5e journées internationales d'analyse statistique
des données textuelles
(Lausanne, jeudi 9 mars)

Analyse des données textuelles et Statistique lexicale (Textual Data Analysis and Lexical Statistics)

Dominique Labbé
CERAT-IEP - BP 48 - F38.402 Grenoble Cedex
cerat@iep.upmf-grenoble.fr

Résumé

Cette conférence plaide pour des données textuelles de qualité, normalisées et étiquetées. Elle illustre leur utilité à l'aide d'un exemple : le sens du mot "amour" dans l'oeuvre de Corneille. La technique de l'étiquetage est présentée. Enfin, on évoque la nécessaire coopération entre les chercheurs pour la réalisation des outils de normalisation et d'étiquetage et pour la constitution de corpus de référence.

Summary

This presentation argues in favor of high quality normalized and tagged textual data. An example is given : the sense of the word "love" in Corneille's plays. Then it explains the main principles for normalization and tagging. At least, large cooperation between researchers is needed to elaborate norms and tagging tolls and to create large tagged corpora.

Key Words : textual data, normalization, tagging, corpora

Dans les conférences qui ouvrent un congrès comme le nôtre, il est de bon ton de brosser de vastes fresques portant sur l'état présent de la discipline et sur les développements, nécessairement enthousiasmants, qui se profilent à l'horizon. Permettez-moi de déroger un instant à cette règle et d'abaisser le regard sur nos outils et nos matériaux.

En ce qui concerne les outils, précisons d'emblée que la conjonction de coordination qui, dans le titre de cette conférence, relie l'analyse des données et la statistique est de nature booléenne. Il ne s'agit pas de donner à choisir entre la statistique de papa et les méthodes multidimensionnelles modernes mais de voir comment combiner les deux pour faire des choses intéressantes...

Employer le terme "données textuelles" signifie qu'on travaille avec du texte (et non sur la langue par exemple). Si l'on prend l'adjectif "textuel" au sérieux, cela signifie aussi qu'il faut respecter le texte et éviter de lui faire subir des outrages irréparables. Ainsi ne doit-on pas le réécrire entièrement en lettres capitales ou, à l'inverse, réduire toutes les majuscules ou encore : enlever les accents, etc.

Au fond, nos données doivent être de qualité. Disons qu'un bon éditeur ne devrait pas en rougir : pas de coquilles, de fautes d'orthographe, une bonne ponctuation, etc. En plus de cette mise au net, une seule intervention s'impose : ajouter les références précises du texte, éventuellement les numéros de page, de chapitre, etc. Ainsi, pourra-t-on échanger les corpus et comparer les performances des logiciels en utilisant les mêmes étalons.

Mais allons un peu plus loin. Si nous analysons les textes, c'est pour y chercher le sens.

1. A la recherche du sens

Travailler sur le texte revient à postuler que le sens des mots n'est pas à rechercher d'abord dans le composant sémantique de la langue mais dans les usages, toujours singuliers, que fait de ces mots tel ou tel locuteur.

Comment faire pour retrouver ce sens singulier ?

Permettez-moi de prendre un "petit" exemple : les 34 pièces de Corneille, soit 553.191 mots. "Amour" est le substantif le plus fréquemment employé dans ce corpus. Voilà qui ne surprend pas vraiment. Mais quel sens Corneille donne-t-il à ce mot ? Si nous sommes fidèles au principe énoncé ci-dessus, il faut interroger le texte et examiner les mots qui entourent chaque occurrence d’"amour".

L'examen des formes graphiques nous apprend que :

— devant "amour", on trouve par ordre de fréquence décroissante : "l'" (686 fois) puis "d'" (216 fois), "mon" (200), votre (110) et "cet" (82)…

— derrière "amour" : "de" (59 fois), "qui" (48), "est" (40), "pour", "a", etc.

Autrement dit, devant une voyelle Corneille élide l'article "le" et la préposition "de" et remplace "ce" par "cet". Ou encore : il accorde le verbe avec le sujet… Corneille écrit en français ! Ne sourions pas trop vite. Tous ceux qui ont été un jour confrontés aux listes de "segments répétés" le savent : le "bruit" provoqué par les règles de la graphie et de la syntaxe française est si fort qu'il couvre presque complètement le sens. Ainsi, autour d'un verbe conjugué à la première personne, on est certain de rencontrer un pronom "je", etc... Autrement dit, nous vérifions que l'auteur connaît sa grammaire mais rien ne permet d'affirmer que ce verbe est caractéristique de la première personne puisque le même va apparaître à la seconde personne avec "tu", à la troisième avec "il", etc.

Comment mettre de l'ordre dans cette émeute des formes graphiques ?

Par exemple, comment vérifier que "est" et "a" sont bien des verbes auxquels Corneille associe "amour" comme le laisse penser l'examen des formes graphiques ? Donnons à tous les verbes employés par l'auteur — indépendamment de leur mode, de leur temps et de leur personne — la même chance de se trouver associés à "amour" et voyons, dans l'entourage de ce mot, ceux qui sont "trop" employés et ceux qui ne le sont pas "assez" (par rapport à l'hypothèse nulle d'une répartition aléatoire des mots). Ces liaisons positives et négatives forment *l'univers sémantique du mot* (voir : Hubert-Labbé, 1995).

Le résultat de cette opération est présenté en annexe à ce document. Les mots significatifs sont classés par catégories grammaticales et, au sein de chaque catégorie, en fonction inverse de l'intensité du lien les unissant à "amour". Les listes sont limitées aux mots pour lesquels, on peut affirmer, avec moins de 1% de chances d'erreur, qu'il existe un sur-emploi ou un sous-emploi caractéristique.

Nous avons la réponse à la question posée ci-dessus. Par rapport à l'emploi moyen du verbe "être" dans l'ensemble de l'oeuvre, Corneille sous-emploie très significativement ce verbe quand il parle d'amour. Quant à "avoir", il est "banal" au sens de Lafon, c'est-à-dire qu'il n'apparaît pas en positif ni en négatif. L'article "le" et la préposition "de" se trouvent dans le même cas.

Certains résultats pourront paraître triviaux comme l'association de Vénus avec l'amour. Il y a des clichés inévitables ! D'autres résultats sont moins prévisibles. Par exemple : "haine" est le substantif associé le plus fortement à "amour". A la suite des dictionnaires, on classerait plutôt "haine" dans les antonymes d'"amour". Je n'ai pas dépouillé toute la littérature critique sur l'oeuvre de Corneille, mais, dans ce que j'en ai lu, je n'ai pas vu signalée cette association des deux mots. Quoi de plus logique au fond ? Parmi le demi-million de mots de Corneille, "amour" n'est employé "que" 1 888 fois, ce qui n'est déjà pas considérable, et haine 433. Comment repérer que les 92 occasions où ils sont associés sortent vraiment de l'ordinaire ? Seul l'ordinateur peut le faire. Ou encore, la forte association d'"amour" avec l'adjectif "conjugal"…

Et les mots les plus fortement sous-employés ? "seigneur" et "Dieu". Ce n'est pas trivial non plus !

Naturellement, il ne faut pas extraire un mot de la liste mais examiner le réseau sémantique entier. Par exemple, l'importance de la métaphore du "feu" est confirmée par la présence des verbes "éteindre", "allumer", "brûler", "rallumer" ou encore de l'adjectif "éteint" mais aussi de son contraire : la *froideur*. Autre exemple, l'emploi très important du possessif "mon" doit être rapproché du sous-emploi de "ton" ainsi que des pronoms "tu" et "vous" : chez Corneille, on se déclare rarement à la personne aimée, on fait confidence à un tiers à qui on livre son coeur...

Il faut également examiner les oppositions entre les deux listes. Par exemple, quels sont les "objets" d'amour chez Corneille ? La liste est brève : outre la "beauté", il s'agit de : *mère, jeunesse, patrie, amant* et *maîtresse*. Tout le reste est repoussé en dehors de cet univers comme s'il s'agissait d'obstacles : les puissants (au premier chef, les *Romains, Rome* et *César*) : *dieu, roi, prince, comte, chef, état, peuple, autel*… mais aussi et peut-être surtout : *père, frère, soeur, fils, gendre*…

Remarquons également que l'amour se rattache peu au corps (*coeur* étant une métaphore) : la *tête,* l'*oeil* ou la *main* sont en association négative.

Le temps manque pour évoquer toute la richesse du réseau sémantique suggéré par la liste en annexe. Naturellement, cette analyse statistique n'est qu'exploratoire, travailler sur des

données textuelles implique que l'on en revienne toujours au texte. En l'occurrence, il suffit de demander à la machine de rechercher les phrases, ou membres de phrases, contenant "amour" avec le plus d'associations positives (et le moins d'associations négatives). La phrase la plus caractéristique de l'amour chez Corneille évoquera certainement quelques souvenirs :

CHIMENE.

C'est peu de dire aimer, Elvire : je l'adore ;
Ma passion s'oppose à mon ressentiment ;
Dedans mon ennemi je trouve mon amant ;
Et je sens qu'en dépit de toute ma colère,
Rodrigue dans mon coeur combat encor mon père :
Il l'attaque, il le presse, il cède, il se défend,
Tantôt fort, tantôt foible, et tantôt triomphant ;
Mais en ce dur combat de colère et de flamme,
Il déchire mon coeur sans partager mon âme ;
Et quoi que mon amour ait sur moi de pouvoir,
Je ne consulte point pour suivre mon devoir :
Je cours sans balancer où mon honneur m'oblige.
Rodrigue m'est bien cher, son intérêt m'afflige ;
Mon coeur prend son parti ; mais malgré son effort,
Je sais ce que je suis, et que mon père est mort.

(Le Cid, Acte III, scène 3)

En annexe, on trouvera les phrases qui suivent celles-ci parmi les plus caractéristiques de l'univers sémantique de l'amour chez Corneille.

Cependant avec un mot aussi polysémique, plusieurs sens coexistent certainement. Comment les retrouver ? La combinaison de la classification traditionnelle et de l'analyse factorielle semble appropriée à cette recherche. Toutes les phrases contenant "amour" sont rassemblées en un vaste tableau et classées en fonction de leur proximité réciproque sur les plans factoriels (voir le chapitre II de Lebart et Al, 1995). Ainsi, on verra apparaître les principales acceptions d'"amour" chez Corneille, acceptions que l'on illustrera avec un réseau sémantique et des phrases clefs comme nous venons de le faire ci-dessus.

Un mot encore sur cette expérience, la méthode n'est pas parfaite. Après "haine", les substantifs les plus fortement liés à "amour" sont le "tour", le "jour", le "retour" et, un peu plus loin dans la liste, on trouve : la "cour" et le "séjour"… car ces mots offrent des rimes à "amour" quand celui-ci vient en fin de vers. Nos données ne sont donc pas exemptes d'un certain bruit ! Mais au moins avons-nous neutralisé le vacarme provoqué par les règles graphiques et syntaxiques du français.

Comment avons-nous procédé ?

2. Les données textuelles normalisées et étiquetées

Pour obtenir ces résultats il faut normaliser les graphies des mots et les étiqueter.

En premier lieu, la normalisation. Tout logiciel de traitement de données textuelles se trouve confronté à quelques questions que l'usager ignore parfois :

— "L'" et "l'" ou "D'" et "d'" : un ou deux mots différents ?

— "aujourd" et "hui" : deux mots différents ?

— "abattage" et "abatage" : deux mots différents ?

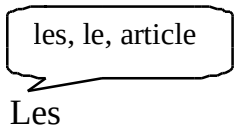
— dans "M. M. Proust", les deux "M" sont-ils équivalents et le point est-il un signe de ponctuation ?

Les auteurs des logiciels de traitement de données textuelles ont généralement choisi de répondre "oui" à ces quatre questions mais les utilisateurs en sont-il conscients ? Le réflexe serait plutôt de répondre "non", car on sait que l'article (ou le pronom) "le" placé en début de phrase (ou de vers) prend une majuscule initiale, mais n'en demeure pas moins un mot commun ; que "aujourd'hui" est un seul mot ; que plusieurs milliers de mots du français ont plusieurs orthographes, que *monsieur Proust* se prénommait *Marcel*, etc. C'est cette seconde voie qu'il faut donc adopter en "normalisant" les graphies.

Attention, la graphie normalisée ne se substitue pas au texte original. On se contente d'attacher aux mots des étiquettes sur lesquelles on écrit les "formes normalisées" comme ceci :

Evidemment, cet étiquetage doit être aussi automatisé que possible pour ne pas perdre son temps et pour éviter les erreurs et les biais. Il s'agit donc d'apprendre à un automate les conventions typographiques, les abréviations, les variantes graphiques, etc.

Pourquoi s'arrêter en chemin ? Le programme effectue l'opération suivante :



L'étiquette comporte d'abord la forme normalisée, puis l'entrée du dictionnaire que l'on appelle aussi "vocable". Naturellement, "les" peut aussi être un pronom. Il faut donc examiner l'entourage de la forme pour déterminer sa catégorie grammaticale… De la même manière, on attache aux verbes une étiquette avec leur infinitif, etc. Autrement dit, on rattache les mots à leur place dans le lexique français, bref on reconstitue le "vocabulaire" du texte, ce qui est plus satisfaisant qu'un simple "formulaire" (recueil des formes graphiques brutes)...

Naturellement, pour opérer cet étiquetage, il faut avoir recours au lexique de la langue. Ainsi s'explique le terme "statistique lexicale".

Quel est l'impact sur le corpus ?

Voici le résultat de cette expérience réalisée, il y a une quinzaine d'années, sur environ un demi-million de mots : les interventions radio-télévisées de Charles de Gaulle (1958-69) et celles du premier septennat de F. Mitterrand (1981-88). Le nombre des "formes brutes" est obtenu par application mécanique du système des caractères délimiteurs sur le texte publié, sans autre modification de notre part que les corrections orthographiques.

Tableau 1. Impact de l'étiquetage sur les effectifs totaux

	"Occurrences"	1. Formes brutes	2. Formes normalisées	3. Vocables	3/1
Ch. de Gaulle	201 907	15 910	12 601	6 473	0,41
F. Mitterrand	305 124	17 150	13 272	7 700	0,45

La dernière colonne indique que l'impact de l'étiquetage est massif. Un peu d'ordre règne dans l'émeute des formes graphiques : par rapport au traitement standard, le nombre des individus différents est réduit d'au moins 55%… On peut entrer un peu dans le détail. Le

tableau ci-dessous compare la distribution des fréquences entre les formes normalisées et les vocables.

Tableau 2. Impact de l'étiquetage sur la distribution des fréquences (corpus Mitterrand)

	1. Formes normalisées	%	2. Vocables	%	2/1
Hapax	5 476	41,3	2 624	34,1	48
De 2 à 4	4 381	33,0	2 400	31,2	55
De 5 à 10	1 256	9,5	821	10,7	65
De 11 à 20	861	6,5	628	8,2	73
De 21 à 30	383	2,9	315	4,1	82
De 31 à 50	343	2,5	317	4,1	92
De 51 à 100	260	2,0	270	3,5	104
Plus de 100	312	2,3	317	4,1	102
Total	13 272	100	7 692	100	58

La réduction la plus nette est obtenue sur les basses fréquences. C'est là un avantage important. En effet, le statisticien n'a rien à dire sur un fait unique, sinon qu'il est unique… En revanche, on voit augmenter le nombre des individus situés dans les classes de fréquence supérieures à 30 (notre gibier favori).

3. Les corpus étiquetés

L'étiquetage des corpus apporte beaucoup de bénéfices. En particulier, elle ouvre plusieurs portes d'entrée dans les textes. Par exemple, on retrouve enfin les substantifs "devoir", "pouvoir", "savoir"… le point cardinal "est" ou la belle saison perdue dans les participes passés de "être", etc. Sans doute, cela peut-il paraître trivial mais il faut rappeler qu'en moyenne plus du tiers des mots d'un texte en français sont "homographes" et que certains sont extrêmement usuels.

De même, le regretté A. Pibarot a montré combien la technique traditionnelle des "segments répétés" gagnait en puissance grâce à cet étiquetage. Le petit exemple sur Corneille aura suggéré bien d'autres pistes encore.

Imaginons maintenant une vaste base de données textuelles normalisées et étiquetées contenant plusieurs millions de mots judicieusement choisis dans les grandes oeuvres du présent et du passé mais aussi dans la presse, sur le net ou dans les conversations quotidiennes. Appelons ce corpus : "échantillon représentatif du français moderne" (puisque la dénomination "trésor" est déjà prise par un recueil de formes graphiques). Grâce à cet étalon de référence, nous pourrons donner instantanément les différents sens d'un mot dans la langue, fournir les citations les plus illustratives et mesurer la singularité de tel ou tel locuteur. Je pense que les lexicologues ou les traducteurs rêvent d'un tel outil !

Derrière ces usagers évidents, combien d'autres frappent à la porte ? Au cours de ce colloque, nous évoquerons sans doute certaines de ces demandes. Les profs qui essaient d'apprendre les langues à leurs élèves ou de leur faire aimer les grandes oeuvres. Les critiques qui s'interrogent sur la construction des phrases, les figures de rhétorique, les métaphores, le style d'un auteur. Les psychologues, les sociologues et leurs inévitables entretiens. Les sondeurs et leurs questions ouvertes. Les journalistes noyés dans leurs collections de journaux

et les dépêches d'agence. Les patrons perplexes devant les archives électroniques impénétrables que leur entreprise a accumulées…

A tous, il faudrait fournir des outils fiables, intelligents, simples d'emploi, intelligibles dans leurs résultats. La normalisation et l'étiquetage des textes, la constitution de corpus de référence sont des points de passage obligés pour extraire la pépite du sens au milieu des tonnes de boue engendrée par les fluctuations graphiques, les règles grammaticales et syntaxiques. Tant que nous n'aurons pas ces étiqueteurs et ces corpus de référence, nos logiciels seront un peu comme la caverne de Platon : ils projettent sur les murs — pardon, sur les "plans factoriels" — des formes vides de contenu.

Naturellement, la machine devra effectuer cette extraction du sens sans que l'usager ait à s'en préoccuper mais elle devra le faire d'une manière sûre, fiable et rigoureusement étalonnée.

En effet, il faut distinguer deux utilisations des données étiquetées. D'une part, il y a l'utilisation standard sur des "corpus kleenex" que l'on jette après analyse (entretiens semi-directifs, questions ouvertes, débats politiques, motions de congrès syndicaux…) Pour cela, on peut envisager une automatisation complète à condition de fixer des normes strictes, des taux d'erreurs réalistes et aussi bas que possible. D'autre part, la réalisation des corpus de référence, l'étude des grandes oeuvres, comme celle que nous venons d'évoquer : le "zéro faute" est impérieux, ce qui interdit l'automatisation totale. Toute intervention manuelle devra être encadrée par des règles également strictes et leur application contrôlée par la machine afin d'éviter les erreurs et les biais.

Ce sera notre conclusion : la normalisation et l'étiquetage des données textuelles sont des nécessités impérieuses. Elles doivent être effectuées selon des standards explicites. Ces standards existent déjà. On les trouvera, entre autres, dans les travaux pionniers de G. Gougenheim, C. Muller, A. Juilland, C. Bernet ou G. Engwall (dans l'ordre de parution) et dans la nomenclature des bons dictionnaires (citons particulièrement celui de Hatzfeld et Darmeisteter).

Malgré ces travaux pionniers, nous ne disposons pour l'instant d'aucun étiqueteur crédible ni de corpus de référence en français alors que, depuis longtemps, nos collègues anglais ont des outils d'étiquetage et plusieurs corpus étiquetés de bonne taille.

Pour le français, les perspectives qu'on vient de tracer sont donc peut-être des chimères. Mais, pour ouvrir un colloque comme celui-ci, il est toujours bon de laisser une petite part au rêve !

Annexe

L'univers lexical de "amour" chez P. Corneille

(Classement par catégories grammaticales et par liaison décroissante, seuil de 1%)

1° Les suremplois :

- **Noms propres** : Vénus, Léon, Zéphir, Psyché
- **Verbes** : céder, éteindre, opposer, allumer, naître, éclater, trahir, paraître, intéresser, aimer, surmonter, succéder, croître, vaincre, étouffer, pardonner, inspirer, tourner, déférer, rallumer, gémir, éprouver, couronner, mériter, brûler, souffrir, presser, faire, fléchir, produire, combattre, seoir, changer, traiter, flatter, vouloir, préférer, devoir, renaître, unir, animer
- **Substantifs** : haine, tour, jour, retour, coeur, excès, amitié, amorce, cour, noeud, estime, tendresse, objet, beauté, séjour, douceur, ardeur, soin, force, pitié, feu, espérance, respect, désir, devoir, loi, idolâtrie, aile, espoir, dépit, mère, excuse, prix, caresse, cause, impatience, faveur, discours, partage, jeunesse, balance, violence, patrie, divorce, feinte, conduite, lien, mérite, raison, transport, froideur, amant, effort, hymen, ambition, maîtresse, passion
- **Adjectifs** : conjugal, parfait, paternel, véritable, fort, extrême, tendre, aimé, doux, éteint, chaste, puissant, fou, mutuel, forcé, vertueux, éternel, solide, aveugle, feint, simple, léger, indigne, aimable, beau, ferme
- **Pronoms** : dont, se, qui, il, lui
- **Adverbes** : peu, toujours, plus, d'autant, aussi, ensemble, tant, quelquefois, si, auprès
- **Déterminants** : mon, premier, tel
- **Prépositions et conjonctions** : que, malgré, ni, soit, pour, vers, dans, quand, contre, car

2° Les sous-emplois :

- **Noms propres** : Romain, Rome, César, Pompée
- **Verbes** : être, dire, aller, laisser, venir, prendre, arriver, attendre, connaître, penser, pouvoir, sortir, revoir, sembler, amener, craindre, hâter, choisir, trancher, couler, falloir, recevoir, prétendre, défaire, secourir, tomber, plaindre, rougir, marcher, pousser, suivre, fuir, punir, ouvrir, chercher, éviter, perdre, garantir, vanter, mentir, achever, pleurer
- **Substantifs** : seigneur, dieu, ciel, roi, adieu, madame, mot, prince, gens, terre, pied, monsieur, humeur, ordre, heure, ami, homme, comte, mort, lieu, temps, sort, tête, loisir, malheur, mort, avis, coup, combat, soeur, traître, destin, frère, ouvrage, bonheur, guerre, fer, zèle, sang, foudre, bataille, ombre, assassin, main, mal, monstre, père, événement, réponse, fois, oeil, victime, chef, vie, nombre, bourreau, soldat, avenir, affection, fils, pas, châtiment, place, porte, conseil, peuple, épée, parole, effroi, sujet, fortune, état, point, autel, comble, artifice, encens, gendre, assurance, vérité
- **Adjectifs** : funeste, prêt, autre, bon, faux, las
- **Pronoms** : tu, vous, ils, en, quoi, nous, y, cela, autre, leur, vous-même
- **Adverbes** : là, demain, bien, vrai, pas, déjà, bas, trop, encore, oui, ici, pourtant, mieux, tout
- **Déterminants** : quel, second, ce, ton, tout, trois
- **Prépositions et conjonctions** : donc, après, voici, jusque, avec, mais, si, sur

Phrases les plus caractéristiques :

"A votre sûreté, puisque le péril presse, j'immolerai ma flamme et toute ma tendresse ; et je vaincrai l'horreur d'un si cruel devoir pour conserver le jour à qui me l'a fait voir ; mais ce qu'à mes désirs je fais de violence fuit les honteux appas d'une indigne espérance ; et la vertu qui dompte et bannit mon amour n'en souffrira jamais qu'un vertueux retour" (Othon, acte 1)

"Tu ne le comprends point ! et c'est ce qui m'étonne : je veux donner son coeur, non que son coeur le donne ; je veux que son respect l'empêche de m'aimer, non des flammes qu'une autre a su mieux allumer ; je veux bien plus : qu'il m'aime, et qu'un juste silence fasse à des feux pareils pareille violence ; que l'inégalité lui donne même ennui ; qu'il souffre autant pour moi que je souffre pour lui ; que par le seul dessein d'affermir sa fortune, et non point par amour, il se donne à quelqu'une ; que par mon ordre seul il s'y laisse obliger ; que ce soit m'obéir, et non me négliger ; et que voyant ma flamme à l'honorer trop prompte, il m'ôte de péril sans me faire de honte" (Don Sanche, acte 3).

"Je vous aime, et non point de cette folle ardeur que les yeux éblouis font maîtresse du coeur, non d'un amour conçu par les sens en tumulte, à qui l'âme applaudit sans qu'elle se consulte, et qui ne concevant que d'aveugles désirs, languit dans les faveurs, et meurt dans les plaisirs : ma passion pour vous, généreuse et solide, a la vertu pour âme, et la raison pour guide, la gloire pour objet, et veut sous votre loi mettre en ce jour illustre et l'univers et moi." (Pulchérie, acte 1).

"Son oeil agit sur moi d'une vertu si forte, qu'il ranime soudain mon espérance morte, combat les déplaisirs de mon coeur irrité, et soutient mon amour contre sa cruauté ; mais ce flatteur espoir qu'il rejette en mon âme n'est qu'un doux imposteur qu'autorise ma flamme, et qui sans m'assurer ce qu'il semble m'offrir, me fait plaire en ma peine, et m'obstine à souffrir" (Mélite, acte 1).

"Donne un plus digne nom au glorieux empire du noble sentiment que la vertu m'inspire, et que l'honneur oppose au coup précipité de mon ingratitude et de ma lâcheté ; mais plutôt continue à le nommer foiblesse, puisqu'il devient si foible auprès d'une maîtresse, qu'il respecte un amour qu'il devroit étouffer, ou que s'il le combat, il n'ose en triompher" (Cinna, acte 3).

"Je ne puis, tant elle a de mépris et d'appas, ni le faire accepter, ni ne le donner pas ; et comme si l'amour faisoit naître sa haine, ou qu'elle mesurât ses plaisirs à ma peine, on voit paroître ensemble, et croître également, ma flamme et ses froideurs, sa joie et mon tourment" (La Suivante, acte 3).

"Vous céder par dépit, et d'un ton menaçant faire voir qu'on pénètre au coeur du plus puissant, qu'on sait de ses refus la plus secrète cause, ce n'est pas tant céder l'objet de son amour, que presser un rival de paroître en plein jour, et montrer qu'à ses voeux hautement on s'oppose" (Agesilas, acte 4).

Bibliographie :

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

ENGWALL Gunnel, 1984, *Vocabulaire du roman français (1962-1968) Dictionnaire des fréquences*, Stockholm, Almqvist-Wicksell International.

GOUGENHEIM Georges et Al, 1964, *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier.

HATZFELD Adolphe, DARMEISTETER Arsène, THOMAS Antoine, 1898, *Dictionnaire général de la langue française du commencement du XVIIe siècle jusqu'à nos jours, Paris, Delagrave*, 1898 environ.

HUBERT Pierre, LABBE Dominique, 1995, "La structure du vocabulaire du général de Gaulle" in BOLASCO Sergio et AL, I*Ie Giornate internazionali di analisi statistica dei dati testuali*, Rome, CISU, II, p 165-176.

JUILLAND Alphonse, BRODIN Dorothy, DAVIDOVITCH Catherine, 1970, *Frequency Dictionnary of French Words*, La Haye, Mouton.

LABBE Dominique, 1990b, *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahier du CERAT.

LABBE Dominique, 1990c, *Le vocabulaire de François Mitterrand*, Paris, Presses de la Fondation nationale des sciences politiques.

LAFON Pierre, 1984, *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.

LEBART Ludovic , MORINEAU Alain, PIRON Marie, 1995, *Statistique exploratoire multidimensionnelle*, Paris, Dunod.

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

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

MULLER Charles, 1977, *Principes et méthodes de statistique lexicale*, Paris, Hachette.

PIBAROT André, LABBE Dominique, 1998, "Les syntagmes répétés dans l'analyse des commentaires libres", *Communication aux 4e Journées d'analyse des données textuelles*, Nice.