La lemmatisation des grandes bases de textes. Un exemple : Corneille, Molière et Racine

Dominique Labbé

L'édition électronique en littérature et dictionnairique, évaluation et bilan

Rouen, Campus universitaire de Mont-Saint-Aignan

17-21 juin 2002

LA LEMMATISATION DES GRANDES BASES DE TEXTES
Un exemple : Corneille, Molière et Racine

Dominique Labbé
CERAT
Institut d'Etudes Politiques
BP 48
38040 Grenoble Cedex 9
dominique.labbe@iep.upmf-grenoble.fr

Avec l'exemple des pièces de Corneille, Molière et Racine, on montre quelques-uns des nombreux usages possibles des bases de données textuelles normalisées et lemmatisées. Elles sont d'une consultation aisée. Elles fournissent de nombreux renseignements sur le vocabulaire, le style, le sens des mots... Pour cela, il faut réduire les graphies multiples et rattacher chaque mot à son entrée de dictionnaire.

Lors de ce colloque, plusieurs communication ont expliqué ce qu'est le balisage et l'étiquetage des textes mais elles ont aussi montré que ces opérations sont coûteuses, au moins en temps. Un tel investissement en vaut-il la peine ? La réponse semble ne pas faire de doute mais le mieux n'est-il pas de le montrer par l'exemple ? C'est pourquoi on commencera par les usages possibles des grandes bases de textes lemmatisées avant d'expliquer la manière dont on obtient ces bases.
Le colloque se tenant à Rouen, le choix de Corneille nous a paru s'imposer. Nous y avons ajouté deux illustres contemporains : Molière et Racine afin de constituer une sorte d'embryon de ce qui pourrait être la base de données textuelles « théâtre classique ». Enfin, signalons que ces oeuvres ont déjà fait l'objet d'études statistiques qui peuvent servir d'utiles points de repère (Bernet 1983 ; Kylander 1995 ; Muller 1967).
Les principales caractéristiques de ce corpus sont récapitulées dans le tableau I ci-dessous. N est la taille totale mesurée en nombre de "mots" : la base ainsi constituée comporte un peu plus d'un million de mots. Ce n'est pas encore une "grande" base de données — comme le laisse entendre le titre donné par les organisateurs à ma communication — mais la taille est déjà trop importante pour être traitée « manuellement ». Cet ensemble devrait permettre à l'auditoire de juger des "bénéfices" de l'opération, de comprendre les procédures à suivre et de juger de la "faisabilité" d'une opération identique portant sur de plus grands volumes.

Tableau 1. Corpus « théâtre classique »

Corneille : 34 pièces
N : 553 190 mots V : 15 535 formes graphiques normalisées et 6 258 vocables
Molière : 34 pièces
N : 364 963 ; V : 16 735 formes graphiques et 8 088 vocables
Racine : 12 pièces
N : 166 626 mots ; V : 10 120 formes graphiques et 4 323 vocables
Total : 80 pièces
N : 1 084 779 mots ; V : 25 692 formes graphiques normalisées et 10 604 vocables

I. LES INTERETS D'UNE BASE DE DONNEES TEXTUELLES LEMMATISEES

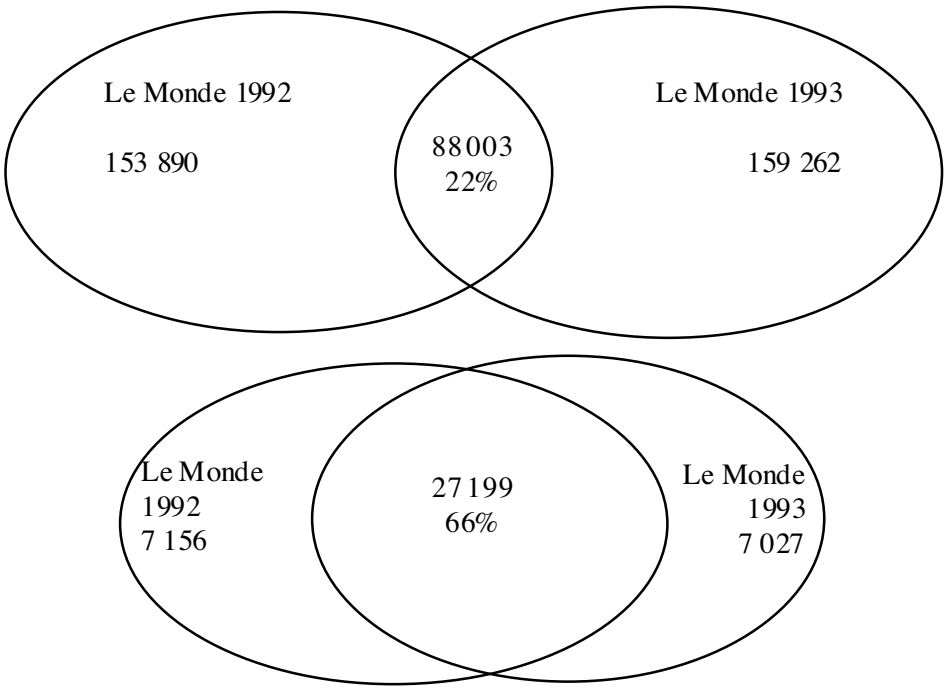
Nous montrerons d'abord que la lemmatisation est indispensable pour permettre la consultation aisée des grandes bases de textes. Au-delà de ce premier bénéfice, la lemmatisation permet une analyse scientifique du vocabulaire d'un auteur, elle aide à retrouver ses principaux thèmes. Enfin, elle ouvre des perspectives nouvelles à la lexicologie ou la stylistique.

Une consultation aisée et sûre des grandes bases

Tout utilisateur des grandes bases actuelles – type Frantext – sait qu'il est pratiquement impossible de retrouver toutes les attestations d'un verbe usuel et quelles ruses, bien souvent déjouées, il doit employer pour parer aux "homographies". De toute façon, à l'issue de procédures longues et complexes, il ne sera jamais certain d'avoir effectivement trouvé toutes les attestations du mot recherché.
Pour le faire comprendre, nous évoquerons une expérience présentée en 1995 par M. Sylberztein sur deux années du journal *Le Monde*, expérience que résument les figures ci-dessous. Sur le CD de 1992, il a décompté 21,8 millions de mots sous près de 242 000 formes différentes et, l'année suivante : 23,2 millions sous 247 000 formes. Mais surtout, la comparaison du vocabulaire sur les deux années fait apparaître un noyau commun ridiculement faible. Le "visiteur" se trouve face à un "vocabulaire" de plus de 400.000 "mots" différents – comment travailler avec un dictionnaire d'une telle taille ? – dont la plupart ne sont utilisés que dans l'une des deux années : le journal aurait-il changé de langue entre 1992 et 1993 ?
Après normalisation des graphies et lemmatisation, le tableau est radicalement différent. Le vocabulaire ne compte plus "que" 41.000 entrées dont les deux tiers sont communes aux deux

années… Dans les fichiers originaux, la grande majorité des formes graphiques "différentes" étaient des coquilles, des variantes pour un même mot (surtout les noms propres étrangers)… L'expérience prouve que la correction orthographique, la normalisation des graphies et la lemmatisation sont des opérations indispensables pour mettre à disposition une information fiable.

Tableau 2. Le "vocabulaire" du Monde sur deux années (en formes graphiques brutes puis en lemmes)



Actuellement les grandes bases de textes sont des "émeutes de formes", comme "le pittoresque est une émeute de détails" (Baudelaire). Sauf pour les amateurs de pittoresque, leur visite est souvent décevante !

Le premier intérêt de la lemmatisation est donc de lever cet obstacle : on peut entrer dans la base comme on consulte le dictionnaire et répondre avec certitude à cette question simple : le mot est-il présent ou non dans le ou les textes considérés et si oui : combien ? Pour répondre à cette question, l'algorithme n'a d'ailleurs pas besoin de consulter la base proprement dite mais l'index des lemmes et des formes.
On sait que, dès l'origine de l'imprimerie, l'index fait partie de l'édition savante des œuvres imprimées. L'édition électronique peut-elle faire moins ? Le tableau 3 donne quelques extraits de l'index des trois œuvres. Il est classé par ordre alphabétique et respecte les habitudes des locuteurs français. Par exemple, si l'on s'intéresse au verbe *être*, l'index indique qu'il apparaît un peu plus de 26 000 fois et détaille l'ensemble de ses flexions. On est sûr qu'il n'y a aucun point cardinal (est, nm), aucune saison estivale (été, nm), aucun substantif "être", ou "sommes" (qui n'a d'ailleurs pas le même sens au masculin et au féminin) ni la conjonction "soit", le verbe "suivre", etc… De même, le verbe *pouvoir* exclut les substantifs, les adverbes, etc.
Je ne suis pas allé chercher un exemple "corsé". J'ai simplement choisi le verbe le plus employé en français. En fait, dans tout texte écrit en langue française, plus du tiers des mots sont "homographes" : une graphie et au moins deux entrées différentes dans le dictionnaire. De plus, les deux tiers des mots ont plusieurs flexions… C'est pourquoi les grandes bases de données textuelles sont si malcommodes.
Grâce à la lemmatisation, on saura avec certitude si le mot recherché se trouve bien dans la base. On peut alors lancer la recherche de ses attestations en combinant vocables et formes graphiques et obtenir une "concordance", c'est-à-dire les mots devant et derrière chaque attestation (en général une ligne d'imprimante : 120 caractères). Par exemple, le verbe "suivre" conjugué à la première et à la deuxième

personne chez Molière (tableau 4). Dans une base en formes graphiques, il aurait fallu extraire ces quelques lignes au milieu de quelque 2000 verbes "être" : tâche évidemment impossible !

Tableau 3. Extraits de l'index alphabétique du théâtre classique :

être (v)	26 321
est	14 422
été	245
être	1 976
fût	197
soit	835
sommes	185
suis	1 934
être (nm)	11
pouvoir (v)	6 695
pouvoir	214
puis	1 090
pouvoir (nm)	506
puis (adv)	84
soit (cj)	157
somme (nm)	2
somme (nf)	18
somme	16
sommes	2
suivre (v)	856
suis	51

Naturellement, si l'on est intéressé par telle ou telle citation, on peut élargir le contexte avant et/ou arrière et il n'est pas forcément inutile d'aller chercher l'original papier dans sa bibliothèque !
Quand un mot est relativement rare et univoque, la consultation des concordances permet de trouver sa signification et les citations intéressantes, sans avoir besoin d'aller plus loin. Par exemple, dans le théâtre classique, la consultation de la concordance suffit pour indiquer que le substantif "pouvoir" ne désigne généralement pas l'aspect politique mais la passion — plus précisément l'influence que le sexe dit "faible" exerce, sur les personnages masculins, grâce à ses "charmes" et, en premier lieu, le *pouvoir* des *yeux*…
Cet exemple appelle deux remarques.
D'une part, la concordance doit restituer exactement le texte original, ainsi les conjugaisons des verbes en "oit", les majuscules initiales de phrase ou de vers, etc… A ce propos, l'extrait présenté dans le tableau 4 montre que notre outil n'est pas parfait : il manque les tirets unissant le verbe et le pronom post-posé (suis-moi). On pourrait aussi souhaiter des indications de page, savoir quel personnage parle, etc… Mais tout cela serait de peu d'utilité si l'on ne disposait pas d'abord de la porte d'entrée par le lemme, sa catégorie grammaticale et ses flexions.
D'autre part, à ce simple stade la lemmatisation a résolu la plupart des problèmes que se posent les usagers des grandes bases de données textuelles. Mais on peut aussi leur offrir quelques renseignements supplémentaires notamment sur le vocabulaire des auteurs étudiés.

Tableau 4. Concordances du vocable : "suis, suivre" dans les comédies de Molière

L'étourdi (Acte 3)		
ivertir. Mais aussi, raisonnons un peu sans violence : si je	suis	maintenant ma juste impatience, on dira que je cède à
Dépit amoureux (Acte 3)		
and je suis regardé, et mon trépas ainsi se verroit retardé.	Suis	moi, traître, suis moi : mon amour en furie te fera vo
, et mon trépas ainsi se verroit retardé. Suis moi, traître,	suis	moi : mon amour en furie te fera voir si c'est matière
Dom Garcie (Acte 32)		
nce que les soins approuvés d'un peu de complaisance, et qui	suis	seulement par d'utiles leçons la pente qu'a le prince
Dom Juan (Acte 5)		
oi qu'il arrive, que je sois capable de me repentir. Allons,	suis	moi. Arrêtez, dom Juan : vous m'avez hier donné parole
Mélicerte (Acte 1)		
nte Daphné ! Trop aimable Eroxène. Acante, laisse moi. Ne me	suis	point, Tyrène. Pourquoi me chasses tu ? Pourquoi fuis
Amphitryon (Acte 2)		
illé même étoit l'autre Sosie, quand il m'a si bien étrillé.	Suis	moi. Je t'impose silence : c'est trop me fatiguer l'es
Faut il… Je ne puis rien entendre : laisse moi seule, et ne	suis	point mes pas. Il faut que quelque chose ait brouillé
Amphitryon (Acte 3)		
; et l'on me dés-Sosie enfin comme on vous dés-Amphitryonne.	Suis	moi. N'est il pas mieux de voir s'il vient personne ?
Georges Dandin (Acte 3)		
t'ai je amené avec moi pour l'entretenir. Monsieur, je vous	suis	… Chut ! J'entends quelque bruit. Claudine. Hé bien ?
Fourberies de Scapin (Acte 1)		
ibonds. Marche un peu en roi de théâtre. Voilà qui est bien.	Suis	moi. J'ai des secrets pour déguiser ton visage et ta v
Comtesse d'Escarbagnas		
c'est trop longtemps, Iris, me mettre à la torture, et si je	suis	vos lois, je les blâme tout bas de me forcer à taire u
Femmes savantes (Acte 3)		
jamais vu ce fol entêtement ; et d'un Grec là-dessus je	suis	le sentiment, qui, par un dogme exprès, défend à

Principales caractéristiques du vocabulaire de Corneille, Molière et Racine

Commençons par les questions classiques qui sont toujours posées à propos d'un auteur : quels sont les verbes ou les substantifs favoris ? Ou encore : quels sont les mots qu'il emploie tout le temps et, au contraire, ceux qui ne sont employés que dans telle ou telle oeuvre… Les "index hiérarchiques" – qui classent les vocables en fonction de leur fréquence d'emploi — fournissent la réponse à la première question et les tableaux 5, 6 et 7 donnent cette information pour les trois auteurs auxquels nous nous intéressons. Puisque les 3 œuvres n'ont pas les mêmes dimensions, les fréquences absolues seraient de peu d'utilité. Pour les rendre comparables, on les convertit donc en fréquences relatives.

Tableau 5. Les dix premiers verbes (fréquence pour mille mots)

Corneille		Molière		Racine	
être	21,8	être	30,4	être	19,0
avoir	18,4	avoir	20,4	avoir	16,5
faire	9,2	faire	11,3	voir	6,2
pouvoir	6,5	dire	6,4	pouvoir	6,1
voir	6,1	vouloir	6,1	faire	5,9
vouloir	4,5	voir	6,0	vouloir	4,3
savoir	3,4	pouvoir	5,7	aller	3,9
dire	3,2	savoir	4,0	dire	3,0
aimer	3,1	aller	4,0	venir	2,9
devoir	2,6	venir	3,0	savoir	2,9

A part de légères différences de classement, les principaux verbes sont communs (*être*, *avoir* et *faire* arrivent pratiquement toujours en tête dans tout texte en français). Corneille privilégie *savoir*, *aimer*, *devoir* ; Molière : *dire* ; Racine : *voir*, *aller*, *venir*… On notera également que chez Molière les trois premiers occupent une place nettement plus forte : c'est une caractéristique du français "oral" et Molière passe justement pour avoir porté sur les planches le parler de ses contemporains.

Tableau 6. Les dix premiers substantifs (fréquence pour mille mots)

Corneille		Molière		Racine	
amour	3,4	monsieur	4,3	seigneur	3,5
coeur	3,2	chose	2,6	coeur	3,2
âme	2,0	coeur	2,6	madame	3,0
oeil	1,9	homme	2,0	oeil	3,0
seigneur	1,9	madame	2,0	dieu	2,9
roi	1,8	amour	1,5	amour	2,6
main	1,6	monde	1,4	roi	2,3
madame	1,6	père	1,3	jour	2,2
jour	1,3	oeil	1,3	fils	2,0
sang	1,2	fille	1,2	père	2,0

La comparaison entre les colonnes du tableau 6 suggère déjà quelques différences thématiques entre les auteurs. Dans les comédies de Molière, *monsieur* et *homme* prennent la place de *seigneur* et de *roi*… Corneille privilégie *amour* et Racine *seigneur* (la manière de s'adresser aux puissants comme on le verra plus bas) et Molière : *Monsieur*… *Dieu* est plus présent chez Racine (5^{e} position) que chez Corneille (11^{e} position) ou chez Molière (19^{e}). Et cela n'est pas simplement dû à ses deux dernières

pièces. On notera également : chez Corneille : *main* et *sang ;* chez Molière : *chose* et *monde* qui sont également des caractéristiques du discours oral (on dirait aujourd'hui : *les gens*) ainsi que *fille ;* chez Racine : *fils* et *père*…

Tableau 7. Les dix premiers adjectifs

Corneille		Molière		Racine	
grand	2,2	grand	1,8	seul	1,7
seul	1,4	beau	1,7	grand	1,1
beau	1,1	bon	1,6	cruel	1,1
doux	0,9	petit	1,0	heureux	1,0
vain	0,8	doux	0,7	cher	0,9
digne	0,8	vrai	0,6	vain	0,8
heureux	0,8	seul	0,5	funeste	0,7
juste	0,7	jeune	0,4	prêt	0,7
bon	0,6	honnête	0,4	triste	0,7
cher	0,6	pauvre	0,4	beau	0,6

Grand est pratiquement toujours l'adjectif le plus employé, mais Racine préfère la "solitude" (ou les expressions construites avec cet adjectif) et Molière ne sous-emploie pas trop *petit* comme le font les deux autres… Par contraste, Racine semble privilégier : *cruel*, *funeste*, *triste* ; chez Corneille : *digne* et *juste* ; chez Molière : *petit*, *vrai*, *jeune*, *honnête*, *pauvre*.

Les spécialistes de ces auteurs sauront commenter ces classements suggestifs mais on peut avoir de sérieuses réserves face à ces listes. Notamment, on sait bien que ces vocables très usuels sont fortement polysémiques. Pour comprendre le sens de l'un ou l'autre d'entre eux, il faut donc regarder avec quels autres mots il se combine. Mais les concordances deviennent des outils difficilement maniables dès qu'elles sont trop volumineuses. On peut donc demander à l'ordinateur de rechercher les combinaisons les plus fréquentes dans le corpus qui nous intéresse.

Les syntagmes sont des combinaisons stables de plusieurs vocables qui conservent leur indépendance contrairement au mot composé et à la locution, où ils la perdent (Pibarot et Labbé, 1998). Puisqu'on dispose des catégories grammaticales, on peut s'intéresser à certaines combinaisons particulières. Dans la langue française, les "pseudo-auxiliaires" — ou "verbes modaux" qui sont suivis d'un infinitif (sur modèle "pouvoir faire") — tiennent une place importante. Le tableau 8 ci-dessous récapitule les "combinaisons favorites" de nos trois auteurs. La requête indique que le premier membre de la combinaison peut apparaître sous toutes ses flexions alors que le second est toujours à l'infinitif ; de plus, on indique à l'automate que des adverbes ou des locutions adverbiales peuvent se glisser dans le couple recherché (par exemple : "faire (très bien) voir", etc). Enfin, étant donné le caractère beaucoup plus discret du phénomène, la fréquence est ici calculée sur la base de 100.000 mots.
Les premiers syntagmes occupent une surface presque double chez Corneille et Molière que chez Racine. Corneille et Molière partagent les trois constructions préférées ce qui ne manquera pas de surprendre l'analyste tant ces constructions sont propres à chacun (un peu à la manière d'une empreinte digitale). Les modalités "pouvoir" et "faire" dominent chez les trois. La troisième est chez Corneille : *devoir* ; chez Molière : "vouloir" ; chez Racine : "aller".
De même, les combinaisons "nom + nom" sont généralement très suggestives quand on recherche les thèmes favoris d'un auteur ou ses "tics" d'écriture (tableau 9).

Tableau 8. Les syntagmes "pseudo-auxiliaire + infinitif" (fréquence pour 100.000)

Corneille		Molière		Racine	
Syntagmes	Fréquence	Syntagmes	Fréquence	Syntagmes	Fréquence
faire voir	33,8	faire voir	31,5	aller voir	12,0
pouvoir être	18,8	pouvoir être	25,5	pouvoir voir	9,6
pouvoir faire	18,4	pouvoir faire	25,5	faire entendre	9,0
faire naître	13,9	vouloir dire	24,9	pouvoir faire	8,4
pouvoir voir	13,4	vouloir faire	19,5	aller chercher	7,8
devoir être	12,7	pouvoir dire	14,5	faire parler	7,8
pouvoir souffrir	10,8	pouvoir avoir	13,7	pouvoir être	7,8
vouloir faire	9,9	aller faire	13,2	venir chercher	7,2
faire connaître	9,6	avoir faire	13,2	faire éclater	6,6
devoir faire	8,7	pouvoir voir	12,3	falloir partir	6,6

Tableau 9. Les syntagmes "nom + nom" (fréquence pour 100.000)

Corneille		Molière		Racine	
Syntagmes	Fréquence	Syntagmes	Fréquence	Syntagmes	Fréquence
grand cœur	17,9	honnête homme	14,0	nom dieu	8,4
grand roi	8,0	coup bâton	11,5	fils Hector	6,6
grand courage	7,1	monsieur Purgon	8,8	soin vie	6,6
grand nom	6,5	dom Juan	8,2	fond cœur	6,0
beau feu	5,6	beau esprit	7,9	sang roi	6,0
beau œil	5,4	beau œil	7,4	tour tour	6,0
grand homme	5,4	dieu merci	7,1	bout univers	4,8
dernier soupir	5,1	grâce ciel	7,1	pied autel	4,2
haut fait	5,1	monsieur Jourdain	6,6	porte palais	4,2
juste courroux	5,1	monsieur Gorgibus	6,3	sang frère	4,2

L'adjectif placé devant le substantif est caractéristique du style de Corneille, caractéristique qu'il partage avec Molière (quand celui-ci ne sacrifie pas aux nécessités du dialogue "parlé"). Pour Racine, il s'agit du syntagme "nom + préposition + nom". Mais comme pour la combinaison "verbe + verbe", les fréquences relatives sont nettement plus faibles chez Racine que chez les deux autres : expression plus variée et renouvellement plus important d'une pièce à l'autre…
Naturellement, il faudrait examiner des listes plus longues pour pénétrer plus avant dans le vocabulaire de chacun de ces auteurs, mais ces exemples auront fait comprendre que l'étude systématique des syntagmes apportera à l'analyse littéraire beaucoup de renseignements précieux. On pourra aller bien au-delà encore dans la recherche du sens.

A la recherche du sens des mots

Il y a quarante ans, dans la polémique qui l'opposait à R. Picard, R. Barthes avait justement fait remarquer que : "Structuralement, le sens ne naît point par répétition mais par différence, en sorte qu'un mot rare, dès lors qu'il est saisi dans un système d'exclusions et de relations, signifie tout autant qu'un terme fréquent" (Barthes, 1966). Bien que trop générale, l'objection comporte une part de bon sens : dans un système comme la langue, le sens d'une unité lui vient d'abord des relations qu'elle entretient avec tous les autres mots du lexique. A condition d'utiliser les textes lemmatisés, la

statistique textuelle peut retrouver ce système d'exclusions et d'associations qui fait le sens d'un mot et qui lui donne son importance réelle dans un corpus donné. Nous donnons en annexe à cette communication une présentation du calcul des univers lexicaux et la méthode sera illustrée avec les substantifs les plus fréquemment employés par Corneille et Racine.
L'algorithme reconstitue le vocabulaire utilisé avec le mot considéré et compare ce vocabulaire avec celui employé dans l'ensemble de l'œuvre, ce qui lui permet d'isoler, avec moins de 1% de chances de se tromper, les mots significativement sur-employés – ils sont attirés par le mot – et ceux qui sont sous-employés (ils sont "repoussés").

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

---

Chez Corneille les emplois du mot « amour » semblent essentiellement organisées autour de trois thèmes : les "feux", les "noeuds", les obstacles (haine, devoir, froideur…) Par exemple, la métaphore du feu est attestée par la présence des verbes : *éteindre, allumer, étouffer, rallumer, brûler…*, des noms : *feu, ardeur, froideur, flamme*, etc. et des adjectifs : *éteint*, *puissant…* Quant aux mots qui ne sont pas associés à l'amour dans l'esprit de Corneille, on y trouve *dieu* (les cieux), le *prince* et le *pouvoir* mais aussi le *père* et le *frère…*
La présence du pronom "tu " dans la liste des spécificités négatives peut étonner. En fait, il signale un procédé constant chez Corneille : les amants se parlent rarement ; habituellement, l'amour est une confidence que l'on fait à un tiers en dehors de la présence de l'être aimé.

On aura certainement été surpris de constater que "haine" est le substantif le plus fortement associé à l'amour. Cette "dialectique" des contraires est un thème très fréquent dans l'œuvre de Corneille. Celle-ci est éclairée par la lecture des phrases les plus significatives. Après avoir établi l'univers du mot, le logiciel relit l'ensemble de l'œuvre, en recherchant les phrases contenant le plus grand nombre de mots appartenant à l'univers de l'amour et le plus petit nombre de mots exclus de cet univers. On ne sera pas surpris par le passage le plus caractéristique (ci-dessous).

Les vers les plus caractéristiques de l'amour chez P. Corneille

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)

La même expérience sur "seigneur", le substantif le plus employé par Racine conduit à des résultats aussi intéressants (ci-dessous).

L'univers lexical de "seigneur" chez Racine
(Classement par catégories grammaticales et par liaison décroissante, seuil de 1%)

Vocables significativement sur-employés dans l'univers

Noms propres : Bérénice, Hermione
Verbes : daigner, tomber, douter, croire, demander, trahir, achever, aimer, songer
Substantifs : reine, mystère, sénat, victime, voeu, vertu, conquête, haine, malheur, loi, nom
Adjectifs : juste, seul, heureux
Pronoms : vous, dont, vôtre
Adverbes : point, si, oui, ne, combien
Déterminants : votre, ce, son,
Conjonctions et prépositions : depuis, mais, sur, entre, si, quand,

Vocables significativement sous-employés dans l'univers

Verbes : devoir, servir, faire, laisser
Substantifs : trône, madame, monsieur
Adjectifs : prêt
Pronoms : ils, qui, lui, tu, il, je, toi
Adverbes : peut-être, enfin, tout
Déterminants : deux, tout, mon, ton
Conjonctions et prépositions : dès, que, pour, et

Phrases les plus spécifiques

BURRHUS.
Vous vous le figurez,
Seigneur ; et satisfait de quelque résistance,
Vous redoutez un mal foible dans sa naissance.

Mais si dans son devoir votre coeur affermi
Vouloit ne point s'entendre avec son ennemi ;
Si de vos premiers ans vous consultiez la gloire ;
Si vous daigniez, Seigneur, rappeler la mémoire
Des vertus d'Octavie, indignes de ce prix,
Et de son chaste amour vainqueur de vos mépris ;
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.
(Britannicus, acte 3)

ARSACE
N'en doutez point, seigneur, tout succède à vos vœux.
(Bérénice, acte 3).

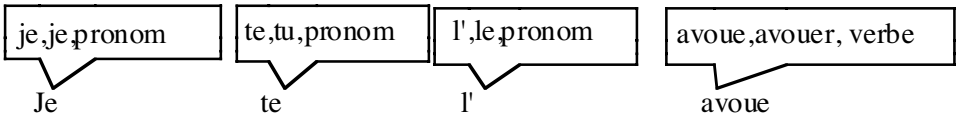
Ces résultats suggèrent un univers particulièrement pauvre, ou plutôt un emploi "banal" du mot "seigneur" : Racine l'associe pratiquement à tout le reste du vocabulaire sauf : "monsieur", "je" et "tu". En effet, les citations relevées comme les plus caractéristiques par le logiciel indiquent que Racine en parsème les propos des courtisans, quand il les fait s'adresser au souverain (noter à ce sujet, que l'initiale du mot est parfois placée en majuscule, ce qui illustre l'importance de ces "détails"). Chez Racine, le mot "Seigneur" a donc pratiquement pour seule fonction de signaler la déférence de celui qui parle envers celui auquel il s'adresse. Or c'est le substantif le plus fréquent ! Ainsi se trouve vérifiée l'intuition de R. Barthes : pour un mot, la répétition n'est pas un gage d'importance sémantique…

Il n'était évidemment pas question, en un temps si court, d'illustrer toutes les pistes ouvertes par la lemmatisation des bases de données textuelles, pistes dont on commence d'ailleurs tout juste à soupçonner l'ampleur et la portée. Grâce à ces bases, on pourra relire les auteurs avec de nouveaux outils, étudier la richesse et la diversité de leur vocabulaire comme l'ont fait C. Muller sur Corneille et C. Bernet sur Racine, repérer les ruptures thématiques (Hubert et Labbé 2002), voire : permettre l'attribution à leur auteur de certains textes dont l'origine est aujourd'hui douteuse ou inconnue…

Ces grandes bases devraient également renouveler la linguistique qui ne travaillera plus avec des modèles idéaux ou des exemples choisis arbitrairement, mais en prenant comme base la langue telle qu'elle est effectivement parlée et écrite par ses usagers (Habert, Nazarenko, Salem 1997). Suivant l'intuition des fondateurs de l'Inventaire de la langue française puis de son Trésor, la lexicographie devrait y trouver les outils pour la recherche automatique des synonymes, des antonymes, des hyperonymes, des exemples canoniques… (Leselbaum, Labbé, 2002). Quant à la grammaire et à la stylistique, les grands corpus étiquetés ouvrent la possibilité d'étudier systématiquement le rôle des catégories grammaticales ou encore les structures de phrases (Monière-Labbé 2002).

II. COMMENT FAIRE ?

La normalisation et la lemmatisation consistent à attacher à chaque mot du texte une étiquette correspondant à son entrée dans le dictionnaire. A titre d'exemple, voici les premiers mots de Mélite (la première comédie de P. Corneille) étiquetés par notre logiciel qui a été mis au point dans les années 1980 pour l'étude du discours politique français contemporain (Labbé, 1990).



"Je" est le "mot" ou la "forme graphique brute" ; elle reste sans changement et à sa place dans le texte. Dans l'étiquette, "je" est la "forme graphique normalisée" et "je,pronom" son "entrée de dictionnaire" ou lemme (mot vedette et catégorie grammaticale).

Normalisation des graphies

Est-il nécessaire de rappeler que**,** avant tout, il est nécessaire d'effectuer une correction orthographique soignée**.** Ceci fait, on isole par des balises tout ce qui n'est pas le texte mais que l'on estime devoir être conservé. Par exemple, les noms des personnages, les numéros des vers, les indications scéniques… doivent être entourés de balises qui indiquent qu'on ne doit pas les compter comme du texte mais qu'ils doivent, éventuellement, être restitués à l'utilisateur. Pour ce balisage, les exposés, qui ont précédé mon intervention, ont montré qu'il est nécessaire de respecter les standards type **X**ML… Nous n'y revenons donc pas.
C'est donc ce texte corrigé et balisé qui est soumis au logiciel. Celui-ci commence par normaliser les graphies multiples d'un même mot, la graphie standard étant placée au début de l'étiquette. Par exemple "puis" et "peux" (verbe pouvoir). Pour les conventions à retenir, voir par exemple, le travail réalisé par le conseil de la langue française sur l'harmonisation des dictionnaires (CLIF, 1988). Il faut également réduire les majuscules initiales de phrases (ou de vers) et celles qui sont données aux "faux noms propres". Répétons-le : nous ne modifions pas le texte original, nous y ajoutons cette information. Par exemple, on ne "corrige" pas Racine : *Seigneur* garde sa majuscule — puisque l'auteur l'a voulu ainsi et cette graphie sera retournée à l'écran ou sur papier — mais, dans l'étiquette, la graphie normalisée sera en minuscule et elle sera codifiée comme un nom commun et non pas un nom propre…
De même le logiciel invite à déployer les abréviations à chaque fois que possible (M. : monsieur, Marcel, Maurice… mètre ?).
Enfin, il agglutine les mots composés selon des règles restrictives. Par exemple, "aujourd'hui" ou "bric à brac" sont considérés comme un seul mot mais pas "pomme de terre". Certes, on peut aussi placer dans le fichier étiqueté, des sortes de "macro-balises" pour isoler ces "locutions", pour indiquer qu'on est face à un bloc ayant son unité même s'il est composé de trois mots ayant leur autonomie. Mais il faut envisager cette solution avec prudence car ce travail est fort lourd, il repose sur des critères largement subjectifs et ne pourra jamais être exhaustif…

Lemmatisation

La lemmatisation rattache chaque forme normalisée à son entrée de dictionnaire à l'aide d'une liste de règles attachées à ce mot ou à un petit groupe de mots ayant des comportements syntaxiques similaires. Pour un exemple, voir en annexe les règles de lemmatisation de "tout", qui est certainement l'un des cas les plus complexes, et le tableau des catégories grammaticales utilisées dans nos étiquettes (annexe à cette communication).
Par exemple :
— sous l'entrée "le, article", on trouve : L', La, Le, Les, l', la, le, les… Mais pas le pronom "le" ni "La Fontaine (Monsieur de…)"
— sous l'entrée "avouer, verbe", on trouve : *avoue*, *avoues*, *avouons*… mais pas : "*avoué*, nom masculin" ;
La lemmatisation est **complète** et **univoque** (tout mot reçoit une étiquette et une seule) **stable** (les mêmes conventions sont strictement appliquées du début à la fin de l'opération) et **sans double compte**.
La nomenclature doit **respecter les conventions lexicographiques** et les habitudes des usagers du français. Mais elle doit être **systématique**, c'est-à-dire qu'elle explicite et "durcit" ces conventions souvent assez "molles" (la mollesse étant une caractéristique qui s'adapte mal à la programmation informatique) !
Par exemple,
— si l'on distingue certains substantifs par leur genre (*garde*, *mode*, *tour*…) alors tous les substantifs doivent avoir un genre et tous les substantifs de même graphie et de même genre doivent être regroupés sous une seule entrée (*grève*, nom féminin n'aura donc qu'une entrée) ;

— toutes les formes verbales ayant même infinitif doivent être groupées ensemble (*voler*, intransitif et transitif : une seule entrée). Sinon, il faudrait distinguer les emplois transitifs et intransitifs pour tous les verbes…
Si nous donnons ces exemples, c'est qu'ils figurent dans les dictionnaires que nous avons le plus souvent employés (*Dictionnaire général* et *Robert*) alors qu'ils forment clairement des violations des principes sur lesquels repose la nomenclature de ces ouvrages…
La lemmatisation doit être aussi **automatique** que possible ce qui oblige à renoncer à fournir certaines informations qui reposent sur une codification essentiellement manuelle (comme les modes et les temps des verbes). A condition de s'en tenir à une nomenclature synthétique (type dictionnaire de langue), en moyenne 99% des mots peuvent être lemmatisés automatiquement. Le "résidu" n'est pas négligeable. Pour la lemmatisation de Corneille, Molière et Racine, cela représente tout de même plus de 10 000 cas non résolus ! Les principales difficultés résident dans les mots les plus usuels ("suis", "tout", "même"…) et dans le caractère idiomatique de la langue : toute "règle" aura des exceptions…
Il serait donc nécessaire de réduire encore, si possible, ce "résidu". En effet, toute intervention manuelle est source d'erreurs ou de fluctuation dans les solutions retenues. Aussi, pour les cas non-résolus, l'algorithme offre des solutions pertinentes et limite au maximum les possibilités d'erreur.
En effet, c'est le principe le plus important : **la lemmatisation doit être sans erreur** (du moins par rapport aux conventions retenues qui, elles-mêmes, devraient être entièrement explicites). Les étiquetages fantaisistes et lacunaires sont contre-productifs car ils engendrent le scepticisme sur l'ensemble des informations contenues dans la base (Hug, 2002).

En conclusion, je voudrais souligner quelques points :
— puisque je m'adresse à un public composé de philologues, je voudrais redire que la (bonne) lemmatisation ne touche pas au texte original mais qu'elle permet de le consulter enfin d'une manière rationnelle. Par exemple, à la place de la concordance rudimentaire que vous avez vue, un outil puissant fournirait, notamment, la possibilité d'appeler en "mode image" la page concernée de l'ouvrage original…
— la nomenclature et les conventions utilisées doivent être parfaitement transparentes. Toute base étiquetée doit être accompagnée de ces informations (voir par exemple, le British National Corpus : Burnard 1995). C'est ce que nous avons modestement tenté de faire, en 1990, quand nous avons mis dans le domaine public les discours de de Gaulle et de Mitterrand lemmatisés ;
— notre nomenclature pourra être jugée trop "rustique", notamment en ce qui concerne les modes et les temps verbaux ou certains mots multifonctionnels (comme "que" par exemple qui est ramené à pronom ou conjonction). Nous répondrons d'une part, que l'étiquette n'est pas fermée — on peut y ajouter toutes les informations grammaticales ou sémantiques que l'on souhaite — et d'autre part, que toute complexification introduit des risques d'erreur supplémentaires et nécessite à chaque fois de réécrire tous les programmes permettant d'interroger cette base et d'effectuer des calculs comme ceux que je vous ai présentés…
— naturellement, mon logiciel est une sorte de "démonstrateur" — comme on dit dans l'industrie pour désigner l'objet qui précède le prototype — et si, un jour, nous pouvons disposer d'un véritable "lemmatiseur" du français, ce sera nécessairement le résultat d'un considérable travail d'équipe réunissant plusieurs laboratoires. Mais une telle équipe peut-elle être formée en France ? L'expérience des 20 dernières années — pendant lesquelles j'ai tenté de nouer des collaborations avec les différents laboratoires travaillant sur cette question — incline plutôt au pessimisme. La quasi-totalité des spécialistes, qui travaillent sur ces questions, adhérent encore au dogme selon lequel la "forme graphique" est l'élément nécessaire et suffisant pour la constitution et l'analyse des bases de données textuelles. Cependant certains commencent à évoluer sur ce point (Brunet 2002), ce qui peut donner un peu d'espoir.
Enfin, puisque C. Bernet a évoqué hier devant vous l'histoire du TLF et de Frantext, permettez-moi de terminer avec un souvenir personnel. En 1965 ou 1966, jeune étudiant à Nancy, j'ai eu la chance d'assister à une conférence de P. Imbs présentant le "Trésor de la langue française". Il voulait convaincre son public, composé majoritairement de juristes, et leur disait que sa base leur permettrait de découvrir l'image du droit ou du juge dans la littérature, ce qui impliquait, disait-il, qu'on puisse retrouver le substantif "droit" à l'exclusion de l'adjectif ou de l'adverbe ou encore, le substantif "juge" à

l'exclusion du verbe homographe. Car indubitablement, P. Imbs pensait, à l'époque, que le TLF serait lemmatisé. La tâche était sans doute plus difficile qu'il l'imaginait, mais est-elle vraiment impossible ?

PS : un forum de discussion sur la lemmatisation est ouvert par la revue électronique *Lexicometrica* :
http://www.cavi.univ-paris3.fr/lexicometrica
Par ce site, on peut également accéder aux actes des JADT 2000 et 2002.

Bibliographie

Barthes R., 1966, *Critique et vérité*, Paris, Le Seuil.

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

Bernet C., 1998, "Les mots placés à la rime dans le théâtre de Racine", in Belin C. et al, *Racine poète*, Poitiers, La Licorne, p 187-202.

Brunet E., 2002,"Le lemme comme on l'aime", in Morin A. et Sébillot P. (dir), *6e Journées d'analyse des données textuelles*, Rennes, IRISA, 2002, 1, p 221-232.

Burnard L., 1995, *Users Reference Guide for the British National Corpus*, Oxford, Oxford University Computing Services.

CONSEIL INTERNATIONAL DE LA LANGUE FRANÇAISE, 1988, *Pour l'harmonisation orthographique des dictionnaires*, Paris, CLIF.

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

Habert B., Nazarenko A., Salem A., 1997, *Les linguistiques de corpus*, Paris, A. Colin.

Hatzfeld A., Darmesteter A., Thomas A., 1898, Dictionnaire général de la langue française du commencement du XVIIe siècle jusqu'à nos jours, Paris, Delagrave.

Hubert P., Labbé D., 1994, "Vocabulary Richness", *Communication au congrès de l'ALLC-ACH*, Paris, La Sorbonne (reproduit dans *Lexicometrica*, 0, 1997).

Hubert P., Labbé D., 1995, "La structure du vocabulaire du général de Gaulle" in Bolasco S. et Al, *IIIe Giornate internazionali di analisi statistica dei dati testuali*, Rome, CISU, II, p 165-176.

Hubert P., Labbé C., 2002, "Segmentation automatique des corpus", in Morin A. et Sébillot P. (dir), *6e Journées d'analyse des données textuelles*, Rennes, IRISA, 1, p 359-370.

Hug M., 2002, "Désambigüisation automatique d'homographes verbe/nom", in Morin A. et Sébillot P. (dir), *6e Journées d'analyse des données textuelles*, Rennes, IRISA, 1, p 371-379.

Juilland A., Brodin D., Davidovitch C., 1970, *Frequency Dictionnary of French Words*, La Haye, Mouton,.

Kylander B.-M., 1995, *Le vocabulaire de Molière*, Goteborg, Acta Universitatis, Gothoburgensis.

Labbé D., 1990, *Le vocabulaire de F. Mitterrand*, Paris, Presses de la FNSP.

Labbé D., 1990, Normes de saisie et de dépouillement des textes politiques, Grenoble, Cahiers du CERAT.

Lafon P., 1984, *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.

Leselbaum J. et Labbé D., 2002, "Lexicographie assistée par ordinateur", in Morin A. et Sébillot P. (dir), *6e Journées d'analyse des données textuelles*, Rennes, IRISA, 2, p 447-456.

Monière D. et Labbé D., 2002, "Essai de stylistique quantitative", in Morin A. et Sébillot P. (dir), *6e Journées d'analyse des données textuelles*, Rennes, IRISA, 2002, 2, p 561-569.

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

Muller C., 1977, *Principes et méthodes de statistique lexicale*, Paris, Hachette

Pibarot A., Labbé D., 1998, "Les syntagmes répétés dans l'analyse des commentaires libres", in Mellet Sylvie (ed), *4e Journées d'analyse des données textuelles*, Nice, 1998, p 507-516.

Silberztein M. 1993, *Dictionnaires électroniques et analyse automatique des textes : le système INTEX*, Paris, Masson.

Silberztein M. 1995,"Dictionnaires électroniques et comptage des mots", in Bolasco S. et al, *IIIe Giornate internazionali di analisi statistica dei dati testuali*, Rome, CISU, I, p 93-101.

Annexe 1

Le calcul des univers lexicaux.

Chaque locuteur, lorsqu'il utilise un mot polysémique, lui imprime un sens spécifique que l'on peut retrouver en étudiant les liens d'association, ou de répulsion, que ce mot entretient avec le reste du vocabulaire. Le total de ces associations positives ou négatives, constitue l'**univers** du mot. Cette dénomination est préférable à "champ lexical" qui est employé par la lexicologie pour désigner un sous-ensemble du lexique de la langue. On a également écarté "monde lexical", utilisé pour désigner un imaginaire psychologique plus qu'une réalité langagière et qui renvoie à un calcul assez différent de celui présenté ci-dessous.

Remarque : la graphie des mots est préalablement normalisée et le texte est lemmatisé. La recherche des univers lexicaux se déroule uniquement sur les vocables.

Soit un corpus composé de N occurrences ("mots").

U est l'ensemble des phrases qui contiennent le vocable étudié.

Le nombre total d'occurrences contenues dans U sera N_u.

Pour un vocable quelconque, employé F fois dans N, on en attend dans U :

$$E_u = F \ x \ \frac{N_u}{N}$$

Si la fréquence constatée (F_u) est différente de la fréquence attendue (Eu), quand peut-on dire qu'il existe une relation d'opposition ou d'association entre ce vocable et l'univers considéré ?

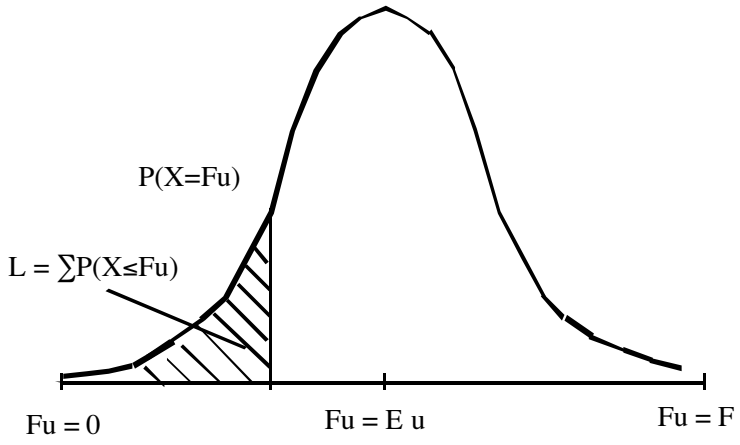
La relation existant entre le vocable et U peut se mesurer à la probabilité de l'événement observé F_u par rapport à l'événement attendu (E_u). La variable aléatoire X suit une loi hypergéométrique de paramètres N, F et F.

$$(1) \quad P(X = F_u) = \frac{\begin{bmatrix} F \\ F_u \end{bmatrix} \begin{bmatrix} N - F \\ N_u - F_u \end{bmatrix}}{\begin{bmatrix} N \\ N_u \end{bmatrix}}$$

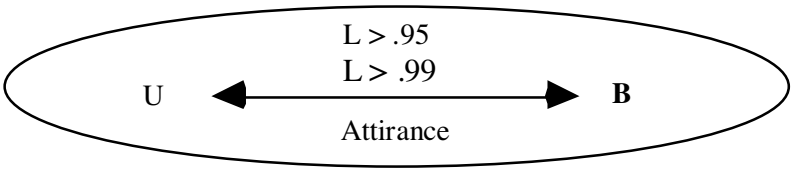
F_u peut varier entre O — aucune occurrence du vocable dans U — et F : toutes les occurrences du vocable sont observées dans U ($0 \leq F_u \leq F$).

En développant (1), on constate que le calcul n'a de sens que si $F < N_u$ et $F < (N - N_u)$. La borne supérieure signifie que le calcul est inutile en cas de corpus monothématique (la plupart des phrases appartiennent à U). La borne inférieure signifie que le calcul doit porter sur des grands univers, ou que, pour des petits univers, on doit exclure les vocables les plus fréquents (les "mots-outils").

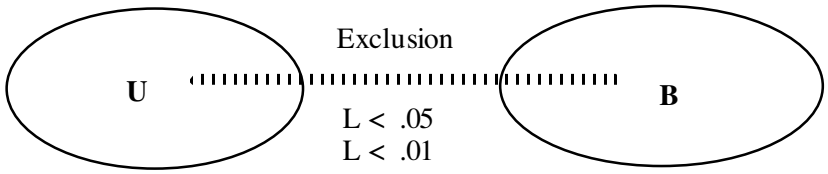
A condition que N, F et U soit suffisamment grands, les valeurs de X se distribueront selon la fameuse "courbe en cloche", avec un mode pour $F_u = E_u$ (graphique ci-dessous).



La liaison (L) entre l'univers et le vocable étudié sera la probabilité pour que la fréquence observée soit **au moins** égale à F_u. C'est la surface comprise sous la courbe.
Un vocable, noté B dans le schéma ci-dessous, sera significativement suremployé dans U lorsque L aura une valeur supérieure à .95 ou à.99 suivant que l'on choisira un risque d'erreur de 5% ou de 1%. L'attirance entre l'univers U et le vocable sera d'autant plus forte que L sera plus proche de 1.



A l'inverse, le vocable sera significativement sous-employé dans U lorsque L aura une valeur inférieure à .05 ou .01 suivant le seuil retenu. Nous pourrons donc affirmer que l'absence relative de B quand A est utilisé ne peut être due au hasard. La force d'exclusion mutuelle sera mesurée par L.



En pratique, le calcul de L se fera par cumul des valeurs de P de la manière suivante :

(3) $avec\ \ F_u > E_u\ :\ L\ =\ P(X \geq F_u) =\ 1\ -\ \sum_{k=F_u}^{Min(F,N_u)} P(X = k)$

(4) $avec\ \ F_u < E_u\ :\ L\ = P(X \geq F_u) =\ \sum_{k=0}^{F_u} P(X = k)$

Deux remarques doivent être faites :

En premier lieu, l'espérance mathématique sera être un entier tel que (Lafon 1984) :

(5) $\frac{(F+1)\ (N_u+1)}{N+2}\ -\ 1 \leq E_u \leq\ \frac{(F+1)\ (N_u+1)}{N+2}$

Les résultats sont donc affectés d'une certaine incertitude dont il faut tenir compte dans l'interprétation. Cela oblige à ne pas descendre dans les trop basses fréquences.
En second lieu, le calcul doit porter sur les mots dont la fréquence totale (F) est telle que leur absence dans U soit significative. C'est le "seuil d'absence significative (Salem 1987) :
avec $F_u = 0$: L < .05 ou L <.01
En effet, il y a peu d'intérêt à effectuer des calculs sur des mots peu fréquents ou sur des univers trop petits. Plus fondamentalement, on rappellera que la distribution normale n'est probablement pas un modèle adéquat pour l'étude de la probabilité de survenue d'un événement très rare au sein d'une population nombreuse et diverse.

Bibliographie :
HUBERT P., LABBE D., 1995, "La structure du vocabulaire du général de Gaulle" in BOLASCO Sergio et AL, *IIIe Giornate internazionali di analisi statistica dei dati testuali*, Rome, CISU, II, p 165-176.
LABBE C., LABBE D., 1994 et 2001, "Que mesure la spécificité du vocabulaire ?", Grenoble, CERAT. Repris dans : Lexicometrica, 3, 2001.
LAFON P., 1984, *Dépouillements et statistiques en lexicométrie*, Genève-Paris, Slatkine-Champion.
SALEM A., 1987, *Pratique des segments répétés*, Paris, Klincksieck.

Annexe II
La lemmatisation de "TOUT"…

Le tableau ci-dessous résume les classifications possibles de "tout"

	Déterminant	Pronom	Adverbe	Nom
A. Tout	x	x	x	x
B. Toute	x	x	x	
C. Toutes	x	x	x	
D. Tous	x	x		

Soit douze cas que l'on peut pratiquement tous résoudre grâce aux règles suivantes :

1. Tout est déterminant quand il est employé dans un groupe nominal et qu'il est accordé aux autres éléments du groupe :

— seul devant un *substantif* , tout est un déterminant normalement "invariable" (il se comporte comme un adverbe). Exemples : "le **tout** Paris", Bernet a analysé **tout** Phèdre, **tout** Les Plaideurs et **tout** Racine. Mais pour des raisons euphoniques : **toute** La Thébaïde, **toutes** les Fleurs du mal (cf. plus bas le même problème avec l'adverbe) … NB la même position devant un adjectif désigne l'adverbe (cf. plus bas) ;
— devant un déterminant autre que l'adjectif indéfini : article + substantif : Muller a lemmatisé **tout** le théâtre de Corneille, **toute** l'oeuvre, **toutes** les tragédies, **tous** les actes… Le possessif : **tous** nos actes, **toute** votre volonté, **toutes** vos volontés… ; le démonstratif : **toutes** ces femmes, **tout** ce mélange.
— se rattache à cette solution, tout employé devant un numéral (avec l'accord) : "**tous** (ou toutes) deux, trois, tout un…
— devant les pronoms démonstratif : **tous** ceux, **toutes** celles, **tout** ça
— derrière une préposition et devant un substantif : à **tout** bout de champ, de **tout** coeur, en **tout** cas… à **toute** force, de **toutes** parts, de **tous** côtés,

2. Tout est pronom lorsqu'il est employé seul ou associé à un groupe verbal.

— Pronom sujet : **Tout** lui convient, **Tous** (e,es) étaient présents(e,es), **tout** se vaut. La seule difficulté vient de tout précédé de "avoir" ou "être" : ils ont **tous** pris quelque chose, elle était **toute** à lui (cf. plus bas la discussion), elles sont **toutes** les mêmes…
— Pronom COD : Ils prennent **tout**, Ils ont **tout** pris
— Pronom COI : Ils sont prêts à **tout**, Ils en ont donné à **tous**(e,es), ils se servent de **tout**
— plus généralement à chaque fois qu'il est précédé d'une préposition sans être suivi d'un groupe nominal ou verbal : "prêt à **tout**" signifie : prêt à faire **tout**… On en aura une confirmation dans le parallélisme avec rien : parler de **tout** et de rien (pronoms tous les deux)
— Sont rattachées à cette solution : les locutions comme "après **tout**", "comme **tout**", "en **tout**" (qui sont analysées comme "préposition+pronom" soit "adv+pronom" bien qu'on puisse à rigueur y voir des substantifs (voir la note sur les homographies entre l'adjectif et le l'adverbe).

NB: dans les grammaires, ce pronom peut **tout** remplacer : un mot, un groupe de mot, une proposition et, en conséquence, se retrouver un peu n'importe où dans la phrase et être associé pratiquement à n'importe quel élément ou à aucun : "nous voulons **tout**" donne : "Ce que nous voulons ? Tout !" Cette mobilité du pronom, véritablement "à **tout** faire" est l'une des deux difficultés de la lemmatisation.

3. Tout est adverbe lorsqu'il est placé devant un adjectif ou employé dans une locution adverbiale ou prépositive.

— avec un adjectif ou un participe ayant valeur d'adjectif :
— un adjectif qualificatif : elle est **tout** heureuse, il est resté **tout** bête, **tout** seul, **tout** cassé, **tout** pantelant, etc.
— un adjectif indéfini : **tout** autre (différent) est le cas de…",
— un numéral : les **tout** premiers jours…
— devant un participe passé : nos prix sont **tout** compris (mais "sont **tous** fixés") ;

Bien qu'invariable l'adverbe est accordé à l'adjectif et au participe féminins pour des raisons "euphoniques" : elle est **toute** contente, elles sont **tout** heureuses et **toutes** confuses, les **toutes** premières fois (ce qui pose alors le problème de l'homographie avec le pronom, cf plus bas la discussion du problème) ;
— devant ou après un adverbe ou une préposition ("locution") : **tout** naturellement, **tout** devant, **tout** derrière (mais : "elle place **tout** devant elle")…
— locution prépositive introduisant un gérondif : **tout** en lisant ce travail, il pensait à… On y rattache : **tout** à coup, **tout** à fait, **tout** de même…, bien qu'on puisse à la rigueur y voir des pronoms indéfinis (cf plus bas la discussion à propos de "elle est **toute** à lui")

4. Tout est substantif quand il est précédé d'un déterminant ou d'un adjectif antéposé et suivi d'autre chose que d'un substantif ou d'un adjectif.
— précédé des déterminants "le" et "un" : le grand **tout** de l'univers, le **tout** est de … , le **tout** pour le **tout**, c'est un **tout**…
— se rattachent à ce cas tout précédé des formes contractes (du, au) qui sont analysés en "de+le" et "à+le". "Du **tout** au **tout**" est lu comme "de le **tout** à le **tout**" (substantif + substantif). Egalement : pas du **tout**. Ce qui entraîne un risque de téléscopage dans des formules comme "elles ne sont pas du **tout** heureuses" où c'est l'ensemble {adv + subst} qui joue le rôle de locution adverbiale.
NB : nous avons décidé que tout précédé d'une préposition et non suivi d'un groupe nominal est analysé comme un pronom indéfini : en tout est pronom, en **toutes** choses, déterminant (cf plus haut 2).

5. Discussion.
Voici les principaux problèmes auxquels s'est heurtée la programmation :

— l'adverbe invariable mais accordé au féminin pour l'euphonie rend impossible l'analyse de certaines phrases : "elles sont **toutes** contrites" peut signifier que pas une seule de ces femmes n'y échappe (pronom) ou que chacune d'entre elles est totalement contrite (adverbe). Que faire ? Le programme s'arrête et laisse l'opérateur se interpréter… En revanche, le problème ne se pose pas pour "tous" qui dans cette construction est sûrement un pronom : ils sont **tous** contrits (pronom) et ils sont **tout** contrits (adverbe).
— "elle était **toute** à lui". Dans cette phrase, "toute" pourrait être analysé comme un adjectif qualificatif — synonyme de "follement amoureuse de lui" — mais pas comme un adverbe (dans ce cas nous aurions : "**tout** à lui"). On ne vas pas créer une cinquième classe (l'adjectif) pour ce seul cas ! En fait, il s'agit bien d'un pronom. Se rattachent à cette même construction : "elle est **toute** à ses pensées".
— les locutions du genre : "**tout** à coup". Dans ces expressions, tout n'est pas précédé d'un déterminant ou d'une préposition, ni suivi d'un substantif : ce n'est donc pas un nom ni un déterminant. Il reste le pronom "indéfini" (cf plus haut : "nous voulons **tout**") ou l'adverbe. On ne peut les analyser comme des "locutions figées". Cette solution entraîne dans des difficultés sans fin car leur nombre est potentiellement infini. Il faut analyser ces locutions comme des "compositions libres" faites de plusieurs mots indépendants.

Les télescopages plus ou moins difficiles à maîtriser…
— à cause des locutions. Par exemple, dans "ils ne se sont pas du **tout** compris", tout est analysé comme "pas de le **tout**" (substantif) bien qu'il soit situé entre l'auxiliaire et le participe.
— "nous avons **tous** nos défauts", "nous avions **tous** nos avions à prendre" (à l'oral, le déterminant se prononce "tout" et le pronom : "tous"… Le problème se pose dans : "nous avons **tous** les défauts (de nos pères et leurs qualités aussi)" et dans "…**tous** les défauts (dont vous nous accusez)". Le pronom s'utilise normalement combiné avec une préposition : "nous avons **tous** des défauts qui nous sont propres", "nous avions **tous** de beaux avions". Une solution raisonnable pourrait donc considérer que, dans la phrase : "Nous avons **tous** nos bagages" — limitée à ces seuls mots — tout est déterminant mais pronom dans : "nous avons **tous** des bagages". A contrario, on peut aussi bien rencontrer : elles ont **toutes** leurs bagages (pronom) et elles ont **tous** leurs bagages (déterminant)…
— Le problème avec "être" est différent. "Nous sommes **tous** leurs obligés (pronom)", "nous sommes **tout** obligés (adv)", "nous sommes **tous** juifs (pronom)", "ils sont **tout** noirs (adverbe)", "il est **tout**

le temps débordé" (déterminant)". Ces femmes sont **toutes** noires peut aussi bien signifier qu'il n'y en a aucune qui ne soit pas noire ou que **toutes** sont entièrement noires… Le programme ne peut que s'arrêter et rendre la main à l'opérateur.

— Dans la pratique, le problème ne se pose pas avec les verbes admettant des COI. On emploie alors la préposition : "Ils demandent **tous** les emplois" (déterminant). Sinon on écrira : "Ils demandent **tous** des emplois, du pain et des jeux : **tous** des Romains !" Dans les oeuvres littéraires au moins, cela marche tout le temps et à tous les coups !

Finalement, le cas de "tout" a été assez long à programmer et à tester à cause des télescopages entre le pronom, le déterminant et l'adverbe, invariable mais accordé, qui se trouvent employés dans des constructions de phrases identiques ou très ressemblantes. En définitive, c'est environ 10% des "tous", "toute" et "toutes" qui restent non codés du fait de ces ambiguïtés.

Annexe III
La nomenclature finale

1. Verbe :
   11 forme fléchie
   12 forme au participe passé
   13 forme au participe présent
   14 forme à l'infinitif
2. Substantif :
   20 « nom propre » (mot à majuscule initiale)
   21 substantif masculin
   22 substantif féminin
3. Adjectif :
   30 Adjectif "pur"
   31 Participe dans un emploi "adjectivé"
5. Pronom :
   51 Personnel
   52 Relatifs, réfléchis, interrogatifs, possessifs, etc.
6. Adverbes
7. Déterminant:
   71 Articles (défini et indéfini)
   72 Numéraux et cardinaux
   73 Possessifs
   74 Démonstratifs
   75 Adjectifs indéfinis

81. Préposition
82. Conjonction
91, 92, 93. Locution, expression, interjection

Ponctuation :
- "p" : ponctuation mineure (interne à la phrase)
- "P" : ponctuation majeure (délimitant la phrase)