34	78	11	victoire
23	55	9	états
54	213	9	gloire
17	34	9	valeur
8	12	7	attaquer
8	13	7	lauriers
15	38	7	peuples
9	16	7	provinces
21	66	7	soeur
241	2029	6	;
18	57	6	ardeur
9	18	6	beaux
13	35	6	exploits
12	36	6	fers
15	50	6	orgueil
32	146	6	rois
16	67	5	amitié
67	477	5	coeur
13	47	5	combat
9	25	5	combattre
8	21	5	conquête
11	38	5	éclat
12	39	5	font
14	56	5	héros
20	81	5	maître
14	50	5	rival
47	282	5	tant
96	738	5	vos
301	2912	4	a
6	15	4	afin
10	44	4	appui
5	13	4	armé
11	41	4	arrêter
76	612	4	au
5	12	4	avouerai
22	127	4	bras
13	58	4	cherche
8	27	4	connaître
14	58	4	courage
189	1811	4	en
19	94	4	ennemi
5	12	4	estime
7	23	4	fierté
22	124	4	grand
7	20	4	moindres
8	28	4	offrir
15	69	4	partout
5	15	4	présenter
9	37	4	princes
190	1756	4	qu'
28	179	4	seul
15	64	4	soupirs
23	130	4	sous
8	25	4	sujets
10	38	4	tombeau
7	23	4	tyran
5	13	4	tyrans
14	71	4	univers
5	15	4	vaincus
11	51	3	armée
5	18	3	avaient
24	159	3	âme
50	403	3	bien
9	42	3	cent
4	12	3	chaleur
7	29	3	combats
19	119	3	courroux
5	16	3	court
12	59	3	croyez
6	24	3	douter
4	12	3	éclate
4	12	3	ferait
4	12	3	forcer
4	12	3	invincible
5	16	3	magnanime
5	16	3	oubliez
5	17	3	passage
101	915	3	sa
6	23	3	sert
93	825	3	ses
5	16	3	siens
6	23	3	vaincu
17	95	3	voulez

Andromaque

57	57	24	pyrrhus
40	40	20	hector
189	586	19	à
36	37	19	hermione
30	32	17	oreste
21	21	14	andromaque
43	77	14	grecs
19	22	13	epire
15	15	12	foire
13	13	11	phoenix
82	334	10	fils
25	49	10	grèce
24	44	10	troie
17	37	8	sois
7	13	6	enlever
11	23	6	ingrate
20	84	5	allons
8	17	5	consens
11	35	5	épouse
27	128	5	haine
281	2365	5	il
395	3423	5	je
43	234	5	non
11	35	5	voilà
152	1266	4	:
9	34	4	abandonne
12	46	4	aimer
11	44	4	charmes
13	54	4	coups
23	119	4	courroux
13	53	4	craint
19	87	4	époux
7	20	4	hélène
9	27	4	infidèle
105	766	4	lui
9	28	4	oublier
207	1756	4	qu'
7	22	4	refus
8	22	4	transport
12	52	4	venger
343	3174	3	-
333	3057	3	?
54	403	3	bien
6	21	3	conquête
5	16	3	couronner
10	44	3	destin
41	275	3	enfin
9	36	3	mépris
6	19	3	misère
9	37	3	rage
6	19	3	rendu
5	16	3	siens
10	42	3	songe
26	154	3	t'
15	79	3	temple

Les plaideurs

115	115	46	monsieur
32	38	22	bon
18	18	18	affaire
18	18	18	messieurs
17	17	18	procès
657	7416	15	.
12	12	15	oh
14	15	15	petit
13	13	15	sergent
207	1623	14	!
12	13	14	exploit
23	45	14	là
71	395	12	...
73	403	12	bien
13	21	12	fort
17	40	11	juge
85	586	10	à
12	25	10	cela
9	14	10	chose
11	22	10	homme
11	19	10	juger
98	741	10	on
8	12	10	partie
15	35	10	voilà
37	186	9	hé
7	13	8	maison
48	321	8	père
11	39	7	comment
31	201	7	faire
6	13	7	matin
7	17	7	personne
8	21	7	six
6	15	6	bas
9	28	6	cause
22	140	6	oui
7	26	5	arrêt
9	40	5	chez
30	229	5	donc
8	32	5	monde
11	61	5	porte
6	22	5	prends
5	13	5	sorte
5	15	5	soyez
10	49	5	tête
9	44	5	venir
6	18	5	vingt
10	52	5	voici
93	1266	4	:
7	39	4	beau
45	465	4	c'
20	170	4	dit
10	69	4	donne
5	21	4	droit
124	1700	4	est
13	92	4	êtes
38	374	4	fait
17	131	4	fille
5	18	4	parole
59	668	4	pas
5	21	4	pied
6	26	4	plaît
21	184	4	rien
23	215	4	suis
4	12	4	tôt
21	191	4	veux
49	612	3	au
5	26	3	celui
18	170	3	comme
4	18	3	donnez
8	50	3	fais
5	25	3	fera
3	12	3	irais
215	3423	3	je
4	20	3	mots
4	20	3	passer
4	17	3	présent
7	41	3	témoins
6	32	3	trois
17	161	3	va
10	72	3	voyez

Britannicus

72	76	26	néron
37	37	19	narcisse
34	34	18	britannicus
33	33	18	burrhus
29	29	17	junie
32	36	16	césar
32	44	15	empereur
19	19	14	octavie
17	17	13	agrippine
17	18	12	pallas
12	12	11	claude
28	65	9	cour
35	117	8	rome
12	26	6	auguste
30	124	6	empire
84	500	6	madame
14	40	6	secrets
9	23	5	sénat
10	25	5	soupçons
104	738	5	vos
71	447	5	yeux
159	1266	4	:
5	12	4	auraient
5	12	4	confier
8	23	4	désirs
5	12	4	instruit
5	13	4	jeunesse
8	23	4	liberté
26	139	4	mère
13	53	4	palais
5	13	4	poison
13	52	4	regards
124	915	4	sa
107	825	4	ses
24	128	4	soit
10	36	4	vertus
451	4037	4	vous
9	35	3	aïeux
5	16	3	appartement
6	19	3	aucun
8	30	3	bonté
5	16	3	changement
6	19	3	chaque
10	43	3	choix
6	19	3	commence
5	16	3	exil
5	15	3	intelligence
34	220	3	jour
7	26	3	lit
19	106	3	longtemps
17	84	3	ose
44	300	3	quoi
76	579	3	seigneur
11	47	3	tandis
37	247	3	toujours

Bérénice

66	66	27	titus
46	46	23	bérénice
20	20	15	paulin
16	16	14	arsace
52	117	14	rome
15	15	13	phénice
10	14	9	orient
41	141	9	reine
14	23	9	sénat
11	22	7	adieux
385	3423	7	je
8	15	6	déclarer
30	139	6	moment
10	27	6	moments
17	53	6	partir
34	155	6	pleurs
83	579	6	seigneur
8	18	6	séparer
14	53	5	adieu
8	21	5	amoureux
6	14	5	constance
25	118	5	prince
6	18	4	(
6	18	4	)
302	3057	4	?
6	16	4	appartement
8	32	4	chargé
14	65	4	cour
6	20	4	demain
10	44	4	empereur
23	124	4	empire
13	59	4	espoir
10	35	4	grandeur
27	158	4	hélas
132	1196	4	me
11	49	4	mot
5	12	4	pars
31	205	4	puis
16	71	4	univers
9	36	4	vertus
391	4037	4	vous
15	84	3	allons
51	422	3	amour
9	43	3	attendre
5	18	3	aveu
9	42	3	cent
9	41	3	cesse
42	326	3	cette
59	477	3	coeur
5	16	3	couronner
5	17	3	départ
17	98	3	dire
68	573	3	elle
6	23	3	entretien
6	21	3	fuis
30	205	3	jamais
8	36	3	jusques
4	13	3	nécessaire
6	24	3	pouvais
11	58	3	puisse
26	184	3	rien
8	37	3	taire
4	12	3	tourment

Bajazet

64	64	25	bajazet
54	54	23	roxane
34	34	18	amurat
22	22	15	atalide
18	18	13	osmin
17	17	13	sultan
18	18	13	sultane
17	17	13	vizir
15	15	12	zaère
13	13	11	acomat
13	13	11	sérail
12	12	11	sultans
15	25	9	rivale
21	55	7	esclave
72	422	6	amour
7	13	6	lettre
13	30	6	périls
7	15	5	amants
6	12	5	esclaves
7	16	5	malheureuse
17	64	5	ordre
97	681	4	ai
19	80	4	amant
13	51	4	amis
23	105	4	discours
46	275	4	enfin
23	108	4	était
10	38	4	indigne
70	500	4	madame
5	12	4	marque
34	192	4	mort
9	27	4	prompt
7	21	4	prompte
6	17	4	respects
7	21	4	récit
5	12	4	sacrifier
8	25	4	soupçons
6	17	4	tromper
33	173	4	vie
7	23	4	vouloir
323	2912	3	a
5	16	3	adresse
6	21	3	amante
10	44	3	bontés
7	26	3	bout
77	573	3	elle
17	86	3	eût
24	134	3	foi
13	61	3	grâce
13	58	3	ingrat
11	48	3	jaloux
5	14	3	jeunes
17	84	3	malgré
85	648	3	même
102	798	3	moi
12	53	3	palais
5	16	3	préparer
33	205	3	puis
5	14	3	reconnaissance
24	137	3	soins
10	43	3	sortir
5	16	3	tiens
67	495	3	tu

Mithridate

49	49	22	pharnace
50	59	20	romains
31	31	18	xipharès
24	24	16	mithridate
17	17	13	monime
15	15	12	arbate
29	117	6	rome
13	33	6	vaisseaux
7	16	5	prétendre
450	4037	5	vous
30113101		4	,
61	422	4	amour
15	66	4	devoir
10	40	4	dû
50	334	4	fils
135	1066	4	m'
10	34	4	maintenant
17	77	4	malheurs
9	34	4	mêmes
49	321	4	père
15	59	4	place
117	908	4	plus
13	50	4	rival
39	238	4	roi
31	169	4	sais
6	18	4	sentiments
6	19	3	asie
5	14	3	cessez
5	15	3	chemins
15	80	3	dois
17	94	3	ennemi
401	3770	3	et
20	108	3	funeste
5	16	3	histoire
17	88	3	hymen
16	83	3	malheur
141	1196	3	me
138	1162	3	mon
12	53	3	partir
6	21	3	pied
7	24	3	porter
7	24	3	pourtant
8	29	3	prétends
11	48	3	soldats
5	15	3	verrais

Iphigénie

232	586	25	à
58	70	21	achille
40	40	20	calchas
72	131	18	fille
21	21	14	iphigénie
16	19	11	agamemnon
13	13	11	aulide
71	239	11	dieux
15	19	10	arcas
28	62	10	autel
15	20	10	vents
27	77	8	grecs
12	20	8	hélène
20	44	8	troie
20	54	7	camp
145	917	7	ma
9	15	7	oracle
66	321	7	père
8	15	6	argos
7	12	6	félicité
11	24	6	sacrifice
14	37	6	victime
23	87	5	époux
22	88	5	hymen
29	139	5	mère
174	1284	5	qui
119	838	5	votre
14	51	4	armée
8	22	4	daignez
13	49	4	grèce
10	32	4	respect
7	19	4	rougir
51	320	4	sang
462	4037	4	vous
5	14	3	armer
8	31	3	autels
9	36	3	barbare
14	68	3	bonheur
6	19	3	captive
5	15	3	connaît
33	211	3	contre
17	89	3	cruel
8	31	3	fureurs
5	14	3	mer
5	15	3	présenter
14	68	3	prix
5	15	3	rivage
95	748	3	une
9	33	3	vaisseaux
9	35	3	voilà

Phèdre

39	39	20	hippolyte
36	36	20	phèdre
32	34	18	thésée
19	19	14	oenone
17	17	13	aricie
12	12	11	athènes
13	22	8	monstre
11	22	7	bords
45	239	6	dieux
323	2912	5	a
13	45	5	coupable
7	16	5	exil
11	35	5	feu
52	334	5	fils
7	17	5	mortelle
9	35	4	affreux
90	681	4	ai
12	51	4	cacher
8	30	4	crimes
7	23	4	criminel
17	87	4	époux
6	17	4	flots
7	21	4	fuis
5	12	4	fuyant
14	66	4	horreur
6	17	4	horrible
5	12	4	inimitié
129	1057	4	j'
96	764	4	mes
7	22	4	mortel
12	45	4	oeil
7	23	4	offense
8	27	4	remords
5	13	4	reproche
6	15	4	rivage
28	161	4	te
66	495	4	tu
226	2098	4	un
92	748	4	une
5	17	3	accuser
6	21	3	amante
11	52	3	as
5	15	3	charme
5	16	3	corps
4	12	3	enfers
6	22	3	farouche
11	54	3	flamme
8	31	3	fond
7	25	3	frères
5	17	3	innocent
9	41	3	mal
9	40	3	mortels
10	45	3	odieux
7	27	3	osez
4	12	3	pars
50	382	3	quel
6	22	3	redoutable
9	39	3	souvent
25	156	3	tes
25	157	3	toi
20	121	3	voeux

Esther

30	30	21	esther
22	22	18	aman
15	15	15	mardochée
14	15	14	juif
61	246	12	dieu
18	28	12	juifs
14	27	10	cieux
52	238	10	roi
39	160	9	á
9	17	8	persans
11	23	8	sion
8	13	8	soeurs
8	15	7	israël
16	54	7	terre
6	13	6	audacieux
24	110	6	devant
7	16	6	filles
8	21	6	impie
12	41	6	innocence
157	1526	6	les
24	128	6	peuple
11	34	6	race
7	17	6	sage
28	156	6	tes
5	12	5	douce
79	694	5	du
7	23	5	éternelle
212	2305	5	la
19	110	5	notre
13	53	5	palais
5	12	5	pendant
6	17	5	sacrés
6	18	5	saints
24	148	5	ta
7	19	5	terrible
18	118	4	"
6	23	4	avis
12	63	4	bouche
93	943	4	des
4	12	4	dépouille
8	34	4	douceur
12	57	4	front
13	67	4	fut
5	17	4	innocent
26	205	4	jamais
249	2931	4	le
27	188	4	leurs
6	23	4	lumière
5	15	4	méchants
6	22	4	mortel
4	13	4	or
6	25	4	pays
6	23	4	portes
5	16	4	puissants
6	22	4	sceptre
5	17	4	souffre
7	26	4	sujet
22	157	4	toi
20	139	4	ton
6	25	4	tremble
13	71	4	univers
5	15	4	vains
10	45	4	zèle
236	2912	3	a
14	96	3	ainsi
34	312	3	aux
7	36	3	barbare
9	52	3	enfants
18	134	3	heureux
7	36	3	honneurs
6	29	3	humains
7	37	3	intérêt
19	146	3	jours
8	40	3	mortels
5	20	3	nature
4	15	3	pères
10	61	3	porte
5	20	3	puissant
5	22	3	redoutable
5	19	3	soient
5	22	3	sommes
15	103	3	voix

Athalie

156	246	29	dieu
55	72	20	enfant
38	39	19	david
53	79	18	temple
27	27	16	abner
26	26	16	joad
24	24	15	joas
21	21	14	athalie
20	20	14	mathan
17	17	13	josabet
19	20	13	prêtres
15	15	12	eliacin
27	52	11	enfants
13	13	11	jéhu
17	23	11	saint
12	14	10	prêtre
43	160	8	á
42	146	8	rois
11	17	8	sainte
11	21	7	impie
9	15	7	méchants
12	23	7	sion
110	694	6	du
9	20	6	femme
8	14	6	héritier
19	57	6	loi
7	13	6	nourri
8	15	6	parents
9	18	6	saints
85	495	6	sur
17	62	5	autel
9	25	5	es
6	13	5	étrangère
7	15	5	israël
48	276	5	nos
11	34	5	race
29	141	5	reine
46	238	5	roi
14	42	5	songe
7	15	5	troupe
22	118	4	"
323	2912	4	a
10	38	4	bienfaits
36	219	4	cet
9	27	4	cieux
515	4803	4	de
118	943	4	des
5	12	4	encens
25	124	4	grand
11	45	4	heure
16	66	4	horreur
9	28	4	juifs
8	26	4	livre
27	146	4	mains
8	26	4	naissance
25	128	4	peuple
5	12	4	poignard
8	23	4	portes
5	13	4	sacrée
6	17	4	sacrés
5	13	4	soeurs
26	128	4	soit
5	12	4	sortons
30	171	4	temps
26	139	4	ton
5	12	4	vengeur
7	20	4	vérité
13	50	4	voila
6	20	3	aspect
5	16	3	bandeau
27	170	3	comme
6	19	3	commence
28	173	3	déjà
14	65	3	eux
14	68	3	lieu
50	344	3	où
5	15	3	pères
113	915	3	sa
25	148	3	ta
6	19	3	traits

ANNEXE 3 : LES TRAGÉDIES DE CORNEILLE ET RACINE PAR RAPPORT À LA BASE FRANTEXT (TLF)

La liste suivante présente les termes qui distinguent le mieux les tragédies de nos deux auteurs d'un extrait de la base Frantext. Cet extrait est constitué de textes du XIXe et XXe siècles. On compare la fréquence des termes dans notre corpus de tragédies par rapport à la fréquence dans cet extrait de Frantext. En première colonne on lit la fréquence du terme dans le corpus de tragédies et dans la troisième la mesure de l'écart.

1338	seigneur	184.2
743	encor	171.4
284	courroux	160.9
11089	vous	152.0
220	trépas	122.0
1658	vos	111.3
1056	madame	99.3
367	voeux	98.9
628	dieux	90.5
2285	votre	89.9
152	souffrez	86.2
670	gloire	78.1
177	régner	74.5
255	o	73.6
5941	a	72.9
313	trône	72.6
189	tyran	71.1
878	sang	71.0
1464	amour	70.8
5482	?	70.7
456	haine	68.3
208	venger	68.3
2778	ma	67.8
7961	;	67.3
131	sceptre	67.1
3518	mon	65.9
389	rois	64.0
330	époux	62.6
420	crime	62.3
204	funeste	61.3
319	ose	60.1
356	choix	60.0
277	pleurs	60.0
3249	m'	58.7
151	rival	58.6
152	punir	58.3
1442	coeur	56.2
165	soupirs	55.7
1755	point	55.4
135	ingrat	54.5
1723	mes	54.4
179	vainqueur	54.3
804	roi	53.7
305	amant	51.1
8839	je	51.1
361	reine	50.2
104	jusques	49.9
333	digne	49.6
3700	si	49.5
105	perfide	48.4
216	ardeur	47.9
287	lieux	46.7
181	dessein	46.2
105	craignez	45.4
219	perte	44.9
1222	trop	44.4
218	fureur	43.9
147	indigne	43.9
222	craindre	43.8
348	empire	43.8
78	destins	43.5
159	supplice	43.3
256	flamme	43.0
224	rang	42.9
191	malheurs	42.2
114	fers	42.1
738	quel	42.0
152	illustre	41.9
156	romains	41.9
125	périr	41.8
441	sort	41.8
178	vengeance	41.2
177	vôtre	40.8
179	jaloux	40.7
229	soins	40.4
280	vain	40.2
4394	:	39.6
3520	me	39.3
152	cruel	39.2
243	don	39.2
199	destin	38.6
332	dois	38.6
75	fureurs	38.2
543	vois	38.1
160	généreux	37.7
365	prince	37.7
72	exploits	37.4
182	mien	37.4
763	quoi	37.3
620	fils	37.0
67	hait	36.6
2278	moi	36.6
228	victoire	36.6
598	ta	36.2
234	faveur	36.1
661	veux	36.0
106	traître	35.9
770	âme	35.7
653	ciel	35.6
325	vertu	35.1
161	feux	34.8
73	flatte	34.3
113	fatal	34.2
73	outrage	34.2
76	offense	34.1
97	auguste	34.0
77	prompt	34.0
158	crains	33.8
273	espoir	33.5
114	rends	33.2
585	aime	33.0

149	hé	33.0
443	tes	32.9
130	péril	32.7
137	craint	32.6
70	maxime	32.5
107	charmes	32.3
81	périls	32.3
266	hélas	31.9
311	devoir	31.6
81	haïr	31.6
57	rivale	31.5
317	doux	31.4
84	trahir	31.4
85	ordonne	30.6
153	appui	30.5
508	veut	30.3
362	honneur	30.2
93	songez	30.2
119	zèle	30.2
77	desseins	30.1
72	mériter	30.1
1037	faut	29.9
64	soupire	29.8
6498	qu'	29.3
76	artifice	29.2
358	foi	29.0
119	obéir	28.9
212	soin	28.8
194	ennemis	28.7
146	maux	28.6
330	juste	28.5
5346	ce	28.4
785	tant	28.3
365	frère	28.1
80	cède	27.8
176	coeurs	27.8
53	déplaire	27.8
60	infidèle	27.6
98	adore	27.5
83	miens	27.3
3142	j'	27.2
129	plaire	27.2
940	yeux	27.2
188	héros	27.1
115	refus	27.1
355	dieu	26.9
9231	que	26.9
129	aimez	26.8
142	couronne	26.6
727	mort	26.6
678	ton	26.5
58	cruels	26.2
62	jugez	26.2
271	aimer	25.8
129	lâche	25.8
83	barbare	25.6
77	glorieux	25.6
249	courage	25.4
53	infamie	25.4
199	rend	25.4
78	fuit	25.3
62	tyrans	25.3
4605	pour	25.2
177	secours	25.2
207	princesse	25.1
117	esclave	25.0
153	fidèle	24.9
70	gendre	24.9
92	injuste	24.9
177	pouvez	24.9
187	su	24.9
161	laissez	24.8
58	flatter	24.7
53	rebelle	24.7
355	heureux	24.6
171	quels	24.6
45	téméraire	24.6
45	montrez	24.5
166	trouble	24.5
70	transports	24.3
230	colère	24.2
170	éclat	24.2
293	laisse	24.1
292	soeur	24.1
95	cruelle	24.0
250	allez	23.8
252	lois	23.8
44	respects	23.8
94	vaincre	23.8
615	ah	23.7
100	rigueur	23.6
56	aspire	23.3
49	amante	23.2
95	conquête	23.2
159	mépris	23.2
112	victime	23.2
96	audace	23.1
84	honteux	23.1
58	autels	23.0
660	enfin	23.0
248	voyez	22.9
828	voir	22.7
177	honte	22.6
62	bienfaits	22.5
820	nos	22.5
334	rendre	22.5
75	odieux	22.3
7311	en	22.2
697	père	22.2
43	serments	22.1
48	hautement	21.9
142	combat	21.8
188	discours	21.7
210	perdre	21.7
42	accable	21.6
127	désirs	21.6
280	fais	21.5
63	joug	21.5
232	malheur	21.5
54	monarque	21.3
95	crimes	21.2
247	prix	21.2
65	rougir	21.2
178	horreur	21.1
57	mortels	21.1
56	noeuds	21.1
58	perds	21.1
43	ressentiment	21.1
74	tremble	21.1
49	prompte	21.0
42	suivez	21.0
145	mérite	20.9
113	rage	20.9
41	secourir	20.9
53	hais	20.7
81	amants	20.6
219	secret	20.6
101	fuite	20.5
1284	peut	20.3
62	plains	20.3
57	pourrez	20.3
55	saurai	20.3
440	doit	20.1
43	prétends	20.1
71	honneurs	20.0
153	souffrir	20.0
56	criminel	19.9
115	regret	19.9
42	rigoureux	19.9
2664	sa	19.9
45	meure	19.8
100	plaindre	19.8
48	gage	19.7
64	infâme	19.7
76	sitôt	19.7
69	justes	19.6
43	faveurs	19.2
75	abandonne	19.1
225	douleur	19.1
601	main	19.1
366	pouvoir	19.1
555	t'	18.9
37	furie	18.8
47	gardez	18.7
39	contraindre	18.6
118	promis	18.6
50	renommée	18.6
3152	son	18.6
171	orgueil	18.5
38	ravir	18.5
83	devez	18.4
176	puisqu'	18.4
52	soupçons	18.4
104	douleurs	18.3
78	aveu	18.2
40	douceurs	18.2
190	malheureux	18.2
55	meurs	18.2
128	naissance	18.2
129	offre	18.2
87	grecs	18.1
66	obéissance	18.1
168	pitié	18.1
332	donne	18.0
1788	ai	17.9
112	estime	17.9
659	te	17.9
114	temple	17.9
231	mourir	17.8
64	hommage	17.7
45	constance	17.6
40	impie	17.6

104	parlez	17.6
33	disgrâce	17.5
89	effroi	17.5
66	mourant	17.5
86	taire	17.5
40	horreurs	17.4
44	insolence	17.4
67	romain	17.4
343	vient	17.4
128	toutefois	17.3
96	choisir	17.2
60	éclater	17.2
40	ravie	17.2
44	épargner	17.1
48	injure	17.1
35	recevez	17.1
94	sévère	17.1
61	mortel	17.0
1499	tu	16.9
182	amitié	16.8
164	ennemi	16.8
51	ennemie	16.8
1239	tous	16.8
43	triompher	16.8
111	vertus	16.8
116	attend	16.7
118	états	16.7
119	fasse	16.7
42	tourments	16.7
98	assurer	16.6
78	arrêt	16.5
93	fruit	16.5
290	maître	16.5
90	secrets	16.5
70	comble	16.4
48	dépens	16.4
67	injustice	16.4
33	supplices	16.4
219	faites	16.3
260	malgré	16.3
50	pourriez	16.3
139	respect	16.3
205	viens	16.3
34	insolent	16.2
57	troubler	16.2
434	nom	16.1
87	princes	16.1
116	sauver	16.0
214	voulez	16.0
31	enfers	15.9
1488	fait	15.9
41	percer	15.9
37	puni	15.9
48	sénat	15.9
42	apprends	15.8
37	croître	15.8
40	vains	15.8
35	redouter	15.7
35	apaiser	15.6
99	coupable	15.6
62	foudre	15.6
35	frayeur	15.6
184	puisse	15.6
31	aveuglement	15.5
144	cherche	15.5
83	dépit	15.5
51	prétendre	15.5
53	vaine	15.5
677	cet	15.4
61	trembler	15.4
63	combats	15.3
86	camp	15.2
77	menace	15.2
480	avez	15.1
48	jalouse	15.1
72	obstacle	15.1
96	sujets	15.1
35	trahit	15.1
40	consentir	15.0
86	douter	15.0
30	évite	15.0
713	quelque	15.0
32	dérobe	14.9
249	grâce	14.9
94	souffre	14.9
61	sûreté	14.9
183	effort	14.8
55	forcer	14.8
48	ôter	14.8
83	conseils	14.6
33	portez	14.6
70	verra	14.6
52	dignes	14.5
40	illustres	14.5
63	vaincu	14.5
38	déplorable	14.4
76	innocence	14.4
29	ordonner	14.4
129	noble	14.3
51	oppose	14.3
88	prenez	14.3
59	prépare	14.3
465	sais	14.3
212	objet	14.2
93	offrir	14.2
49	ennuis	14.1
38	rendra	14.1
149	suivre	14.1
101	faiblesse	14.0
38	promettre	14.0
28	surprend	14.0
60	fierté	13.9
699	grand	13.9
36	promet	13.9
61	résoudre	13.9
269	tel	13.9
31	combattu	13.8
35	condamne	13.8
37	mérité	13.8
63	monstre	13.8
38	ôte	13.8
85	pensez	13.8
41	verser	13.7
379	bras	13.6
37	cherchez	13.6
37	lâches	13.6
63	pourrai	13.6
99	sert	13.6
144	servir	13.6
407	vu	13.6
29	couronné	13.5
269	peuple	13.5
38	renonce	13.5
39	allégresse	13.4
79	ambition	13.4
51	joindre	13.4
744	moins	13.4
479	toi	13.4
194	triste	13.4
103	défendre	13.3
77	donnez	13.3
1653	sans	13.3
133	armes	13.2
285	bonheur	13.2
28	criminelle	13.2
113	désespoir	13.2
30	irrite	13.2
30	abuser	13.1
67	arracher	13.1
242	joie	13.1
26	mourrai	13.1
72	perd	13.1
89	prête	13.1
67	sien	13.1
164	adieu	13.0
481	assez	13.0
33	épargne	13.0
364	mains	13.0
31	vouliez	13.0
69	conserver	12.9
551	contre	12.9
72	autel	12.8
56	exil	12.8
432	mieux	12.8
344	pu	12.8
86	sacrifice	12.8
68	sire	12.8
70	aisément	12.7
62	choeur	12.7
31	expose	12.7
85	naître	12.7
169	puissance	12.7
89	règne	12.7
41	repentir	12.7
30	vanter	12.7
49	verrai	12.7
137	croit	12.6
45	répandre	12.6
172	cours	12.5
120	crainte	12.5
35	dût	12.5
36	irrité	12.5
203	puisque	12.5
79	excès	12.4
59	rompre	12.4
48	achève	12.3
84	presse	12.3
64	souverain	12.3
38	unir	12.3
104	venez	12.3
85	aveugle	12.2
3854	n'	12.2

44	poison	12.2
27	suffrage	12.2
35	trahi	12.2
25	victorieux	12.2
34	douteux	12.1
32	émouvoir	12.1
28	obliger	12.1
45	surprendre	12.1
30	arbitre	12.0
64	attache	12.0
56	nommer	12.0
93	prêt	12.0
27	consent	11.9
65	entretien	11.9
25	garantir	11.9
165	justice	11.9
1227	leur	11.9
77	mienne	11.9
96	prends	11.9
44	saura	11.9
39	accuser	11.8
192	allons	11.8
102	paraître	11.8
58	témoins	11.8
61	verrez	11.8
32	dérober	11.7
57	grâces	11.7
176	larmes	11.7
76	quitte	11.7
48	respire	11.7
50	confus	11.6
66	espérer	11.6
36	oblige	11.6
178	paix	11.6
87	plaît	11.6
85	cacher	11.5
28	déja	11.5
483	seul	11.5
55	soupir	11.5
24	verrais	11.5
35	châtiment	11.4
59	combattre	11.4
41	éclate	11.4
51	légitime	11.4
28	perdant	11.4
84	triomphe	11.4
88	attends	11.3
33	encens	11.3
104	espère	11.3
234	ait	11.2
846	ont	11.2
44	prêts	11.2
302	quelle	11.1
52	siens	11.1
72	tombeau	11.1
88	bonté	11.0
137	vouloir	11.0
72	cieux	10.9
28	percé	10.9
23	rivaux	10.9
29	salaire	10.9
39	disposer	10.8
33	marques	10.8
47	remède	10.8
37	dispose	10.7
54	partage	10.7
35	arrache	10.6
168	loi	10.6
50	achever	10.5
45	innocent	10.5
43	querelle	10.5
100	tendresse	10.5
23	blesse	10.4
23	offerte	10.4
26	osais	10.4
32	vaincus	10.4
55	emporte	10.3
40	exposer	10.3
49	permettez	10.3
74	puissant	10.3
23	saurez	10.3
32	tourment	10.3
38	veuille	10.3
98	douceur	10.2
62	épouse	10.2
494	ici	10.2
35	joint	10.2
22	taureaux	10.2
55	cache	10.1
23	criminels	10.1
35	défend	10.1
26	funestes	10.1
290	lieu	10.1
401	va	10.1
24	assassins	10.0
221	cher	10.0
147	coups	10.0
36	esclavage	10.0
100	fera	10.0
91	grandeur	10.0
28	grandeurs	10.0
25	ingratitude	10.0
39	inspire	10.0
28	nuire	10.0
163	savez	10.0
40	souhaite	10.0
120	univers	10.0
92	avis	9.9
121	connais	9.9
101	croyez	9.9
25	infortuné	9.9
109	montrer	9.9
132	attendre	9.8
38	engage	9.8
89	erreur	9.8
25	honorer	9.8
20	offenser	9.8
36	quiconque	9.8
40	traite	9.8
33	vole	9.8
25801	.	9.7
54	épouser	9.7
27	haines	9.7
31	noeud	9.7
34	plaint	9.7
68	reconnaissanc	9.7
39	trahison	9.7
117	voyant	9.7
45	attendez	9.6
201	bientôt	9.6
25	éternels	9.6
54	étonne	9.6
39	loisir	9.6
56	pareils	9.6
23	rende	9.6
25	renvoie	9.6
23	sortez	9.6
28	abattu	9.5
20	croirais	9.5
88	espérance	9.5
2697	mais	9.5
24	sachez	9.5
24	salutaire	9.5
51	soumis	9.5
27	souvenez	9.5
37	tienne	9.5
25	trompée	9.5
76	violence	9.5
24	défends	9.4
36	fatale	9.4
135	fortune	9.4
27	partez	9.4
30	présents	9.4
44	puissent	9.4
24	sanglant	9.4
20	servie	9.4
57	témoin	9.4
29	anime	9.3
25	charmé	9.3
118	cru	9.3
36	éteindre	9.3
23	excite	9.3
22	faille	9.3
24	naissant	9.3
80	pourra	9.3
35	ravi	9.3
56	tristes	9.3
37	volontés	9.3
326	état	9.2
21	frivole	9.2
27	hâter	9.2
25	innocents	9.2
33	souveraine	9.2
71	suit	9.2
24	condamnée	9.1
24	divorce	9.1
86	droits	9.1
44	gardes	9.1
23	incertaine	9.1
33	armé	9.0
61	avantage	9.0
81	efforts	9.0
102	jusque	9.0
253	ordre	9.0
48	osé	9.0
27	respecte	9.0
46	satisfaire	9.0
19	captif	8.9
34	insensible	8.9
31	oubliez	8.9
40	partager	8.9
34	provinces	8.9

25	séduire	8.9
41	voudra	8.9
47	auriez	8.8
22	bourreaux	8.8
65	égal	8.8
21	emportement	8.8
97	entends	8.8
43	excuse	8.8
34	pardonnez	8.8
311	peine	8.8
26	réduite	8.8
84	sois	8.8
18	abaisse	8.7
20	acquitter	8.7
19	murmurer	8.7
18	offensé	8.7
32	puissants	8.7
46	vaisseaux	8.7
58	assure	8.6
915	autre	8.6
25	climats	8.6
24	honteuse	8.6
60	ignore	8.6
22	infortune	8.6
55	précieux	8.6
28	répandu	8.6
111	tantôt	8.6
32	transport	8.6
25	versé	8.6
37	amène	8.5
87	aurai	8.5
24	doive	8.5
20	honore	8.5
32	souvient	8.5
26	vainqueurs	8.5
140	dites	8.4
39	vienne	8.4
213	voit	8.4
35	assassin	8.3
22	disputer	8.3
283	effet	8.3
32	lâcheté	8.3
67	pourrais	8.3
18	souhait	8.3
20	soulager	8.3
18	alarme	8.2
29	confidence	8.2
135	connaître	8.2
45	sauvé	8.2
32	tyrannie	8.2
68	affreux	8.1
22	conquêtes	8.1
37	exprès	8.1
26	implacable	8.1
50	miracle	8.1
55	remords	8.1
69	rendez	8.1
31	retient	8.1
33	romaine	8.1
26	adieux	8.0
19	aversion	8.0
41	céder	8.0
148	cour	8.0
13237	et	8.0
59	éviter	8.0
83	extrême	8.0
29	guerrier	8.0
83	ordres	8.0
24	vôtres	8.0
37	consoler	7.9
26	entreprendre	7.9
62	envoie	7.9
21	gémir	7.9
31	sacrés	7.9
36	accorder	7.8
38	commande	7.8
28	commander	7.8
104	entière	7.8
30	instruit	7.8
29	oser	7.8
42	prévenir	7.8
63	rare	7.8
36	serment	7.8
21	suspect	7.8
37	tonnerre	7.8
110	valeur	7.8
96	aimé	7.7
33	balance	7.7
48	dangereux	7.7
20	examine	7.7
50	marque	7.7
108	porter	7.7
35	trouvez	7.7
82	songe	7.6
36	alliance	7.5
23	incertain	7.5
25	méchants	7.5
26	monstres	7.5
46	murmure	7.5
180	parle	7.5
43	parts	7.5
32	réduire	7.5
318	reste	7.5
70	effets	7.4
44	égale	7.4
26	instruire	7.4
135	intérêt	7.4
31	laissons	7.4
23	pressant	7.4
16	puissiez	7.4
73	saurait	7.4
29	accuse	7.3
142	aura	7.3
595	donc	7.3
263	donner	7.3
23	due	7.3
74	garder	7.3
20	interprète	7.3
32	justifier	7.3
26	pompe	7.3
44	prudence	7.3
28	acquérir	7.2
36	attacher	7.2
34	farouche	7.2
17	flatté	7.2
50	impatience	7.2
22	souverains	7.2
17	tien	7.2
41	faibles	7.1
33	récompense	7.1
26	scrupule	7.1
100	succès	7.1
17	tarder	7.1
64	courir	7.0
33	défaite	7.0
109	envie	7.0
84	faible	7.0
103	mémoire	7.0
25	miracles	7.0
16	ombrage	7.0
108	sentiments	7.0
18	bienheureux	6.9
31	cendre	6.9
41	humains	6.9
99	palais	6.9
26	poignard	6.9
58	touche	6.9
32	victimes	6.9
16	adoucir	6.8
59	change	6.8
18	conçois	6.8
262	dis	6.8
16	éclaircir	6.8
249	êtes	6.8
20	impitoyable	6.8
17	inflexible	6.8
17	jurer	6.8
342	mal	6.8
18	odieuse	6.8
22	presser	6.8
15	approuver	6.7
114	chère	6.7
25	complice	6.7
16	descendue	6.7
60	dirai	6.7
88	faux	6.7
68	intérêts	6.7
42	jouir	6.7
88	montre	6.7
4575	ne	6.7
34	reconnais	6.7
72	rendu	6.7
21	rudes	6.7
17	tâchez	6.7
16	traîtres	6.7
21	allume	6.6
26	char	6.6
30	confondre	6.6
25	cruauté	6.6
111	empereur	6.6
24	excuser	6.6
936	faire	6.6
137	garde	6.6
19	gêner	6.6
86	maîtresse	6.6
23	mortelle	6.6
66	permis	6.6
3185	plus	6.6
43	refuse	6.6
94	sein	6.6
40	tue	6.6
78	voie	6.6

37	accepte	6.5
37	admire	6.5
36	assurance	6.5
24	défiance	6.5
17	détestable	6.5
17	épris	6.5
41	jalousie	6.5
36	liens	6.5
44	nomme	6.5
20	séduit	6.5
87	soldats	6.5
79	titre	6.5
51	aviez	6.4
14	avouez	6.4
51	explique	6.4
34	pardonne	6.4
34	redoutable	6.4
28	restes	6.4
43	ruine	6.4
26	sacrifier	6.4
44	sincère	6.4
23	tendresses	6.4
36	timide	6.4
65	arrête	6.3
74	arrêter	6.3
28	beautés	6.3
23	choisit	6.3
44	connaissez	6.3
51	désordre	6.3
15	éperdue	6.3
37	intéresse	6.3
15	pouviez	6.3
85	regards	6.3
20	répondez	6.3
41	secrète	6.3
22	souhaiter	6.3
23	accablé	6.2
53	accepter	6.2
34	assuré	6.2
17	audacieux	6.2
209	autant	6.2
16	cherchons	6.2
29	contrainte	6.2
33	couler	6.2
33	coûte	6.2
21	coûté	6.2
64	destinée	6.2
17	dragon	6.2
64	ennui	6.2
33	éteint	6.2
26	guerriers	6.2
15	perce	6.2
18	renaître	6.2
58	soyez	6.2
17	sus	6.2
96	tiens	6.2
16	timides	6.2
35	verrons	6.2
44	auront	6.1
151	chercher	6.1
26	félicité	6.1
31	flanc	6.1
18	fuyant	6.1
22	généreuse	6.1
112	présence	6.1
36	promesse	6.1
85	recevoir	6.1
63	agir	6.0
33	asile	6.0
26	avertir	6.0
17	cruelles	6.0
23	déesse	6.0
55	pleurer	6.0
25	promptement	6.0
259	seule	6.0
19	adorer	5.9
23	atteinte	5.9
40	ayez	5.9
31	brûle	5.9
55	chagrin	5.9
24	craintes	5.9
17	dépouille	5.9
52	écouter	5.9
35	flammes	5.9
91	met	5.9
35	moindres	5.9
13	remerciements	5.9
19	soupçonner	5.9
174	vais	5.9
22	attachée	5.8
17	contents	5.8
13	dispense	5.8
17	dissiper	5.8
70	écoute	5.8
30	fière	5.8
17	infaillible	5.8
13	intrépide	5.8
24	pourront	5.8
32	préfère	5.8
14	profane	5.8
17	revivre	5.8
17	sanglante	5.8
45	serez	5.8
26	seriez	5.8
45	soutenir	5.8
16	trêve	5.8
21	vertueux	5.8
32	arme	5.7
19	assurée	5.7
48	aurez	5.7
26	caprice	5.7
16	craignais	5.7
55	défense	5.7
38	échappe	5.7
13	équité	5.7
14	goûte	5.7
20	héritier	5.7
23	innocente	5.7
59	surpris	5.7
63	aimable	5.6
40	fameux	5.6
14	légitimes	5.6
39	livrer	5.6
25	méprise	5.6
21	pleinement	5.6
15	redire	5.6
54	suivi	5.6
48	bords	5.5
20	entretiens	5.5
28	générosité	5.5
13	impuissante	5.5
17	irais	5.5
74	juge	5.5
18	paternel	5.5
78	quitter	5.5
70	race	5.5
71	raisons	5.5
44	reproche	5.5
13	sacrilège	5.5
23	soutient	5.5
78	suffit	5.5
13	bénir	5.4
109	cesse	5.4
48	changement	5.4
381	déjà	5.4
33	favorable	5.4
18	fiers	5.4
18	immortelle	5.4
13	imprime	5.4
29	plainte	5.4
17	redoute	5.4
40	réduit	5.4
13	remets	5.4
34	réponds	5.4
13	servira	5.4
101	tient	5.4
22	verrait	5.4
21	adorable	5.3
30	crédit	5.3
69	demeure	5.3
15	hasards	5.3
23	ignorer	5.3
19	nourri	5.3
60	obtenir	5.3
33	parlons	5.3
16	précipiter	5.3
13	prescrit	5.3
15	résolue	5.3
22	soupçon	5.3
636	suis	5.3
26	suivie	5.3
22	attire	5.2
15	impatient	5.2
13	impunément	5.2
62	inutile	5.2
47	irai	5.2
16	préférer	5.2
53	amoureux	5.1
249	aujourd'	5.1
12	conquérant	5.1
66	désormais	5.1
32	détruit	5.1
17	deviens	5.1
17	dévore	5.1
58	frères	5.1
249	hui	5.1
12	irritée	5.1
12	mourra	5.1
16	obtient	5.1
26	opposer	5.1
77	ouvrage	5.1
49	plaisirs	5.1

19	promets	5.1
79	repos	5.1
39	souci	5.1
24	splendeur	5.1
27	tragédie	5.1
43	attente	5.0
48	conduire	5.0
12	confesse	5.0
36	dépend	5.0
13	eussiez	5.0
306	eût	5.0
12	exposée	5.0
24	feront	5.0
13	honoré	5.0
24	invincible	5.0
124	partout	5.0
149	plutôt	5.0
14	promise	5.0
62	sensible	5.0
58	suprême	5.0
26	vigueur	5.0
47	voudrait	5.0
35	abandonner	4.9
18	creuse	4.9
159	croire	4.9
18	dépôt	4.9
24	échappé	4.9
28	engager	4.9
27	entretenir	4.9
15	festin	4.9
54	jette	4.9
23	poursuit	4.9
1695	ses	4.9
17	sultan	4.9
27	chers	4.8
12	combattants	4.8
32	confusion	4.8
12	délivre	4.8
15	détours	4.8
20	dissimuler	4.8
140	exemple	4.8
15	faiblesses	4.8
18	fermeté	4.8
20	forcée	4.8
24	franchise	4.8
127	laisser	4.8
12	méritait	4.8
13	parer	4.8
31	prétend	4.8
37	refuser	4.8
19	suivis	4.8
14	tarde	4.8
73	traits	4.8
36	troupe	4.8
7023	un	4.8
18	arraché	4.7
43	avoue	4.7
24	chaînes	4.7
16	commandé	4.7
24	conserve	4.7
22	consulter	4.7
15	détourne	4.7
14	divins	4.7
25	exécuter	4.7
24	ferez	4.7
12	gages	4.7
65	nouveaux	4.7
12	ôté	4.7
59	oublier	4.7
16	prendra	4.7
26	saurais	4.7
12	soutien	4.7
38	superbe	4.7
42	aimais	4.6
12	ajoutez	4.6
16	assassinat	4.6
19	bourreau	4.6
14	captivité	4.6
16	chères	4.6
13	complices	4.6
42	fier	4.6
36	gêne	4.6
30	hauts	4.6
15	mette	4.6
827	quand	4.6
21	reprocher	4.6
31	séparer	4.6
25	sexe	4.6
13	successeur	4.6
18	vaines	4.6
48	brave	4.5
59	changer	4.5
58	charme	4.5
15	clartés	4.5
14	conter	4.5
17	déteste	4.5
13	étale	4.5
49	ferai	4.5
17	flambeau	4.5
74	heureuse	4.5
12	impuissant	4.5
86	morts	4.5
12	négliger	4.5
24	retire	4.5
110	retour	4.5
12	tourmenter	4.5
14	triomphant	4.5
15	aïeux	4.4
42	aimée	4.4
24	complaisance	4.4
24	conte	4.4
21	déclarer	4.4
15	foire	4.4
13	frémit	4.4
50	gagner	4.4
28	juifs	4.4
363	soit	4.4
20	sors	4.4
14	souhaité	4.4
58	apprendre	4.3
14	ardents	4.3
23	attaquer	4.3
14	camps	4.3
11	consulté	4.3
14	défait	4.3
17	étonnée	4.3
33	faisons	4.3
162	font	4.3
25	guérir	4.3
22	incertitude	4.3
72	peuples	4.3
12	puissamment	4.3
13	recourir	4.3
18	reculer	4.3
20	rendue	4.3
45	retraite	4.3
13	rougeur	4.3
14	satisfaite	4.3
110	sortir	4.3
125	sujet	4.3
26	déclare	4.2
33	détruire	4.2
27	éloigner	4.2
39	majesté	4.2
21	nôtres	4.2
124	pourrait	4.2
17	poursuite	4.2
31	préparer	4.2
13	prodige	4.2
16	remparts	4.2
28	rendent	4.2
176	sait	4.2
35	tromper	4.2
11	animer	4.1
22	attendent	4.1
142	aurais	4.1
14	bassesse	4.1
20	chagrins	4.1
17	condamner	4.1
18	détourner	4.1
36	écoutez	4.1
41	épée	4.1
66	esprits	4.1
12	frein	4.1
66	moments	4.1
38	orage	4.1
14	plaira	4.1
12	priver	4.1
25	rivage	4.1
30	satisfait	4.1
84	soudain	4.1
344	t	4.1
59	tels	4.1
148	voulu	4.1
96	afin	4.0
48	content	4.0
25	destinées	4.0
201	grands	4.0
88	haute	4.0
21	lasse	4.0
60	noms	4.0
12	ouvrez	4.0
222	prendre	4.0
22	reçue	4.0
32	remplir	4.0
40	servi	4.0
55	adresse	3.9
11	attendrai	3.9
14	blâme	3.9
11	bûcher	3.9
67	cris	3.9
13	embrassant	3.9

11	estimer	3.9
48	finir	3.9
16	imprudence	3.9
49	juger	3.9
36	mets	3.9
14	plonge	3.9
39	proie	3.9
10	racheter	3.9
12	rendrait	3.9
11	retarder	3.9
14	retenue	3.9
771	rien	3.9
11	trompez	3.9
139	vivre	3.9
23	conservé	3.8
16	contrée	3.8
60	dedans	3.8
13	diffère	3.8
26	ferais	3.8
10	nymphes	3.8
11	pâlir	3.8
12	plonger	3.8
45	récit	3.8
19	réparer	3.8
32	rude	3.8
11	semé	3.8
32	vrais	3.8
11	adultère	3.7
108	combien	3.7
11	combler	3.7
14	compassion	3.7
16	direz	3.7
12	feriez	3.7
528	jour	3.7
37	nobles	3.7
52	prière	3.7
34	reçoit	3.7
21	reçus	3.7
13	réservée	3.7
21	tremblant	3.7
15	troublée	3.7
14	vil	3.7
22	voudrez	3.7
16	appelez	3.6
13	blesser	3.6
14	diligence	3.6
10	donnons	3.6
19	dons	3.6
15	dureté	3.6
24	emporter	3.6
47	éternel	3.6
10	favoriser	3.6
18	imiter	3.6
10	lustre	3.6
28	mers	3.6
10	orgueilleuse	3.6
11	osant	3.6
10	persévérance	3.6
20	plaintes	3.6
16	plu	3.6
11	prendrai	3.6
36	retenir	3.6
27	sache	3.6
2105	tout	3.6
10	triomphante	3.6
18	accueil	3.5
18	aisé	3.5
143	bruit	3.5
18	chéri	3.5
17	choisis	3.5
19	communs	3.5
15	conquis	3.5
12	couronnes	3.5
9	déploie	3.5
17	écouté	3.5
11	effrayer	3.5
10	emplois	3.5
15	fâcheux	3.5
32	furieux	3.5
22	obstacles	3.5
13	promit	3.5
33	services	3.5
9	signaler	3.5
29	traiter	3.5
21	voler	3.5
138	demande	3.4
17	juré	3.4
15	ménager	3.4
160	mettre	3.4
10	murmures	3.4
555	puis	3.4
42	remettre	3.4
14	repousser	3.4
16	respecter	3.4
103	souvenir	3.4
43	tâche	3.4
11	approuve	3.3
25	attendais	3.3
16	batailles	3.3
18	borne	3.3
22	briser	3.3
9	comblé	3.3
20	crus	3.3
24	enlever	3.3
13	excusez	3.3
13	frémir	3.3
122	front	3.3
11	fui	3.3
33	fuir	3.3
23	inutiles	3.3
28	jure	3.3
15	menaces	3.3
45	paru	3.3
22	prêter	3.3
10	prêtes	3.3
16	rendant	3.3
18	revers	3.3
25	sépare	3.3
20	trésors	3.3
65	unique	3.3
32	vents	3.3
37	absolu	3.2
32	avouer	3.2
9	cédant	3.2
31	choisi	3.2
42	croyais	3.2
12	deviendra	3.2
33	emploi	3.2
10	enferme	3.2
11	fassent	3.2
15	mépriser	3.2
67	moitié	3.2
14	obéit	3.2
212	parler	3.2
35	projets	3.2
13	prononce	3.2
14	ranger	3.2
19	remise	3.2
10	trame	3.2
19	troubles	3.2
19	verras	3.2
15	verse	3.2
9	vît	3.2
221	cependant	3.1
15	console	3.1
48	côtés	3.1
37	couvert	3.1
191	crois	3.1
11	déchirer	3.1
12	drapeaux	3.1
33	embrasser	3.1
10	empare	3.1
10	emprunter	3.1
9	enlevée	3.1
24	esclaves	3.1
19	essai	3.1
34	fausse	3.1
15	froideur	3.1
20	impuissance	3.1
12	maximes	3.1
14	résiste	3.1
36	vieil	3.1
13	barbarie	3.0
12	blessée	3.0
15	brille	3.0
9	licence	3.0
14	lie	3.0
1613	même	3.0
15	postérité	3.0
23	regrets	3.0
24	résolu	3.0

SOMMAIRE DES FIGURES

Figure 1 : Indicateurs statistiques de caractérisation des corpus 11
Figure 2 : Distribution des fréquences sur les quatre corpus 16
Figure 3 : Distribution des fréquences sur les cinq corpus (double logarithme) 17
Figure 4 : Régressions sur la distribution des fréquences : variation de R2 en fonction des seuils 18
Figure 5 : Régressions sur la distribution des fréquences : variation de la pente en fonction des seuils 19
Figure 6 : Droites de régression correspondant aux cinq corpus 19
Figure 7 : Distribution des fréquences de 1 à 100 avec et sans lemmatisation 23
Figure 8 : Méthode de classification en fonction des profils lexico-sémantiques 30
Figure 9 : Classification descendante sur quatre pièces 33
Figure 10 : Projection sur le premier plan factoriel des oeuvres de Racine (analyse 1) 39
Figure 11 : Champs lexicaux dans les oeuvres de Racine 41
Figure 12 : Projection sur le premier plan factoriel des oeuvres de Racine (analyse 2) 43
Figure 13 : Projection sur le second plan factoriel des oeuvres de Racine (analyse2) 43
Figure 14 : Profils lexicaux-sémantiques des oeuvres de Racine 44
Figure 15 : Accroissement du vocabulaire : écarts réduits (Ch. Bernet) 46
Figure 16 : Accroissement du vocabulaire : données observées par rapport aux données théoriques (D. Labbé) 48
Figure 17 : Les thématiques de la tragédie : Corneille et Racine 52
Figure 18 : Répartition des thèmes selon les pièces : Corneille et Racine 53

LES MOTIVATIONS DES VOLONTAIRES BÉNÉVOLES À UNE GRANDE ÉTUDE ÉPIDÉMIOLOGIQUE

Pascale HÉBEL

Je remercie vivement Saadi Lahlou pour ses précieuses relectures ainsi que le Docteur Serge Hercberg qui a fourni le matériel nécessaire à l'analyse.

Sommaire

1. LE CONTEXTE 81

2. TENDANCES GÉNÉRALES DU DISCOURS 83

3. TYPOLOGIE DES BÉNÉVOLES 86

3.1. Intérêt pour le thème de l'étude 88

3.2. L'altruisme déclaré 92

4. CHEMINEMENT DES MOTIVATIONS 97

CONCLUSION 101

BIBLIOGRAPHIE 102

SOMMAIRE DES FIGURES 104

1. LE CONTEXTE

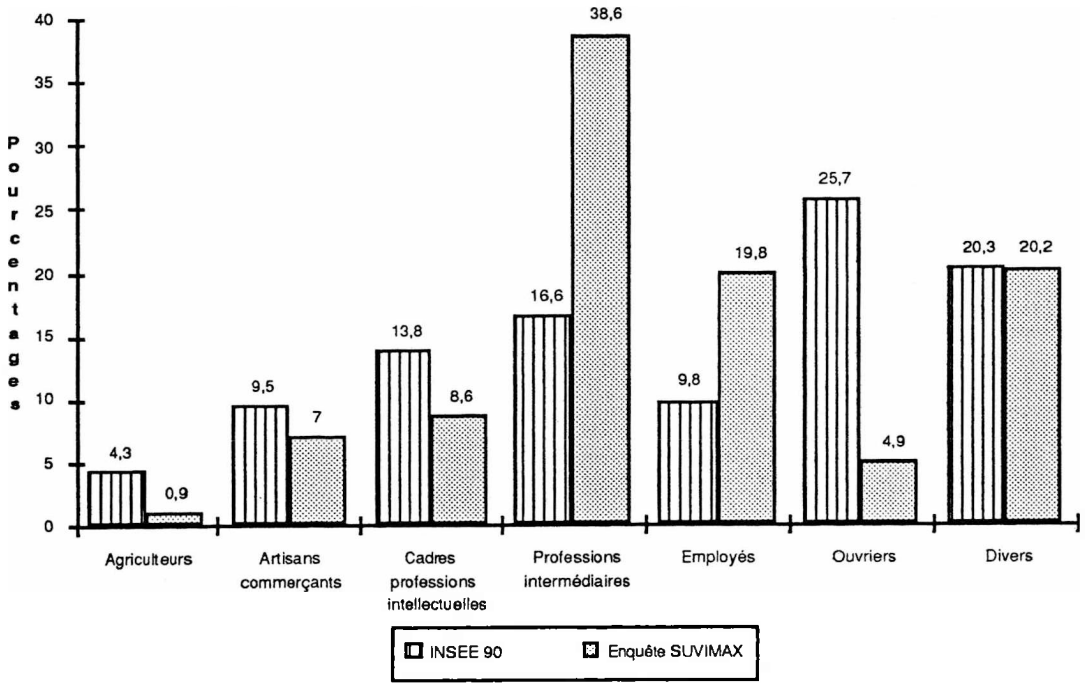
SU.VI.MAX (SUplémentation en VItamines et Minéraux AntioXydants) est un projet d'étude épidémiologique de grande ampleur. Ses objectifs sont doubles :

- préciser les relations existant entre l'alimentation, la santé et les grands problèmes de santé publique rencontrés en France ;
- tester l'efficacité d'une intervention nutritionnelle (apport supplémentaire de vitamines et minéraux antioxydants à des doses nutritionnelles) pour la prévention de la mortalité et des grandes maladies (cancers, maladies cardio-vasculaires, cataracte, infections...). L'étude SU.VI.MAX implique 100 000 sujets adultes volontaires dont 15 000 seront suivis pendant huit ans.

La phase de test méthodologique prévoit l'étude de divers aspects du protocole sur 1 000 sujets, notamment l'absorption de capsules (vitaminées ou placebo) et l'utilisation du système d'enquête par Minitel...

Pour la mener à bien, un appel à volontaires a été lancé le 18 janvier 1993 par une campagne de presse. 13 000 volontaires ont répondu et rempli un questionnaire. Parmi ceux-ci, 1 005 ont été retenus, par tirage sur la base de quotas régionaux.
L'échantillon des volontaires est caractérisé par une surreprésentation des femmes appartenant aux professions intermédiaires et une sous représentation des ouvriers et des agriculteurs (Hercberg et al. (1994) et graphique n°1).

Graphique n°1 : Distribution comparée de la population des volontaires par rapport à celle de la population générale en 1990 en fonction des CSP



On propose aux volontaires d'exprimer les motifs de leur adhésion à un tel projet qui sollicite une assiduité permanente et un investissement conséquent si l'individu est choisi pour participer à l'étude proprement dite qui durera huit ans. Pour cerner les motivations des bénévoles la question ouverte suivante a été insérée dans le formulaire de consentement :

"Expliquez brièvement pourquoi vous souhaitez participer à cette étude"

Aucune variable sur l'appartenance religieuse n'a été insérée dans le questionnaire et ne pourra donc apparaître comme variable explicative de l'implication dans cette enquête.
Cette partie du rapport analyse les réponses à cette question ouverte en utilisant une nouvelle méthode statistique basée sur l'analyse lexicale. Dans la première section, une analyse globale du type de réponses est proposée. Dans la seconde, une typologie des individus est dressée et permet de proposer un schéma du fonctionnement des motivations des bénévoles.

2. TENDANCES GÉNÉRALES DU DISCOURS

Les personnes interrogées ont bien compris la question et y ont répondu avec application. Le taux de non-réponse est très faible, les réponses sont variées et longues. En voici des exemples :

> * Depuis quelques années, je suis partisan d'une alimentation saine et équilibrée que je complète par un apport en vitamines A, E et C ainsi qu'en oligo-éléments. De plus en tant que grand-père je désire sauvegarder la vie de mes petits enfants des maladies les plus mortelles actuellement. J'espère vivement que ma candidature sera retenue afin d'aider la recherche.

> * Parce que c'est le seul type d'études fiables en grandeur réelle dont les données puissent infirmer ou confirmer les études et recherches cliniques en hôpital, et parce que je souhaite me sentir utile à la recherche médicale dans un but de pouvoir agir avant les accidents et maladies par la prévention qui est trop peu développée en France.

Pour analyser les 1 005 réponses, nous utilisons l'analyse lexicale (Reinert, 1983, Lahlou 1989, Beaudouin et Lahlou, 1993). Cette méthode permet de classer ensemble les phrases qui contiennent des mots ayant la même racine[1].

Le tableau n°1 fournit la liste des mots les plus fréquents dans les réponses.

---

[1] Les racines désignent les formes lemmatisées. Par exemple *"vitamine"* désigne : vitamine, vitamines, vitaminé et vitaminés.

Tableau n°1 : Fréquence des racines

Fréquence	Mots	Fréquence	Racine	Fréquence	Racine
383	recherche	86	prendre	43	père
288	santé	82	problème	43	convaincu
286	intérêt	81	utile	42	famille
276	aliment	71	projet	41	traite
254	faire	67	expérience	41	particulier
249	participer	67	suivre	40	moyen
231	vitamine	67	connaître	40	importance
213	maladie	64	donné	40	général
213	étude	63	contribution	39	sujet
188	médical	62	nutrition	39	cardio-vasculaire
186	aide	59	risque	38	sens
177	cancer	59	décédé	38	carence
162	an	58	persuade	37	travail
134	apport	57	équilibre	37	aime
120	prévention	55	mère	36	sensibilisé
112	médecine	55	important	36	active
112	bon	55	permettre	35	temps
96	oligo-éléments	54	trouve	35	meilleur
96	grand	50	rôle	34	sang
95	science	50	actuel	34	progresse
92	souhait	49	domaine	34	hygiène
89	vie	48	personnel	34	désir
89	bien	47	concerne	34	action
87	minéral	47	améliore	33	pouvoir
87	avance	44	enfant	33	nutritionnel

Les thèmes qui apparaissent dans les raisons invoquées sont directement liés au sujet de l'étude. La fréquence du terme "recherche" montre une prise à leur compte des objectifs de l'étude par les individus. Les grandes tendances qui se dégagent des réponses sont les suivantes.

Pour la recherche médicale

En toile de fond, la motivation principale est un intérêt pour la recherche médicale. Ceci est cohérent avec le comportement des Français, puisqu'en 1994, 61 % des dons ont été destinés au secteur santé (Collerie de Borely, 1994) et qu'en 1990, ils n'étaient que de 54 % (Nisak, 1992). On observe ici que le cancer est mentionné quatre fois plus que les maladies cardio-vasculaires. Ceci est cohérent avec le fait que les Français donnent plus pour la lutte contre le cancer que pour le combat contre les maladies cardio-vasculaires, dont la fréquence et le taux de morbidité sont pourtant bien plus grands (Nisak, 1992). Faut-il y voir une influence de la "part de voix" médiatique de ce sujet ?

Pour une meilleure alimentation

Plus d'un quart des bénévoles citent l'alimentation, les vitamines ou les minéraux. Le sujet même du projet, à savoir l'étude de l'effet d'un apport vitaminique et minéral sur la santé, a mobilisé une grande partie des volontaires. On doit rapprocher cet engouement du discours nutritionnel, diffusé ces dernières années, associant le bien-être à une alimentation équilibrée (Hébel, 1993). On notera que les femmes, plus sensibles que les hommes aux arguments diététiques (Lahlou, 1989) sont plus représentées dans l'échantillon étudié (63,7 %) que dans la population française (52,1 %, INSEE, 1992).

3. TYPOLOGIE DES BÉNÉVOLES

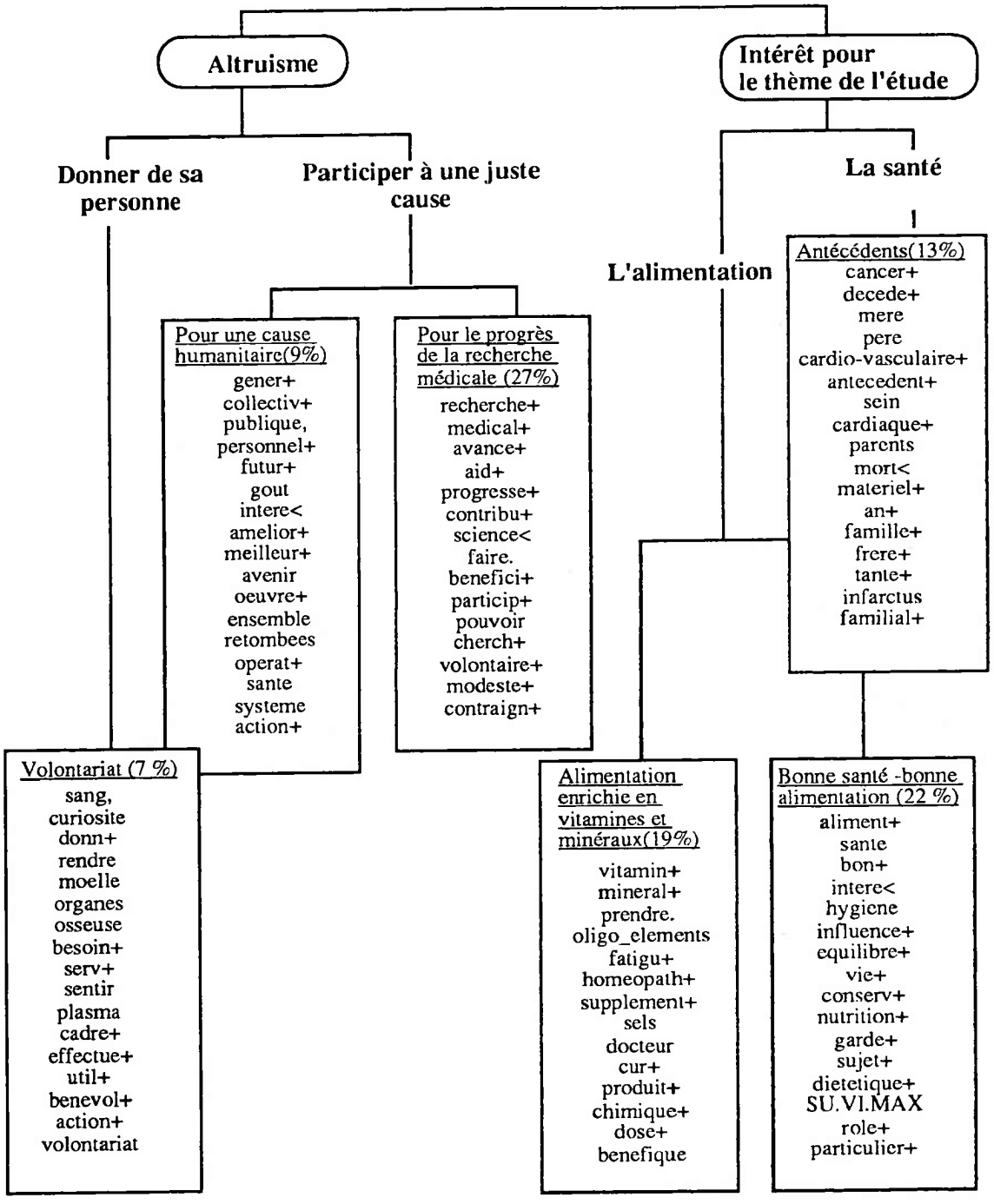
L'analyse statistique des réponses nous a conduit à retenir six classes. Le discours des volontaires se sépare en deux pôles :

- celui des "**intéressés par le thème de l'étude pour eux mêmes**". On retrouve dans cette catégorie des individus qui ont peur d'être malades et qui veulent donc être suivis, et d'autres qui trouvent l'étude intéressante en soi, soit pour le thème des enrichissements en vitamines et en oligo-éléments, soit pour le thème général d'une bonne alimentation pour vérifier le dicton :

  "On creuse sa tombe avec ses dents".

- celui des "**altruistes déclarés**" qui affichent leur volonté d'aider. D'une manière générale, les raisons invoquées sont de trois types :
  - par habitude de participer à des actions médicales (donneurs de sang, de plasma, ...) ;
  - par bienfaisance ;
  - pour aider la recherche médicale.

Graphique n°2 : Arbre de classification illustré par les mots les plus caractéristiques[1]



1 3 % des volontaires n'ont pas été classés, leurs réponses étaient ou trop courtes ou trop éloignées des réponses de la population globale.

3.1. INTÉRÊT POUR LE THÈME DE L'ÉTUDE

Plus de la moitié des individus se sont déclarés volontaires par intérêt personnel. L'homme cherche à satisfaire avant tout ses besoins physiologiques (Maslow, 1943). Il n'est pas étonnant alors de constater que 41 % des individus adhèrent (au sens où l'engagement est caractérisé par l'adhérence à l'acte, Joule et Beauvois, 1987) au projet SU.VI.MAX parce qu'il répond à leurs préoccupations quotidiennes. L'inquiétude grandissante (Hatchuel, 1992) vis-à-vis des maladies graves est un deuxième stimulus garantissant la ruée des volontaires.

Ce sont principalement des femmes (71 %) qui se trouvent dans cette catégorie.

Ce pôle est celui de l'intérêt bien compris. Les volontaires participent à l'enquête parce qu'ils en espèrent un bénéfice personnel direct ou indirect. Les sujets se sentent personnellement concernés par les objectifs de l'enquête, soit parce qu'ils se préoccupent de leur alimentation et cherchent à connaître de meilleures règles diététiques (donc notamment à participer à une découverte), soit parce qu'ils pensent être "à risques". La première motivation se sépare en deux sous-classes, l'une centrée précisément sur les vitamines, minéraux et les antioxydants, l'autre sur la nutrition en général.

Tableau n°2 : Phrases caractéristiques des classes "Intérêt pour le thème de l'étude"

L'alimentation		La santé
"Alimentation enrichie en vitamines et minéraux" (19 %)	**"Bonne santé - bonne alimentation" (22%)**	**"Antécédents" (13%)**
* C'est une façon bien que très modeste de participer à notre mieux-être social. Me nourrissant très mal (stress, extrême fatigue, manque de temps), j'ai conscience par la carence que j'en ressens de la nécessité des vitamines et oligo-éléments. * La nourriture actuelle me semble insuffisamment riche en vitamines. L'absorption régulière de vitamines plus oligo-éléments devrait compenser les carences alimentaires et permettre d'être en meilleure santé. * L'influence des vitamines et oligo-éléments sur la santé m'intéresse par rapport à la prévention possible de certains cancers et par rapport à la vieillesse. * Je pense que la vie actuelle est stressante et que notre fatigue permanente vient d'un manque de vitamines. J'aime beaucoup l'homéopathie et les médecines naturelles. * Menant une vie assez active, étant parfois contrainte de subir une certaine alimentation pas toujours équilibrée ou manger sur le pouce, je pense qu'un apport en vitamines et oligo-éléments devrait être bénéfique. Et bien sûr aussi une certaine curiosité car mon premier contact avec la vie active a été dans le domaine des caroténoïdes.	* Par curiosité et pour vérifier le dicton que l'on creuse sa tombe avec ses dents. * De par ma formation paramédicale et travaillant dans un service de Médecine préventive en université, tout ce qui concerne la nutrition et ce qui s'y rapporte est pour moi d'un très grand intérêt, persuadée que le maintien de la bonne santé dépend de l'équilibre alimentaire entre autres. * En tant que professeur de biologie, j'ai souvent axé mon cours sur ce problème, allant jusqu'à le mettre en application pratique, surtout sur l'importance du petit déjeuner chez les enfants soumis au ramassage scolaire, en pays de montagne, d'autant plus que j'ai subi un déséquilibre pendant 39-45, au front de guerre. J'ai tenté de faire comprendre que bien des maladies pourraient être évitées si la mère de famille pouvait avoir le temps nécessaire à consacrer à la fois à la nourriture matérielle et intellectuelle (mère de famille étant reconnue comme une travailleuse pouvant cotiser pour la S.S. et la retraite). Par ailleurs, nous avons la chance de pratiquer une alimentation biologique en raison des nombreux agriculteurs et de la proximité de l'usine Celnat.	* Fille et petite fille de femmes ayant eu des maladies dont les origines sont mal connues, je m'interroge (cancer et maladie dégénérative des cellules nerveuses : maladie d'Alzheimer ou autre ?) Pour recevoir, il faut soi-même donner - éthique chrétienne. * J'ai aussi des antécédents lourds : mon père et ma mère sont décédés du cancer, aussi je suis très intéressée par tous les aspects de prévention de cette maladie. J'ai envie de soutenir et de participer à la recherche contre les maladies de façon directe et interactive. J'ai une très bonne image et un investissement de la recherche. * J'ai toujours été intéressé sur le plan de la santé par les moyens permettant d'obtenir un état physique satisfaisant pour vivre assez harmonieusement et surtout prévenir les maladies graves. Mes parents sont décédés relativement jeunes (mère 54 ans cancer, père 64 ans artério-sclérose avec amputation d'une jambe). L'attirance pour les vitamines et minéraux a été fréquente chez moi, mais ne sachant pas dans quelles proportions et comment établir un mode d'emploi efficace, je ne l'ai pas mis en pratique.

3.1.1. Intérêts pour l'alimentation

Ce pôle de l'alimentation est celui de l'intérêt bien compris. Les volontaires participent à l'enquête parce qu'ils en espèrent un bénéfice personnel direct ou indirect. Les sujets se sentent personnellement concernés par les objectifs de l'enquête soit parce qu'ils se préoccupent de leur alimentation et cherchent à connaître de meilleures règles diététiques (donc notamment à participer à une découverte) soit parce qu'ils pensent être "à risques". La première motivation se sépare en deux classes, l'une centrée précisément sur les minéraux et les antioxydants, l'autre sur la nutrition en général.

Dix neuf pour-cent des individus (dont 71 % de femmes) sont motivés par le thème de l'**"alimentation enrichie en vitamines et minéraux"**. Cette sensibilisation aux aspects nutritionnels et plus particulièrement à la supplémentation en vitamines et minéraux traduit une préoccupation persistante du maintient de la forme passant par une alimentation raisonnée. Si le phénomène de l'allégé est passé de mode, la recherche d'une alimentation équilibrée et de conseils nutritionnels est toujours présente dans l'esprit des consommateurs (Lahlou, 1990). On ne sera pas étonné de constater en consultant le tableau n°3 que les individus de cette classe sont surtout des femmes consommant régulièrement des vitamines et surveillant de près leur état de santé et leur alimentation. Leurs motivations sont donc cohérentes avec un comportement de "prise en main" raisonnée de leur alimentation.

Avec un point de vue plus général, les personnes de la classe **"Bonne santé - bonne alimentation"** s'intéressent à la nutrition comme condition nécessaire d'accession à un état de santé optimal. Pour ces personnes, la durée de vie a pu en grande partie se prolonger grâce aux progrès réalisés par la science nutritionnelle et il faut continuer dans cette voie là. Cette motivation correspond donc à un désir de recherche plus général, qui se décline éventuellement dans des aspects pédagogiques.

Les individus de ces deux classes prennent régulièrement des vitamines ou des minéraux (74 %) et croient aux avantages d'une supplémentation de l'alimentation et aux valeurs d'une alimentation équilibrée. Les personnes citant les avantages des apports vitaminiques ont un état de santé moins bon que celles qui croient en l'alimentation atout santé. En effet, dans la classe "Alimentation enrichie en vitamines et minéraux", 46 % des individus sont sous

traitement médical (dans la population générale on n'en trouve que 39 %) et 28 % ont eu des maladies sérieuses (contre 24 % dans la population totale). Par contre, les individus de la classe "Bonne santé - bonne alimentation" se disent en bonne santé pour 97 % (dans la population totale, 95 % se classent aussi dans cette modalité) et leurs réponses aux questions sur le traitement médical et sur les maladies sérieuses ne sont pas significativement différentes de celles de la population totale (voir tableau n°3).

Tableau n°3 : Profils des individus des classes "Intérêt pour le thème de l'étude" [1]

L'alimentation		La santé
"Alimentation enrichie en vitamines et minéraux "	**"Bonne santé - bonne alimentation"**	**"Antécédents"**
Caractéristiques socio-démographiques		
Région : Méditerranée, Est Femme Pas de Minitel à domicile CSP : Non Actives Nombre d'enfants : deux Age : 55 ans et plus Poids : Entre 55 et 65 kilos A la Sécurité Sociale	Femme Professions intermédiaires Poids : Moins de 56 kilos Age : De 35 à 44 ans	Femme Employés Deux enfants Age : De 35 à 44 ans Région : Ouest A la Sécurité Sociale
Caractéristiques médicales		
Mammographie (< 3 ans) Sous traitement médical Tension Artérielle (d) : faible Consultation médecin : >5 f /an Cholestérol : Entre 2,2 et 2,8 g/l Tension Artérielle (d) : forte Maladies sérieuses : oui Tension Artérielle (s) : faible	Cholestérol : Entre 1,8 et 2,2g/l Non fumeur Consultation médecin : 2-3 /an En bonne santé Pas de médecin traitant Consultation médecin : 4-5 /an Télévision : non réponse Contraception : Pilule, stérilet Cholestérol : entre 2,8 et 5	Mammographie (< 3 ans) Contraception : Pilule Pas de cholestérol Taille : entre 1,6 et 1,65 m Non fumeur Pas sous traitement médical Taille : Moins de 1,60 m
Caractéristiques alimentaires		
Minéraux vitamines: qq. j/an Surveillance alimentation Suit un régime	Surveillance alimentation Minéraux vitamines: qq. j/an Suivez vous un régime : Non	Pas de surveillance alimentaire Quel régime : Pas précisé
Caractéristiques de diffusion		
Participer à SUVIMAX ?Peut-être Pas de Minitel à domicile	Participer à SUVIMAX ? Oui Appel entendu à R.T.L. et à France Info	Participer à SUVIMAX? Peut être Pas de Minitel à domicile Appel entendu à France 2 et France Inter

[1] Les modalités des variables sociologiques et médicales les plus représentatives des classes sont citées par ordre décroissant d'importance.

3.1.2. Préoccupations de santé

La deuxième préoccupation qui apparaît est celle des maladies graves telles que le cancer et les maladies cardio-vasculaires. L'objectif même du projet SU.VI.MAX, qui est d'évaluer l'influence des apports en vitamines et en minéraux sur les maladies, interpelle la peur collective vis-à-vis du cancer principalement. L'appréhension de la maladie est argumenté par l'antécédence médicale. L'inquiétude s'exprime surtout dans une population d'employés et la préoccupation alimentaire dans une population dominée par les professions intermédiaires.
Ce sont, curieusement, les plus jeunes des personnes interrogées qui expriment cette inquiétude.
Les modalités qui apparaissent dans le tableau n°3 sont issues d'un test statistique. Elles ne sont donc pas exclusivement présentes dans les classes correspondantes, elles y sont seulement surreprésentées. Il faut donc faire attention à la caricature qui peut être déduite trop hâtivement de ces résultats.

3.2. L'ALTRUISME DÉCLARÉ

Si l'engagement des individus du groupe précédent s'explique par l'espoir de retombées de leur investissement au travers d'un suivi médical, le reste des bénévoles(43 %) se déclare beaucoup plus altruiste.

Tableau n°4 : Phrases caractéristiques des classes "Altruisme"

Donner de sa personne	Participer à une juste cause	
"Volontariat" (7 %)	"Pour une cause humanitaire" (9 %)	"Pour le progrès de la recherche médicale" (27 %)
* Je suis donneur de sang bénévole. Je suis inscrit au fichier des donneurs de moelle osseuse. Je suis en bonne santé. Je ne suis pas idiot puisque je suis douanier. Alors pourquoi pas : j'ai l'avenir devant moi. * Importance de l'objet de la recherche et de ses applications futures. Curiosité (être sujet d'une recherche). Habitué au volontariat (vie associative, don du sang). * Membre actif d'une association locale des donneurs de sang bénévoles, j'aime me sentir utile pour une bonne cause. * J'ai donné plus de 130 fois du sang ou du plasma. C'est un état d'esprit de servir. * Très attirée par le volontariat et le besoin d'être utile aux autres et peut-être soulager des besoins de science pour tous. * Comme tout le monde, je considère la santé comme le bien le plus précieux. Toute action pour la préserver m'intéresse. Je serais très heureux, si ma candidature est retenue, d'y contribuer. A toutes fins utiles, je donne sang et plasma depuis de nombreuses années.	* Pour l'importance dans la recherche préventive qu'elle représente. Pour que les générations présentes et futures soient protégées de ces maladies. * Participer à cette étude est intéressant pour différentes raisons. Apporter sa petite contribution à la recherche. Aider les générations futures (mes enfants) à se soigner. Curiosité personnelle et désir de vivre une expérience utile. * Sur un plan personnel, il me semble intéressant d'aller chercher des améliorations par des voies non médicamenteuses. Par ailleurs, participer à une expérience de cette ampleur me motive particulièrement, car c'est une action intelligente, qui peut, je l'espère, faire avancer les connaissances. *Parce que nous sommes dans une société d'abondance et qu'il y a souvent une contradiction entre le niveau de nos connaissances et la manière dont on les met en oeuvre. C'est valable pour notre système d'éthique, notre mode de vie, nos capacités à faire des choix.	* Je souhaite faire quelque chose pour la recherche, mes moyens financiers ne me permettent pas de participer autant que je voudrais donc je suis très heureuse de pouvoir participer de cette manière. Je suis très motivée et très intéressée. * Je souhaite participer à cette étude, car les progrès de la recherche médicale passent par la participation financière et active de l'état et de la population et par la sensibilisation de cette dernière aux enjeux en cause. * Cela m'intéresse, toute recherche médicale m'intéresse. J'ai eu moi-même des jumeaux au bout de 14 ans de mariage grâce à des recherches considérables pour aider les femmes stériles. Je pense que si je peut être utile, c'est mon devoir.

3.2.1. Volontariat

A travers l'enquête SU.VI.MAX, les désirs de dévouement à autrui s'expriment pour 7 % des individus par une adhérence physique. L'individu fait "prêt" de son corps à la science et répond ainsi à son besoin d'estime, comme l'exprime la réponse :

"J'aime me sentir utile pour une bonne cause"

Pour les personnes de cette classe, l'aide personnelle à la recherche médicale est analogue à une action humanitaire, elle passe par le dévouement physique. Il faut souligner qu'un grand nombre d'individus de cette classe font habituellement des dons de sang. La participation à SU.VI.MAX est donc en continuité avec une pratique habituelle d'aide bénévole pour la médecine.

L'acquis culturel est indéniable dans le cas de l'espèce humaine, où l'altruisme réciproque apparaît comme une évidence. Le raisonnement de l'homme se substitue peut être à un mécanisme biologique existant chez l'animal, mais le résultat est le même. Il semble bien que ce soit un tel mécanisme positivement adaptatif pour l'espèce que nous voyons ici à l'oeuvre.

L'individu typique de cette classe a une vie professionnelle indépendante (agriculteur (44,4 % des agriculteurs sont dans cette classe), artisan, commerçant, (14,3% sont dans cette classe)), il ne doit rien à personne et a sans doute besoin d'être reconnu par les autres.
D'où vient, en nos temps individualistes, cette position altruiste ? L'origine biologique de l'altruisme a été mise en évidence lors de l'observation d'animaux. S'il est souvent observé en direction du groupe familial parce qu'il permet de rendre plus efficace la survie des gènes, il existe aussi un altruisme entre individus non apparentés à condition qu'il soit payé de retour, c'est ce que Jaisson (1993) appelle "l'altruisme réciproque". L'avantage d'un tel comportement est l'efficacité en terme de survivance génétique comme l'explique Eibl-Eibesfeldt (1984, p. 391).

Tableau n°5 : Profils des individus des classes "Altruisme"

Donner de sa personne	Participer à une juste cause	
"Volontariat"	"Pour une cause humanitaire"	"Pour le progrès de la recherche médicale"
Caractéristiques socio-démographiques		
Homme Agriculteur exploitant Pas d'enfant Artisans, Comm., Chefs d'entreprise Région : Est Pas à la Sécurité Sociale	Homme Nombre d enfants aucun Région : Est Artisans, Comm., Chefs d'entreprise A la Sécurité Sociale	Homme Région : Bassin parisien Régime de S.S. : général Ouvriers Région parisienne Pas d'enfant Age : De 45 à 49 ans
Caractéristiques médicales		
Pas de cholestérol Tension Artérielle (D) : normale Fumeur depuis 20 à 30 ans Pas de mammographie Poids : Entre 65 et 75 kilos Pas de cholestérol Pas sous traitement médical Taille : entre 1,65 et 1,72m Tension Artérielle (S) : normale Poids : Plus de 75 kilos Médecin traitant : non Consultation médecin <2 f / an Nbre de cigarettes /j : <10	Pas de maladie sérieuse Poids : Entre 65 et 75 kilos Pas de mammographie Taille : entre 1,65 et 1,72m Cholestérol : 2,2 et 2,8 g/l Minéraux-vitamines : 1 f /j Nbre années fumeur : 30-40 ans Taille : plus d 1,73 m Fumeur Consultation médecin <2 f / an Tension Artérielle (D) : faible Nbre années fumeur <10 ans Tension Artérielle (S) : faible	Poids : plus de 75 kilos Entre 10 et 20 cigarettes / jour Consultation médecin : 4-5 f/an Bilan santé : oui Régime S.S. : général Nbre années fumeur : 10-20 ans Taille : plus d 1,73 m Pas de mammographie Tension Artérielle (D) : normal Tension Artérielle (S) : fort
Caractéristiques alimentaires		
Pas de régime alimentaire Pas surveillance alimentaire	Pas de régime alimentaire	Minéraux - vitamines : N.S.P.
Caractéristiques de diffusion		
Appel entendu sur TF1 Appel entendu à RMC	Appel entendu sur France Inter Minitel à domicile Participer SUVIMAX : peut-être	Presse Non réponse Appel entendu sur France 3 Utilise exceptionnellement Minitel Appel entendu sur R.T.L.

3.2.2. Participer à une juste cause

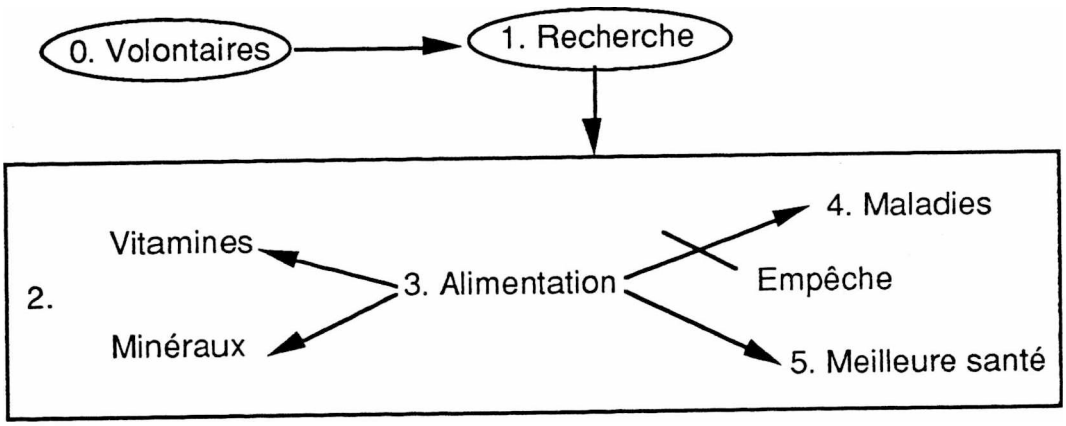
Comme on peut le voir sur le graphique n°2, une classe représentant 27 % des individus se détache de la classe précédente par son contenu plus général quant à son engagement altruiste. Les individus de cette classe s'engagent dans le projet SU.VI.MAX pour le progrès de cette recherche. Il s'agit surtout d'hommes de la région parisienne surveillant leur état de santé. Ce sont plutôt des ouvriers. Pour 9 % des personnes sélectionnées, les raisons de l'engagement sont plus générales. Il s'agit de défendre l'intérêt public et de se dévouer pour les générations futures. Ces bénévoles surveillent peu leur état de santé.
On a donc affaire ici à des engagements pour des raisons idéologiques, dont la tonalité est humanitaire.

Pour 36 % des individus, la justification de leur engagement est donc leur sensation d'accomplir une action humanitaire. Ce regain d'intérêt pour les autres au travers d'organisations caritatives a été observé par Lahlou, Collerie de Borely et Beaudouin (1993) et Collerie de Borely (1994).

4. CHEMINEMENT DES MOTIVATIONS

Schématiquement, les objectifs de l'enquête SU.VI.MAX se résument suivant le graphique n°3. Les volontaires insistent finalement sur la partie qui les intéresse le plus et se positionnent en accordant plus ou moins d'importance à l'un ou plusieurs des points. Ainsi, les individus de la classe "Alimentation enrichie en vitamines et minéraux" se sont focalisés sur les points 2 et 3 du graphique n°3, les individus de la classe "Bonne santé - bonne alimentation" sur les points 3 et 5 ; ceux de la classe "Antécédents" sur le point 4 ; ceux de la classe "Bénévolat" sur le point 0 ; ceux de la classe "Intérêt public" sur les points 0 puis 4 et 5, et enfin ceux de la classe "Recherche médicale" sur les points 1, 4 et 5.

Graphique n°3 : Schéma de l'enquête SU.VI.MAX



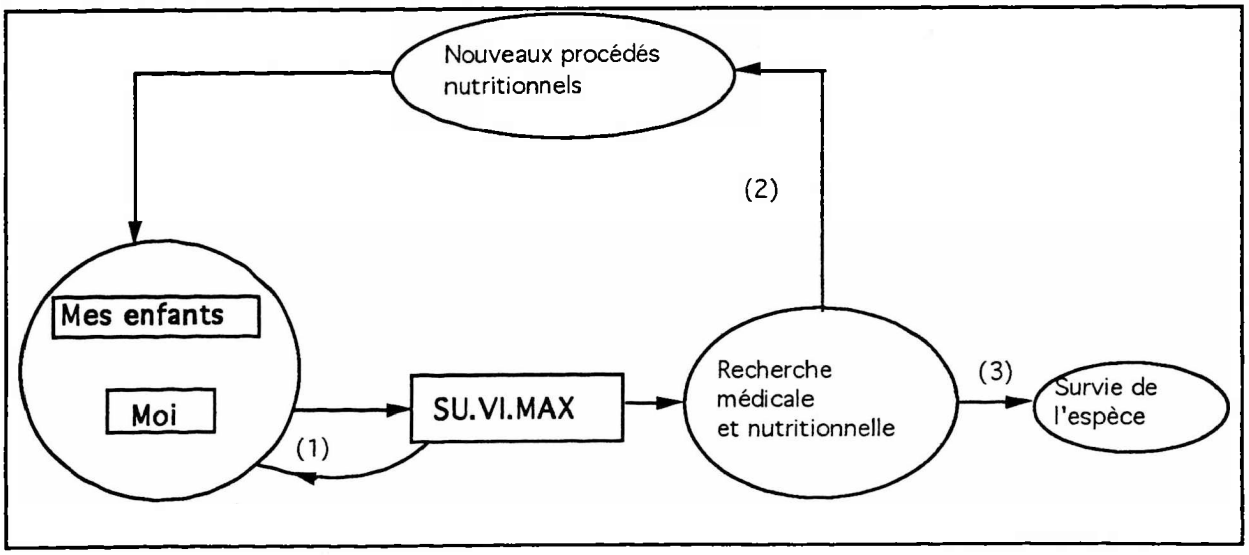
Cette focalisation s'explique par l'histoire des individus et leur situation personnelle : ils orientent la représentation sur la partie finale qui peut leur servir d'objectif (ce sont des représentations privilégiées comme les définit Lahlou (1994)). Ainsi, il est clair que les femmes plus préoccupées par la diététique et les individus atteints de maladies à risques s'intéressent au but "meilleure santé par l'alimentation" du dispositif SU.VI.MAX et se retrouvent dans le pôle "Besoins alimentaires". L'adhésion aux autres points du dispositif est moins évidente et une analyse plus approfondie des motivations serait nécessaire. La variété

des justifications révèle en fait une diversité des perceptions du projet : chacun voit midi à sa porte. La séparation entre les classes n'est pas toujours très claire. La réponse :

> "J'ai toujours été très intéressée par la diététique : j'aime bien manger, j'aime faire la cuisine et j'aime que les bonnes choses soient aussi saines. Bien vivre et bonne santé. Je suis intéressée par la valeur nutritive des aliments et j'aimerais savoir mieux gérer la façon de se nourrir. J'ai aussi des antécédents lourds : mon père et ma mère sont décédés du cancer, aussi je suis très intéressée par tous les aspects de prévention de cette maladie. J'ai envie de soutenir et de participer à la recherche contre les maladies de façon directe et interactive. J'ai une très bonne image de l'investissement de la recherche."

traduit un élargissement des préoccupations qui va de soi aux autres et est à cheval sur plusieurs classes. La personne interrogée part d'une explication égocentrique concentrée sur le besoin de base : l'alimentation, puis les objectifs de l'étude lui reviennent à l'esprit et elle pense à elle en tant que victime possible de la maladie. Enfin, si l'étude peut permettre de faire avancer la recherche et défendre une cause humanitaire, cela devient une argument supplémentaire. Ce phénomène a déjà été observé lors du traitement lexical d'autres questions ouvertes générales (Lahlou, Collerie de Borely et Beaudouin, 1993).

Graphique n°4 : Système de fonctionnement des motivations de bénévoles



Lorsque l'individu ne s'est pas libéré de ses contraintes primaires qui sont celles de ses anxiétés vis-à-vis de son état de santé, il peut trouver en feed-back de son adhésion au projet SU.VI.MAX un contrôle médical et une alimentation supplémentée et surveillée. S'il a plus d'aisance et peut s'intéresser à des objectifs plus vastes et élevés comme le permet la générosité ("qualité qui élève l'homme au-dessus de lui-même et le dispose à sacrifier son intérêt personnel, son avantage à celui des autres, à se dévouer pour son prochain", Robert Électronique, 1991), il s'investira pour les générations futures et donc pour la survie de la société. On retrouve ici la topique des motivations décrite par Lahlou, Collerie de Borely et Beaudouin (1993) qui situe les désirs des consommateurs centrés en priorité sur la survie du corps, puis sur la sauvegarde de la situation familiale et enfin de la société.

Un objet unique, SU.VI.MAX, fait donc réagir différemment les individus. Tous attendent une gratification, mais à des niveaux de préoccupation plus ou moins élevés : utilitaire (être suivi), stratégique (bien s'alimenter, se soigner), altruiste (améliorer le sort de l'espèce). Cette dernière zone correspond aux individus qui ont maîtrisé la plupart des contraintes externes et qui cherchent à se réaliser en prenant en compte des considérations existentielles. Les individus entrant dans cette sphère seraient dans ce modèle d'analyse plus représentés parmi les professions supérieures. Or ici, les catégories socioprofessionnelles les plus caractéristiques de cette tendance sont celles des artisans, commerçants, chefs d'entreprises, des agriculteurs et des ouvriers. La topique des motivations "hiérarchiques" de Maslow (1943) ne semble pas bien fonctionner ici. La comparaison des mots les plus fréquents dans ces différentes classes dans le tableau n°6 en est une preuve flagrante.

Tableau n°6 : Mots caractéristiques de certaines catégories socioprofessionnelles

Agriculteurs 8 %	Artisans commerçants, chefs d'entreprise 7 %	Ouvriers 5 %	Cadres supérieurs 8 %	Professions intermédiaires 20 %
curiosité	**action+**	**désir+**	particulier+	rôle+
travail+	simple+	**science<**	type+	oligo-éléments
aid+	enquête+	surveill+	meilleur+	persuade+
gener+	occasion	**collectiv+**	consomm+	persuade+
seul<	niveau+	**avance+**	sérieus+	vitamin+
informe+	résult+	nutrition+	importance+	santé
don+	vasculaire+	possible+	motiv+	aliment+
mange+	partie<	nécessaire+	développe+	famille+
serv+	**donn+**	form+	expérience+	intere<
mal.	**vie+**	aim+	sein	concern+
savoir	âge+	nutritionnel+	modeste+	étud+
physique+	risqu+	père		personnel+
bienfait+	intére<	activ+		prévent+
complément+	cardio	aid+		problèm+
surveill+	**curiosité**	équilibre+		risqu+
essai+	nutritionnel+	enfant+		important+
minéral+	science<	contribu+		convaincu+
âge+	prévent+	santé		

On peut supposer, qu'à l'heure actuelle, la maturité des individus est forte et a atteint la plupart des couches sociales. Les professions artisanales, commerçantes et de chefs d'entreprises (incluant les exploitants agricoles) ont, sans doute, besoin de plus de reconnaissance que les professions moins indépendantes et souvent mieux reconnues comme le sont les professions intermédiaires et celles des cadres supérieurs. Participer à la collectivité est le besoin ressenti par les ouvriers qui ont du mal à s'affirmer aussi dans leur vie professionnelle. Leur besoin principal est celui de l'estime de soi. Les professions intermédiaires, correspondant souvent à des personnes du secteur médical, sont très représentées mais ont un discours beaucoup plus pragmatique. Les autres catégories ont un discours beaucoup moins marqué et sont réparties également dans les différentes classes de discours.

Rappelons que tous les individus sont de la même tranche d'âge et qu'il n'y a donc pas d'effet de génération.

CONCLUSION

L'enthousiasme général des Français pour le projet SU.VI.MAX s'explique d'abord par l'intérêt pour le thème de l'étude : la corrélation entre les maladies et l'alimentation. Ces deux thématiques font partie des préoccupations actuelles.

Le résultat le plus intéressant est sans doute de constater que 43 % des bénévoles agissent par philanthropie. Ce résultat est d'autant plus étonnant qu'il est plus tentant pour l'homme d'agir pour lui que pour les autres comme l'exprime Rousseau (1862) :

"Ma foi, pas si bête ! chacun pour soi dans ce désert d'égoïsme qu'on appelle la vie"

Ce courant de pensée entre dans la zone de motivation "exploratoire" comme la définissent Lahlou, Collerie de Borely et Beaudouin (1993). Pour Comte-Sponville (1992), la générosité apparaît comme besoin d'amour puis comme une exigence, d'un point de vue moral, comme un devoir.

L'analyse des réponses à la question sur les motivations des bénévoles de l'enquête SU.VI.MAX apporte des éléments importants quant aux tendances actuelles du comportement des individus.

- Les Français croient aujourd'hui aux vertus de l'alimentation comme atout santé.
- La peur des maladies graves dont le cancer reste toujours présente dans leur esprit.
- Après la montée de l'individualisme dans les années 1980, les années 1990 voient un retour à l'humanisme et à une solidarité raisonnée.

BIBLIOGRAPHIE

CONSEIL NATIONAL DE LA VIE ASSOCIATIVE, (1993).- *Bilan de la Vie Associative en 1990-1991*. CNVA.

BEAUDOUIN V., LAHLOU S., (1993).- *Analyse lexicale. Outil d'exploration des représentations. Réflexions illustrées par une quinzaine d'analyse de corpus d'origines très diverses.* CRÉDOC, Cahier de Recherche n°48, Paris.

COLLERIE DE BORELY A., (1994).- *Prix, qualité, service : Les arbitrages du consommateur.* CRÉDOC, Cahier de Recherche, N°58, Paris.

COMTE-SPONVILLE A., (1992).- *Fondation de France*, N° 70, juillet.

EIBL-EIBESFELDT I., (1984).- *Biologie du comportement.* Diffusion Ophrys. Naturalia et biologia.

HATCHUEL G., (1992).- *Les grands courants d'opinions et de perceptions en France de la fin des années 1970 au début des années 1990.* CRÉDOC, Collection des rapports. N°116, Paris.

HÉBEL P. ET RACAUD T., (1993).- "Analyse des outils de communication nutritionnelle", *L'analyse lexicale : outil d'exploration des représentations - Résultats illustratifs.* CRÉDOC, Cahier de Recherche n°48BIS, Paris.

HERCBERG, S ; HÉBEL, P ; PREZIOSI, P ; BRIANÇON S. ; FAVIER, A. ; GALAN, P. ; MALVY, D. ; ROUSSEL A.M. ; SCHWARTZ L., (1994).- "Les motivations des volontaires répondant à un appel particulier pour participer à une étude d'intervention dans le domaine de la prévention nutritionnelle : résultats d'un pré-test du projet SU.VI.MAX". A paraître dans la *Revue d'Épidémiologie et de Santé Publique.*

JAISSON P., (1993). *La fourmi et le sociobiologiste.* Éditions Odile Jacob.

JOULE R.V. ET BEAUVOIS J.L., (1987).- *Petit traité de manipulation à l'usage des honnêtes gens*. Vies sociales. Presse Universitaire de Grenoble.

LAHLOU S., (1989).- *Le Comportement alimentaire des Français*. Rapport au programme Aliment 2000. CRÉDOC.

LAHLOU S., (1989).- *Si / alors "Bien manger". Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus constitué des associations libres de 2000 individus*. CRÉDOC, Cahier de Recherche N°34, Paris.

LAHLOU S.,(1994).- *Penser manger*. Thèse. EHESS. A paraître.

LAHLOU S., COLLERIE DE BORELY, A. BEAUDOUIN V., (1993).- *Où en est la consommation* aujourd'hui. ?. CRÉDOC, Cahier de Recherche, N°46, Paris.

NISAK C., (1992).- Fondation de France, Numéro 70, juillet.

MASLOW A.H., (1943).- "A Theory of Human Motivation", *Psychological Review*, Vol 50.

REINERT M., (1983).- "Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte". *Les cahiers de l'analyse des données*. Vol VIII, n°2.

ROUSSEAU J.J., (1862).- *Émile* (IV, Note).

ROBERT ÉLECTRONIQUE, (1991).- Version Macintosh.

SOMMAIRE DES FIGURES

Graphique n°1 : Distribution comparée de la population des volontaires par rapport à celle de la population générale en 1990 en fonction des CSP 82
Tableau n°1 : Fréquence des racines 84
Graphique n°2 : Arbre de classification illustré par les mots les plus caractéristiques 87
Tableau n°2 : Phrases caractéristiques des classes "Intérêt pour le thème de l'étude" 89
Tableau n°3 : Profils des individus des classes "Intérêt pour le thème de l'étude" 91
Tableau n°4 : Phrases caractéristiques des classes "Altruisme" 93
Tableau n°5 : Profils des individus des classes "Altruisme" 95
Graphique n°3 : Schéma de l'enquête SU.VI.MAX 97
Graphique n°4 : Système de fonctionnement des motivations de bénévoles 98
Tableau n°6 : Mots caractéristiques de certaines catégories socioprofessionnelles 100

CAHIER DE RECHERCHE

Récemment parus :

Parcours singuliers : repérer et interpréter les trajectoires atypiques, par Denise Bauer, n°54, octobre 1993.

La modernisation dans les services publics à caractère social, par Marie-France Raflin, n°55, novembre 1993.

Les exclus du mythe américain : l'heure des comptes, par Isabelle Groc, n°56, mars 1994.

Niveau de vie et revenu minimum : une opérationalisation du concept de Sen sur données françaises, par Christine Le Clainche, n°57, avril 1994.

Prix, qualité, service : les arbitrages du consommateur, par Aude Collerie de Borely, n°58, avril 1994.

Approche sectorielle de l'évolution de l'emploi dans l'industrie manufacturière (1988-1992), par Philippe Moati, n°59, mai 1994.

Articles d'études et de recherche : Année 1993, par Michel Messu, Philippe Moati et Robert Rochefort, n°60, mai 1994.

Président : Bernard SCHAEFER Directeur : Robert ROCHEFORT
142, rue du Chevaleret, 75013 PARIS - Tél. : (1) 40.77.85.00

ISBN : 2-84104-008-9

CRÉDOC

Centre de recherche pour l'Étude et l'Observation des Conditions de Vie