Etienne Brunet
UFR Lettres, 98 bd HERRIOT
06204 NICE

Qui lemmatise dilemme attise

Si les murs ont des oreilles et de la mémoire, il suffit de coller son oreille aux murs de cette salle pour entendre à l'avance ce que je vais dire. Car nous sommes dans la salle où a enseigné Charles Muller pendant des années (il y a quelques photos aux murs qui en portent témoignage). Et on ne peut guère parler de lemmatisation ni, plus généralement, de statistique linguistique, sans se référer au maître de la discipline qui fut aussi le maître de ces lieux. Comme il est présentement dans cette salle, il serait naturel que je lui cède la parole à l'improviste, en lui rendant d'ailleurs la monnaie de sa pièce. Car j'ai souvenance d'une situation embarrassante où le maître m'avait placé, avec sa malice coutumière: invité par lui à un déjeuner bien arrosé, j'étais passé de la salle à manger à la salle de cours, sous prétexte de visiter ce célèbre musée linguistique. Or la salle était pleine d'étudiants à qui le maître proposa que je fisse un cours improvisé, à sa place, et l'ordre était immédiatement exécutoire.

- I - La querelle formalistes/lemmatiseurs

1 - Muller est en effet bien mieux qualifié que moi pour parler de cette vieille querelle qui a longtemps opposé les partisans et les adversaires de la lemmatisation. Ceux à qui cette querelle est étrangère et qui ignorent le sens même du mot, trouveront peu de secours dans les dictionnaires, qui n'ont pas encore accepté ce terme technique. Qu'ils sachent qu'on désigne par là l'opération de regroupement qui rassemble les formes différentes appartenant au même vocable - ce que font précisément les auteurs de dictionnaires quand ils établissent leur nomenclature. Ce classement simplifie la recherche dans le dictionnaire et fait gagner de la place et du temps, sans perdre de l'information, car la définition d'une vedette de regroupement vaut pour toutes les formes qui se rattachent au mot défini. Et les sèmes qu'on y rencontre ne dépendent guère des accidents que le genre, le nombre, le mode, le temps ou la personne peuvent imposer à la désinence. Pareil avantage est attendu dans les recherches documentaires ou lexicométriques. Quand les formes verbales sont rassemblées derrière un même lemme, il est plus aisé de les rechercher dans le texte, et de tirer des conclusions de leurs fréquences cumulées. Ces arguments et d'autres plus subtils sont exposés dans la préface que Charles Muller a donnée à l'ouvrage de Pierre Lafon *Dépouillements et statistiques en lexicométrie (Travaux de linguistique quantitative*, n° 24, Slatkine-Champion*)*.

2 - Les arguments opposés se trouvent aussi dans cette préface, et plus encore dans le reste de l'ouvrage. On peut sans doute faire peu de cas des variations de sens qui peuvent parfois accompagner le nombre et introduire des nuances entre *peuple* et *peuples*, *histoire* et *histoires*, *lumière* et *lumières*. Car la polysémie n'est pas propre au nombre et on la retrouve au singulier comme au pluriel, même dans les exemples qu'on vient de citer. Mais les recherches linguistiques, stylistiques ou lexicométriques ne se réduisent pas à la thématique. Si l'on s'intéresse à la syntaxe ou à plus forte raison à la morphologie, il faut avoir accès à la forme même du mot. L'étude des temps et des modes, par exemple, n'est possible que si les formes ne sont pas confondues dans la même entrée. Les index et les bases hypertextuelles doivent pouvoir répondre à la plus grande variété possible des demandes et ne pas limiter a priori le champ des recherches.

3 - Les premiers travaux de lexicométrie se sont principalement intéressés aux mots pleins et aux caractères sémantiques et thématiques des premiers textes dépouillés, dont le traitement était en grande partie manuel. La lemmatisation faisait alors partie des travaux préparatoires, parmi d'autres tâches plus ingrates encore. Et l'on ne songeait pas à s'y dérober, puisqu'on reprenait là la tradition des *index nominum* (ou *rerum*) chers à l'édition érudite. Vint l'époque où la saisie cessa d'être manuelle et où sans grand effort le texte devint disponible sur un support informatique, avec une fidélité électronique qui respectait l'ambivalence des mots. L'effort de désambiguïsation parut alors démesuré eu égard aux bénéfices escomptés. D'autant que les machines et les logiciels avaient évolué et permettaient d'offrir en même temps la forme et le vocable, et donc de cumuler les avantages des deux systèmes en neutralisant leurs inconvénients. On a proposé alors des outils de conjugaison ou de dérivation, fondés sur les modèles du *Bescherelle*, ou même sur le relevé exhaustif des formes possibles pour un même verbe (cette dernière option est le choix de *Frantext*). Le regroupement des adjectifs et des substantifs, même irréguliers, pose moins de problèmes. Plus simplement encore des caches neutralisant la finale des mots ont servi à isoler des racines ou radicaux. Beaucoup de logiciels (par exemple *Tropes* ou *Alceste*) proposent ces fonctions de regroupement qu'un filtrage ultérieur peut affiner. On trouvera ci-dessous les propositions de notre logiciel HYPERBASE qui exploitent de telles possibilités.

Figure 1. Diverses fonctions de regroupement des mots

4 - Seulement le regroupement des formes se fait alors de façon grossière en obligeant les homographes à se ranger en bloc sous une bannière unique ou à choisir le camp opposé. Une telle "lemmatisation" s'exerce malencontreusement sur le dictionnaire, et non sur l'examen du contexte. Ainsi toutes les formes de *marche* ou de *marches* sont rattachées ensemble au verbe ou au substantif, sauf à se référer à des pourcentages extraits du dictionnaire des homographes (publié par le TLF en 1971), et nécessairement approximatifs (issues d'échantillons partiels, les proportions obtenues, trop fragiles, ne peuvent guère s'appliquer à des corpus extérieurs)[1]. On a même renoncé de plus en plus à faire ce commencement de toilettage. Et en offrant des corpus de plus en plus vastes, on a moins de scrupules à se dérober à la désambiguïsation manuelle des homographes, tâche humaine devenue inhumaine par la taille des textes à traiter et, parfois aussi, par la complexité des règles à respecter, lorsque précisément l'orthographe n'a pas de règle constante et que les graphies se multiplient à l'infini. Les deux opérations associées de séparation (ou désambiguïsation) et de regroupement exigent une sagacité particulière dans le cas des textes du XVI^{e} siècle, et plus encore au Moyen Âge, quand l'orthographe est flottante et l'accentuation fantaisiste. Même chez Rousseau, qui offre à cet égard une rigueur et une constance tout à fait remarquable en son temps, nous avons rencontré un nombre considérable de variantes orthographiques. Ainsi son *élève* nous est-il présenté sous trois habits dans l'*Émile*: *eléve* (145 occurrences), *eleve* (l8) et *éleve* (l3), sans compter 3 occurrences qui lui attribuent la majuscule (*Eleve*).

5 - Les meilleurs travaux sont bien sûr ceux qui s'appuient sur des données rigoureusement lemmatisées, comme ceux de Dominique Labbé sur le vocabulaire de Mitterand. Ce sont aussi les plus rares. En outre ils offrent peu de prise à la comparaison, non seulement du fait de leur rareté, mais aussi parce que les exigences en matière de lemmatisation sont variables selon les corpus et les chercheurs, en sorte qu'il est difficile de trouver un standard pour les corpus lemmatisés, chacun proposant ses propres règles. Dans cette tour de Babel qu'est le marché linguistique, les rencontres et les comparaisons sont plus faciles au niveau des planchers que des plafonds, c'est à dire quand les règles sont minimales. Le degré zéro et le niveau le plus bas correspondent à l'absence complète de toute manipulation préalable des données. Et les logiciels les plus répandus fréquentent de préférence cette foire des données brutes, que les scanners et Internet alimentent à foison. Les conditions sont remplies pour qu'on y trouve le pire: un tas informe de mots estropiés,

---

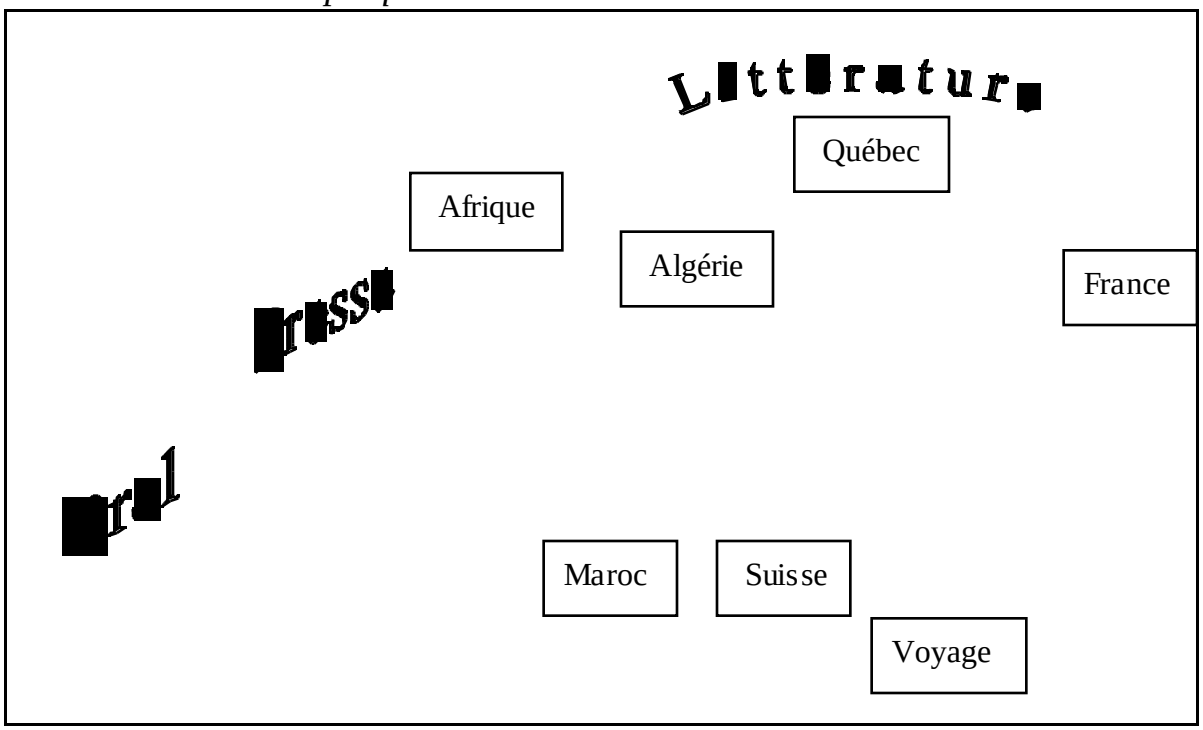
[1] C'est cette formule incertaine que nous avons adoptée, faute de mieux, dans nos monographies sur le vocabulaire de Proust, Zola et Hugo. Du moins s'agissait-il de grands corpus littéraires où la projection des données de Nancy pouvait trouver une justification partielle puisque ces données appartenaient à la même époque et au même usage littéraire de la langue.

douteux, ambigus. Mais l'abondance est là, et une certaine cohérence qui tient au refus de toute intervention où l'arbitraire pourrait s'immiscer. La statistique qui aime les grands nombres et ne répugne pas à l'impureté (son rôle est précisément de filtrer l'entropie) peut-elle se satisfaire de ces conditions? C'est ce que nous nous proposons d'examiner.

- II - Les formes brutes

1 - On partira d'un premier exemple qui rend compte d'un ensemble fort disparate de textes empruntés au monde francophone[2]. L'impureté est ici maximale puisqu'on mêle volontairement l'écrit et l'oral, la presse et la littérature, et les différents pays du globe qui emploient encore le français. Aucun filtrage ne s'est exercé sur les données sinon la nécessaire transposition des données orales. Il est pourtant possible d'en extraire des enseignements même d'ordre morphosyntaxique. Le recours au conjugueur donne par exemple une moisson de verbes dont le total, pour les 50 plus fréquents, atteint 420 000 occurrences, soit le dixième de l'étendue totale (graphique 1 ci-dessous). Certes il y a de l'ivraie dans le grain: par exemple le verbe *être* inclut hors de saison quelques mentions de l'*été*. Mais la décantation n'en est pas moins claire: le verbe abonde à l'oral (les 12 premiers sous-ensembles situés sur la gauche du graphique), et se fait rare dans la presse et les ouvrages techniques (les 17 qui suivent). Dans les textes littéraires son emploi est variable selon les pays, suivant que le style est plus ou moins populaire, et que la part du dialogue est plus ou moins importante.

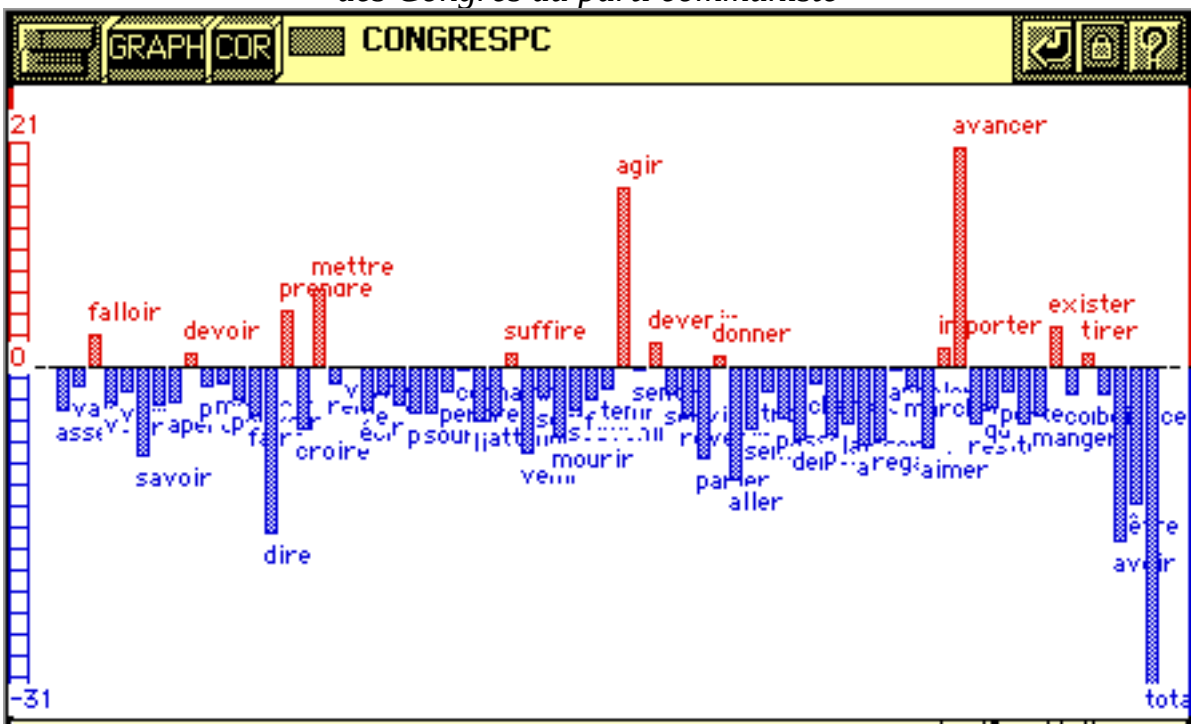
Graphique 1. Les verbes dans la base FRANCIL



[2] Il s'agit d'une recherche soutenue par l'AUPELF dans le cadre du projet FRANCIL.

2 - Si l'on met en gros plan un des sous-ensembles du corpus, celui des résolutions des Congrès du parti communiste, on y mesure la répugnance à l'emploi des verbes, et le recours au substantif, plus apte à exprimer les concepts dont se nourrit la langue de bois. Encore remarquera-t-on le caractère volontariste des quelques verbes qui s'aventurent dans la zone excédentaire : *falloir, devoir suffire, importer, agir, avancer.* Ce dernier élément peut faire sourire quand on mesure le chemin parcouru pendant les 50 ans où le leitmotiv a été répété. On songe à ces héros d'opéra qui chantent "Marchons! Marchons!" en restant sur place.

Graphique 2. Les verbes dans les résolutions des Congrès du parti communiste



3 - Le recours à la procédure du masque pour isoler un radical, et par là une famille de mots, voire un thème, est parfois suffisant. Une fois que la liste est constituée à partir de l'initiale, on peut facilement l'expurger en ajoutant les mots oubliés et en chassant les intrus, par exemple en substituant l'*amant* à l'*aimant* dans une série constituée à partir de l'initiale *aim-*. Ainsi le radical *belg-* proposé au programme aboutit à la collecte de toutes les phrases où interviennent les *Belges* ou la *Belgique*, soit 200 extraits dans le même corpus. Examinés à leur tour par la fonction thématique du logiciel HYPERBASE, ces passages produisent un concentré lexical où s'agrègent les corrélats associés à la belgitude et où la question linguistique est au premier rang. Voir tableau 3.

Tableau 3. Environnement thématique du radical belg- *(ordre hiérarchique)*

Ecart	Corpus	Extrait	Mot	Ecart	Corpus	Extrait	Mot
218.70	90	96	BELGique	12.46	46	4	distinction
176.18	56	61	BELGes	12.34	1798	28	hein
161.99	50	53	BELGe	12.32	47	4	immigrés
43.91	105	21	expressions	12.29	1480	25	crois
37.40	1795	77	français	12.18	48	4	parlementaire
32.88	95	15	journalistes	12.03	145	7	médias
32.34	44	10	francophone	12.02	1028	20	France
28.60	36	8	deviez	11.92	50	4	comparaison
25.96	11	4	qualificatif	11.85	111	6	exprime
25.07	780	34	politiques	11.56	53	4	Bruxelles
23.61	66	9	francophones	10.75	229	8	pratique
23.20	84	10	expriment	10.28	1199	19	silence
19.86	29	5	flamands	10.06	257	8	dirais
19.67	19	4	nonante	9.98	261	8	propres
19.58	444	20	parlent	9.85	332	9	presse
19.06	45	6	Congo	9.67	74	4	socialistes
17.46	401	17	accent	9.66	21944	113	:
15.13	462	16	française	9.27	717	13	niveau
14.56	1212	26	parle	9.10	3795	34	donc
13.71	1727	30	mieux	8.83	399	9	endroit
13.40	40	4	allemande	8.75	89	4	vocabulaire
13.35	1299	25	ailleurs	8.53	200	6	Communauté
12.90	2485	35	hommes	8.32	355	8	sociale
12.67	312	11	entend	8.21	152	5	tendance
12.52	218	9	connaissez	8.07	220	6	régions

III - Les formes étiquetées

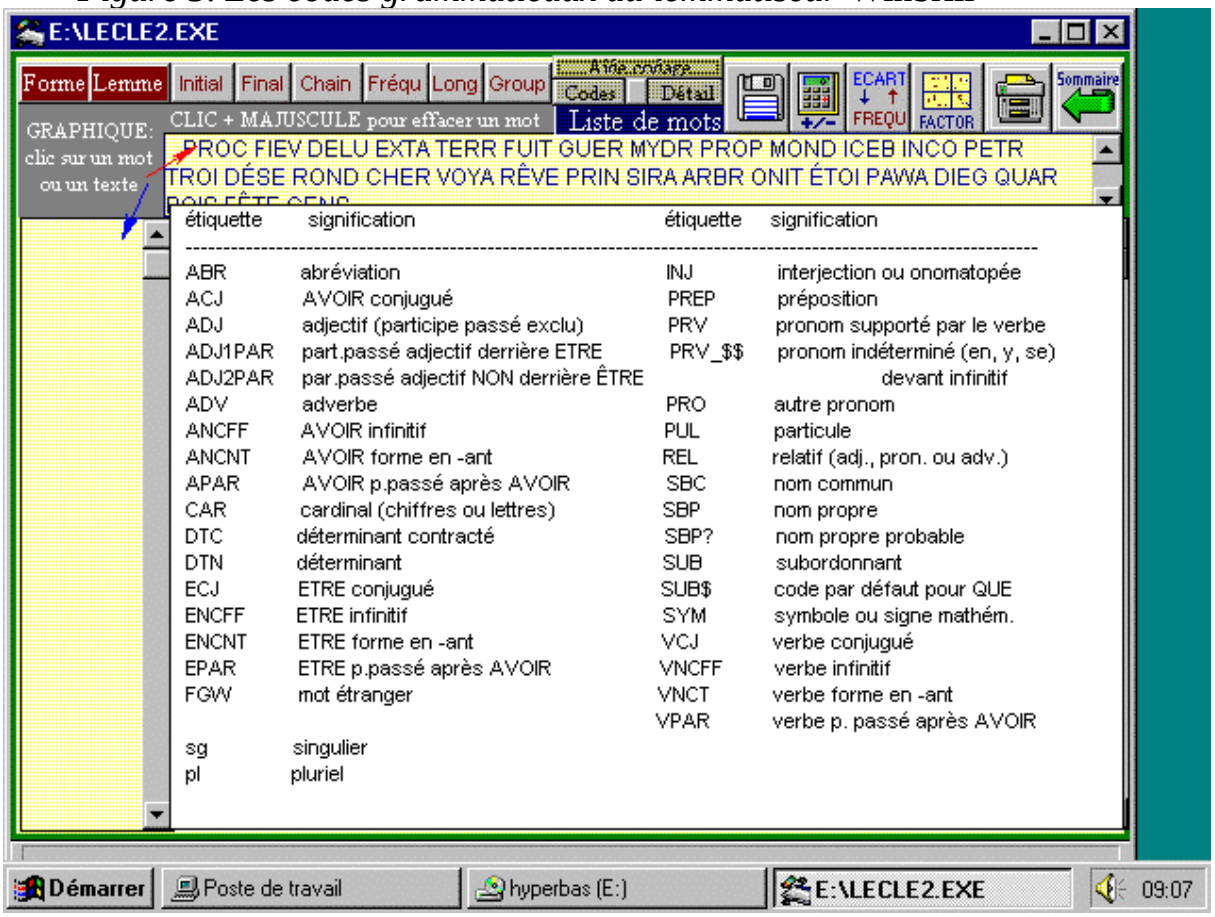
Avec le progrès des traitements informatiques, on peut tirer profit de certains automates qui se chargent non seulement de l'indexation des textes mais même de l'étiquetage des formes, lequel est à la base de la lemmatisation. Cette fois nous utiliserons une base homogène qui rend compte de l'œuvre de Le Clézio, soit trente textes publiés jusqu'ici (tableau 4).

Tableau 4. La composition du corpus Le Clézio

N°	TITRE	OCCURRENCES	VOCABLES	Prob P	Prob Q	ABREGE	CODE
1	Procès	92931	10914	.0428	.9572	PROCES	Pr
2	Fièvre	101667	10914	.0469	.9531	FIEVRE	Fi
3	Déluge	122297	13037	.0564	.9436	DELUGE	Dl
4	Extase	90562	9754	.0417	.9583	EXTASE	Ex
5	Terra	92085	9983	.0424	.9576	TERRA	Te
6	Fuites	104156	10788	.048	.952	FUITES	Fu
7	Guerre	110364	10358	.0509	.9491	GUERRE	Gu
8	Mydriase	9877	1974	.0046	.9954	MYDRIASE	My
9	Prophètes	8781	2039	.004	.996	PROPHETES	Ph
10	Mondo	99093	7160	.0457	.9543	MONDO	Mo
11	Icebergs	2716	820	.0013	.9987	ICEBERGS	Ic
12	Inconnu	124457	9127	.0574	.9426	INCONNU	In
13	Petra	10226	1974	.0047	.9953	PETRA	Pe
14	Trois	15356	2636	.0071	.9929	TROIS	Tr
15	Désert	150732	9186	.0695	.9305	DESERT	Ds
16	Ronde	82871	7004	.0382	.9618	RONDE	Ro
17	Chercheur	133568	10275	.0616	.9384	CHERCHEUR	Ch
18	Voyage	38117	5331	.0176	.9824	VOYAGE	Vo
19	Rêve	86146	10188	.0397	.9603	REVE	RÍ
20	Printemps	69721	6664	.0321	.9679	PRINTEMPS	Ps
21	Sirandanes	2479	814	.0011	.9989	SIRANDANES	Si
22	Arbres	2976	697	.0014	.9986	ARBRES	Ar
23	Onitsha	75326	7847	.0347	.9653	ONITSHA	On
24	Etoile	117930	8536	.0543	.9457	ETOILE	Et
25	Pawana	10594	2004	.0049	.9951	PAWANA	Pa
26	Diego	72657	9725	.0335	.9665	DIEGO	Di
27	Quarantaine	164284	12543	.0757	.9243	QUARANTAIN	Qu
28	Poisson	85290	8262	.0393	.9607	POISSON	Po
29	Fête	70961	9883	.0327	.9673	FETE	FÍ
30	Gens	21717	4188	.01	.99	GENS	Ge
	TOTAL	2169937	59218				

1 - En réalité la base *Le Clézio*[3] a été soumise successivement à deux traitements, avec ou sans étiquetage. Précisons que les étiquettes apposées aux formes du texte se réduisent à des codes grammaticaux dont le détail est reproduit ci-dessous (figure 5). On sera peut-être surpris qu'ils correspondent d'assez loin aux catégories traditionnelles et c'est pourquoi nous avons cru bon de les regrouper pour faciliter la lisibilité. Le logiciel utilisé pour l'étiquetage porte le nom de *Winbrill* : il s'agit d'un produit étranger adapté au français par Josette Lecomte et Gilles Souvay, chercheurs à l'INaLF.

Figure 5. Les codes grammaticaux du lemmatiseur Winbrill

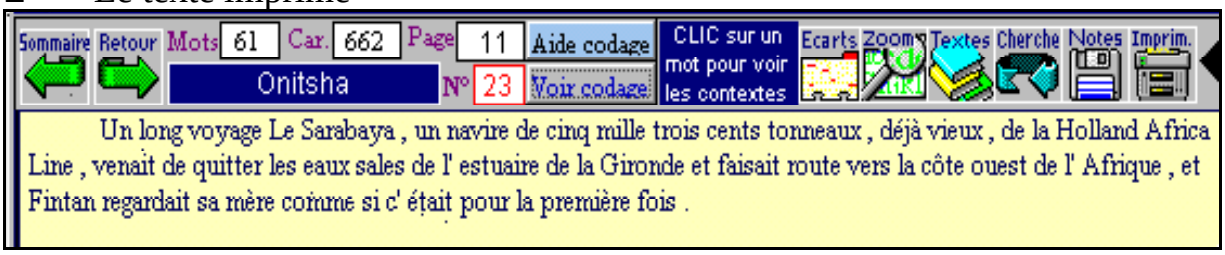


L'opération d'étiquetage est entièrement automatique, ce qui est tout à la fois la force et le handicap de *Winbrill*. Comme il n'y a pas de choix subjectif, les comparaisons sont possibles d'un corpus à l'autre, le traitement étant identique. Mais les erreurs d'analyse sont nombreuses. Et la lemmatisation n'est pas complète, puisque manque la mention du lemme correspondant, ce qui est pourtant le plus facile à faire. Le texte ainsi traité devrait alternativement apparaître sous quatre présentations: 1- formes seules, 2- formes étiquetées, 3 - étiquettes seules (simplifiées), 4 - lemmes seuls.

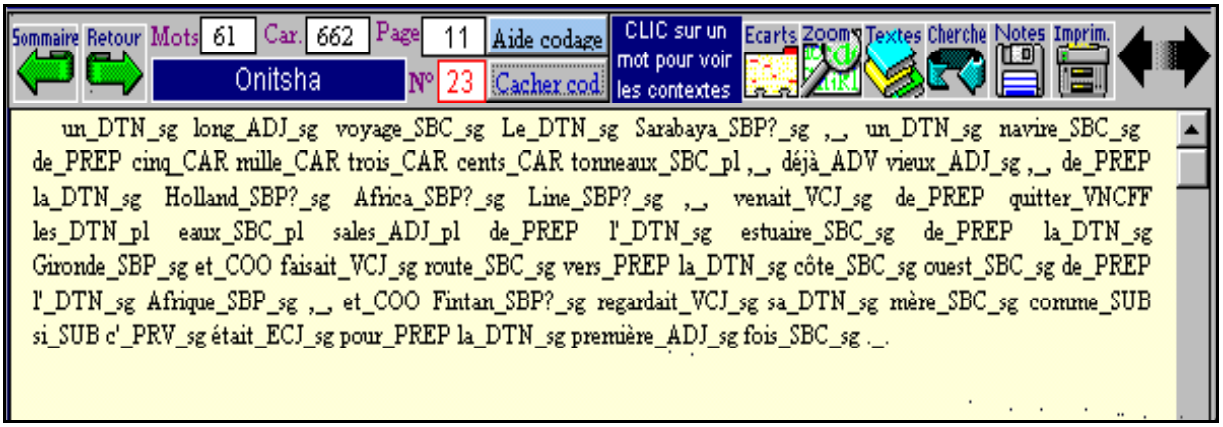
[3] Ce corpus est celui de Margareta Kastberg, qui achève une thèse sur Le Clézio.

Figure 6: Quatre présentations du texte dans une base lemmatisée

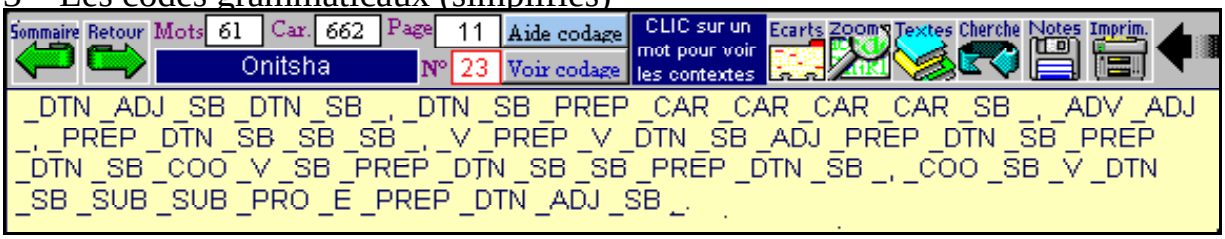
1 - Le texte imprimé



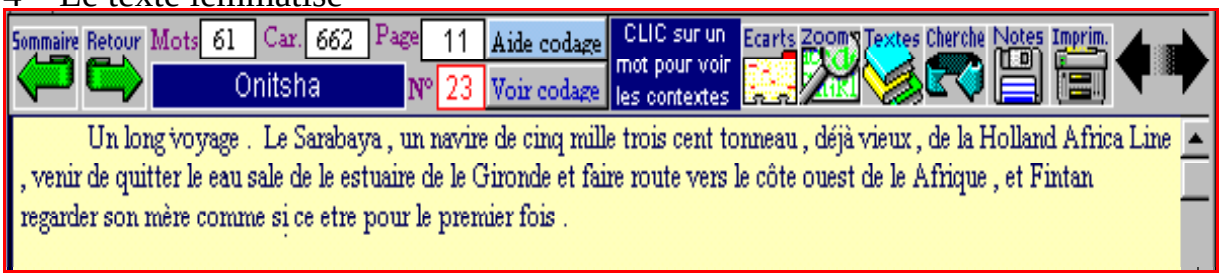
2 - Le texte étiqueté



3 – Les codes grammaticaux (simplifiés)



4 – Le texte lemmatisé



La dernière présentation (texte lemmatisé) est une extrapolation non encore autorisée par *Winbrill*. Mais d'autres logiciels comme *Cordial U* et *Sphinx Lexica* permettent d'accéder à ce dernier stade. Les codes associés aux formes sont d'un grand secours lorsqu'on veut isoler un homographe et par exemple ne plus confondre un *le* article et un *le* pronom. Ils sont aussi indispensables pour étudier les parties du discours et faire apparaître clairement comment le clan du verbe s'oppose à celui des catégories nominales. Ils autorisent même l'observation et la statistique des structures syntaxiques,

comme dans l'exemple ci-dessous qui relève le schéma *Prép+déterminant+adv+adj+subst.*, et en donne la répartition dans le corpus.

Figure 7. Les structures grammaticales (relevé)

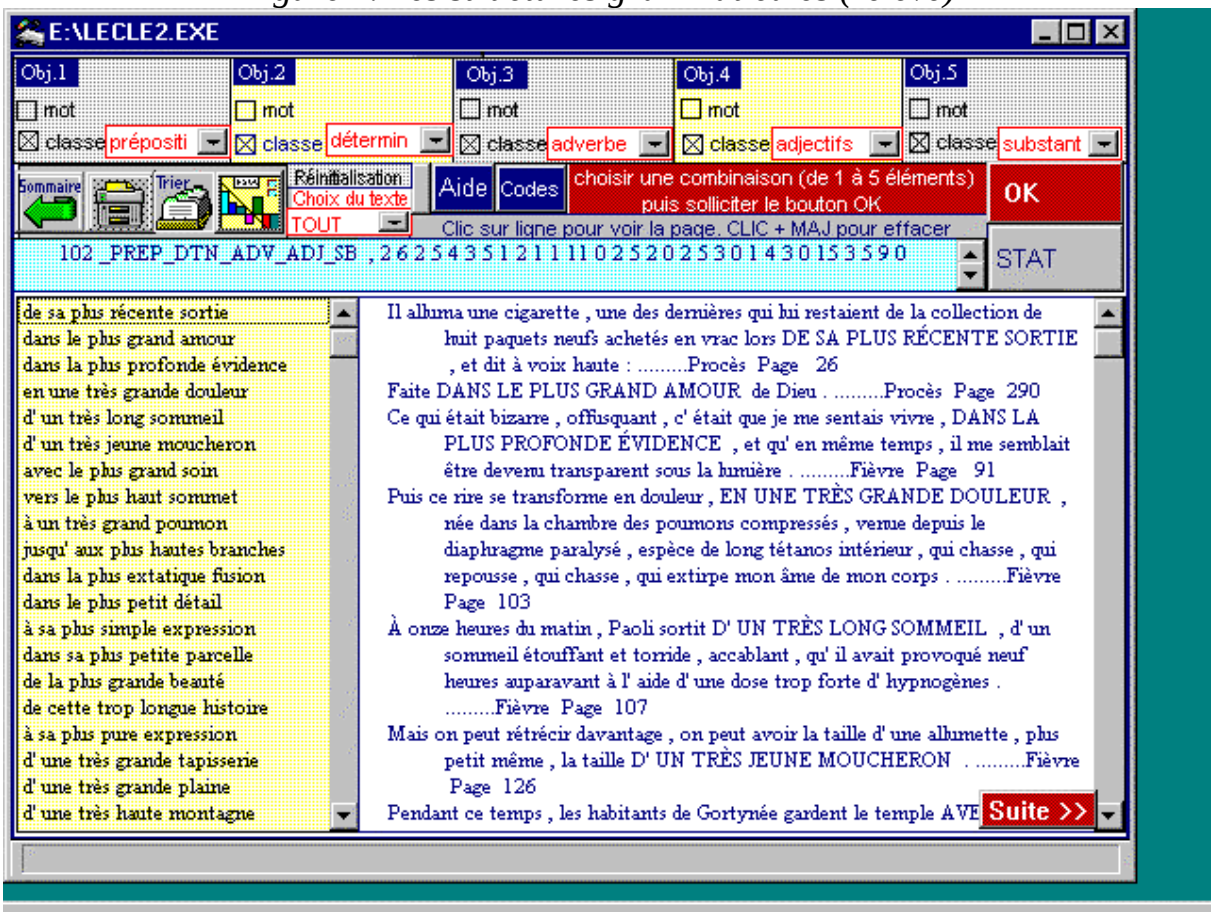
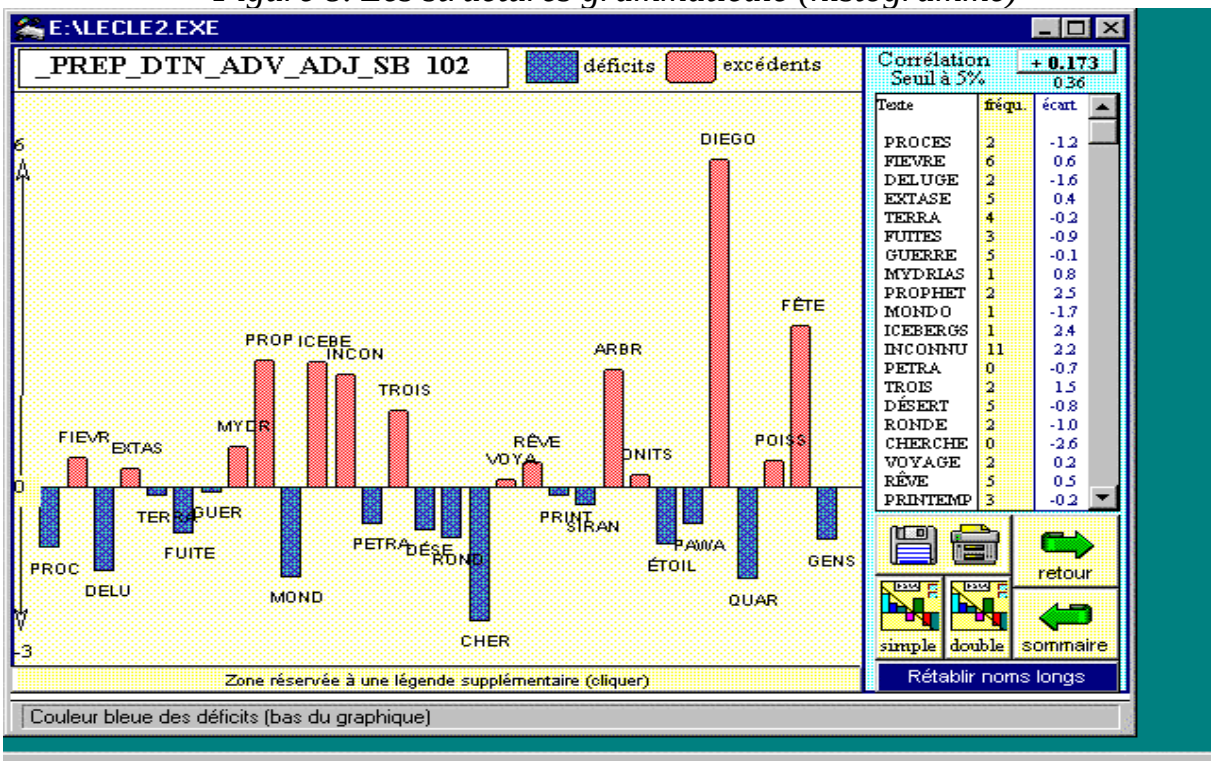
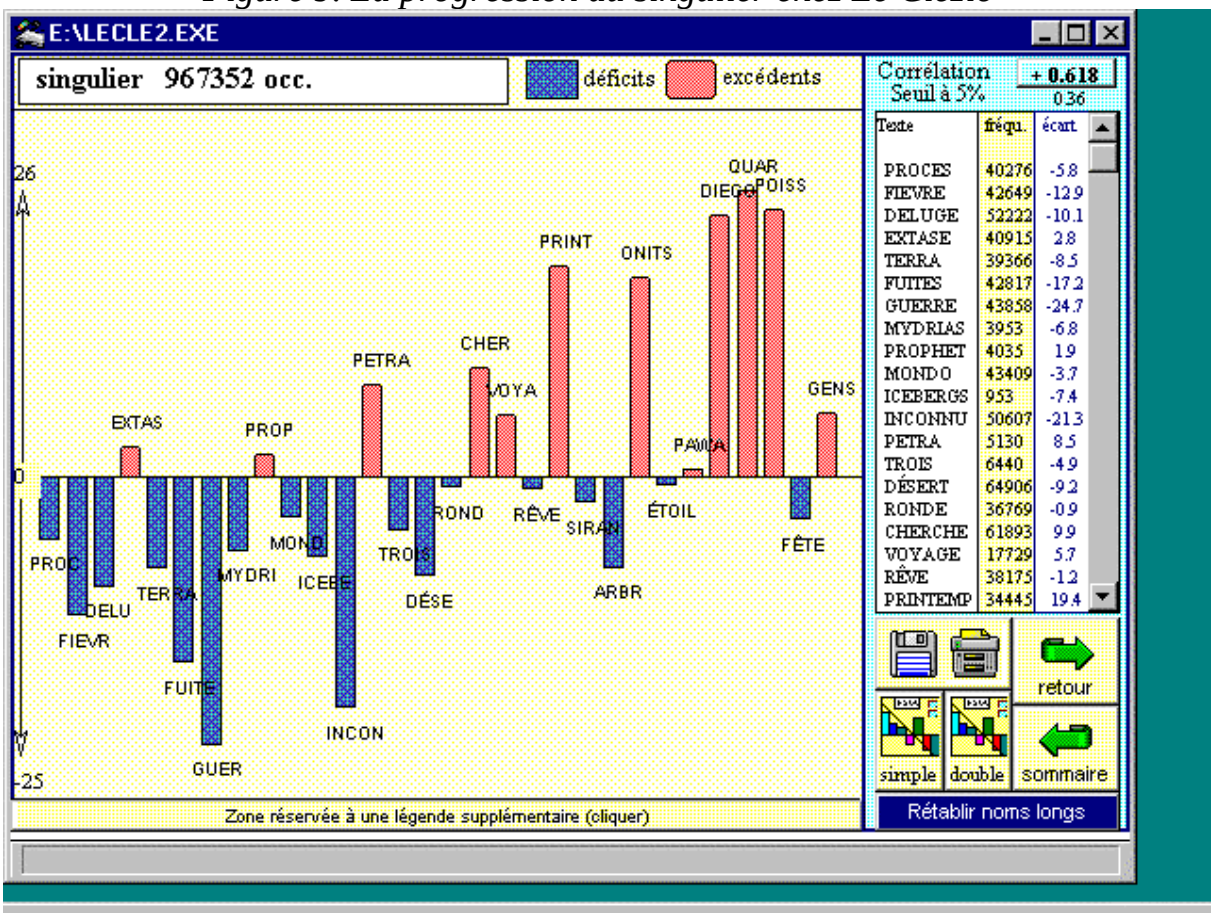


Figure 8. Les structures grammaticale (histogramme)



2 - Les étiquettes peuvent donner lieu à une combinatoire qui reflète les contraintes syntaxiques attachées à la langue ou les habitudes stylistiques propres à un écrivain. Elles se prêtent aussi à des opérations assez puissantes pour rendre compte du genre ou du nombre, là où on en trouve la trace parmi les parties du discours, qu'il s'agisse de verbes, de substantifs, d'adjectifs, de pronoms ou de déterminants. La courbe montante du singulier chez Le Clézio - et corrélativement la décrue du pluriel - ne laisse pas d'intriguer (figure 9). À quoi faut-il attribuer cette tendance, que nous avons observée chez la plupart des écrivains soumis à la même analyse? Est-ce une moindre attention portée à la multiplicité et à la variété des choses et des êtres? Est-ce le goût, croissant avec l'âge et l'expérience, de la généralisation et de l'abstraction? Ce n'est pas le lieu pour en discuter. Bornons-nous à entrouvrir cette porte parmi beaucoup d'autres auxquelles donne accès l'étiquetage des textes. Nous avons trop longtemps piétiné devant ces questions insolubles pour ne pas nous réjouir des possibilités offertes par le codage grammatical et du renouveau de la problématique qu'il entraîne.

Figure 9. La progression du singulier chez Le Clézio



3 - La question se pose néanmoins de l'avantage d'une base étiquetée sur celle qui ne l'est pas, là où le traitement est commun et la comparaison possible. Considérons l'ensemble des mots dans les deux bases Le Clézio dont l'une a 59218 formes étiquetées et l'autre 49773 formes brutes. Qu'on ne s'étonne pas de la différence dans l'effectif observé, l'étiquetage produisant le dégroupement des formes homographes et l'accroissement du vocabulaire. Dans les deux cas on se propose de calculer la distance qui sépare chaque texte du corpus de tous les autres. Cette distance est mesurée par le rapport entre les formes que deux textes ont en commun et celles qui sont propres à chacun. Pour annuler la perturbation que pourrait provoquer la différence d'étendue, le calcul cumule les deux rapports symétriques (du texte A vers B et du texte B vers A) selon la formule:

$d = ((a-ab)/a) + ((b-ab)/b)$

où *ab* désigne la partie commune aux deux vocabulaires *a* et *b* , *a-ab* et *b-ab* recouvrant les parties privatives. Le résultat est un tableau carré des distances qu'on soumet à l'analyse factorielle pour obtenir une sorte de carte géographique où les textes s'assemblent quand ils partagent les mêmes mots et sans doute aussi les mêmes thèmes. La base étiquetée de Le Clézio donne lieu à la représentation graphique de la figure 10.

Le premier axe qui sépare la partie gauche de la partie droite correspond au clivage du genre littéraire: à droite prennent place les romans ou recueils de contes où prime le récit. C'est la veine le plus souvent exploitée par l'auteur et l'aspect le plus connu de son œuvre. À gauche se situent des textes plus descriptifs, où s'exprime la curiosité ethnographique de Le Clézio et en particulier son intérêt pour le peuple indien et le paysage mexicain. Le second facteur correspond à la chronologie. Il parcourt l'espace du graphique de bas en haut - ce qu'on peut vérifier si l'on observe le numéro d'ordre qui accompagne chaque titre. Les premiers (de 1 à 8) correspondent à la production initiale de Le Clézio, qui coïncide avec la faveur du nouveau roman. Les derniers (à partir de *Onitsha*) se groupent au haut du graphique. Il y a donc une évolution très sensible dans la production de l'écrivain.

Or ces deux enseignements sont délivrés pareillement par le graphique 11 qui a été établi sur les formes brutes, dont 10 000 étaient homographes et ambiguës. Cette entropie trouble des données n'a nullement empêché une décantation limpide, et la superposition des deux graphiques est parfaite.

Figure 10. Analyse factorielle de la connexion lexicale (base étiquetée)

Facteur 1 58% facteur 2 17% (distance lexicale 59218 mots)

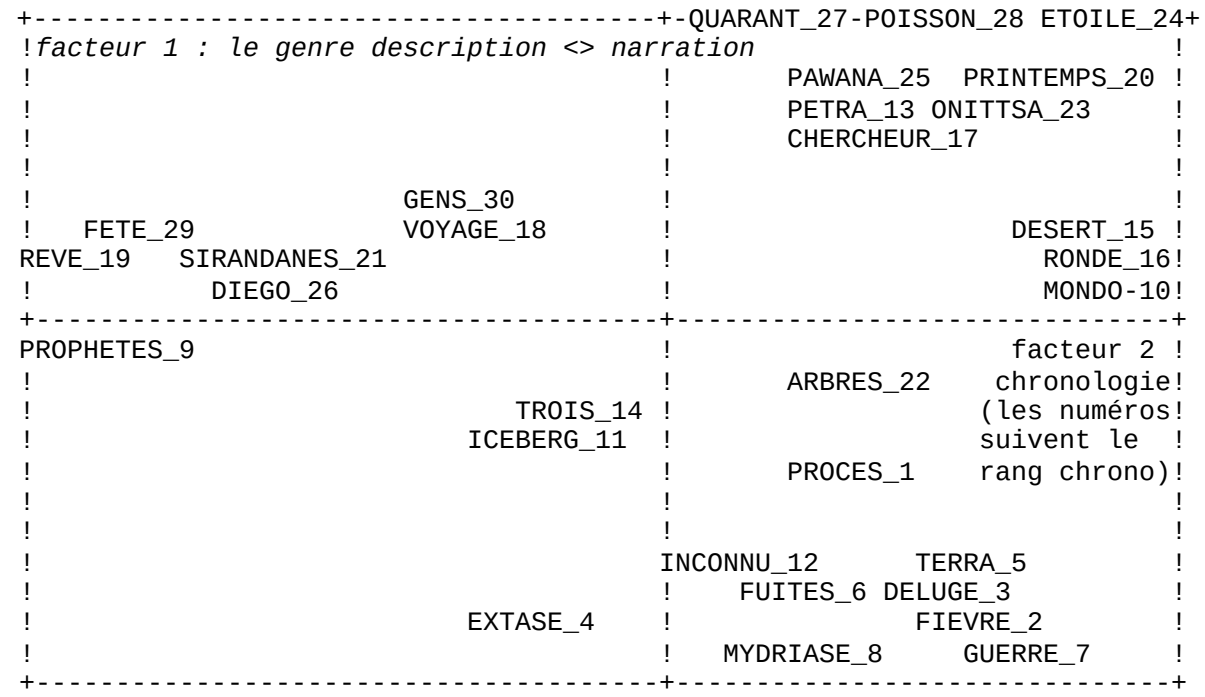
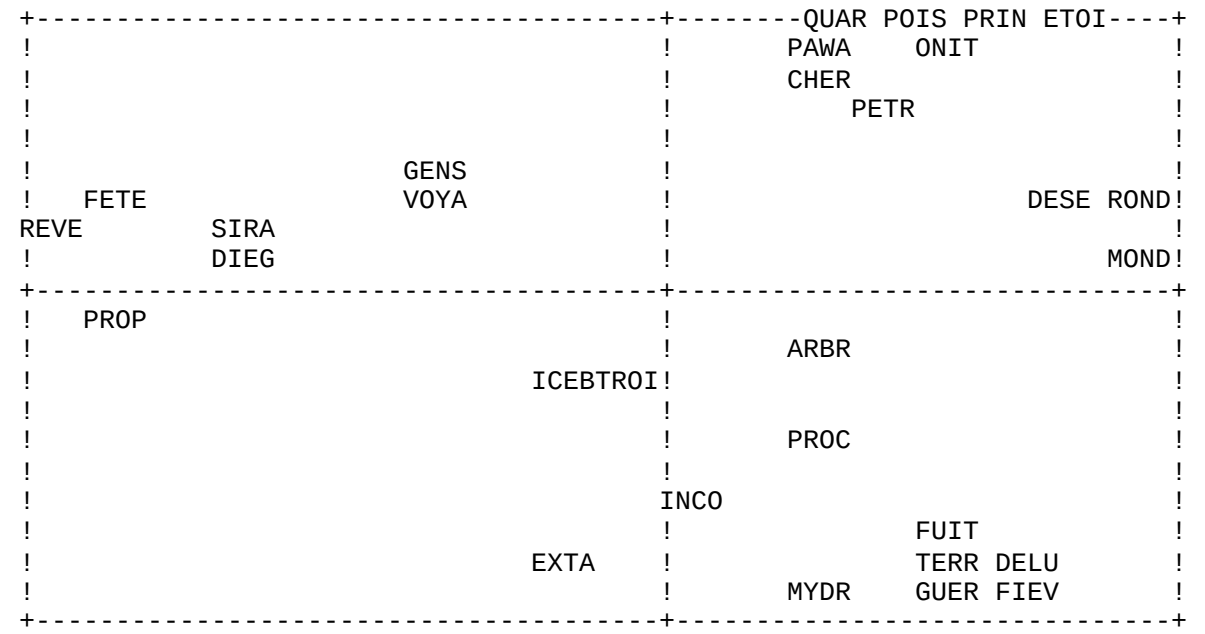


Figure 11. Analyse factorielle de la connexion lexicale (formes brutes)

facteur 1 59% facteur 2 16% (distance lexicale 49773 mots)



André Salem avait déjà observé cette convergence dans une expérimentation qui a précédé la nôtre et dont les conclusions sont semblables: "À partir de tableaux qui présentent des différences importantes dans la mesure

où ils résultent de comptages effectués sur des unités elles-mêmes très différentes, les méthodes de la statistique textuelle (analyse de correspondance, classification hiérarchique, spécificités chronologiques) produisent des résultats très proches"[4].

4 - Il en est ainsi de ce que Salem appelle le TLE (tableau lexical entier) dont rendent compte les figures superposables 12 et 13.

Figure 12. Analyse factorielle du dictionnaire (3000 formes étiquetées)

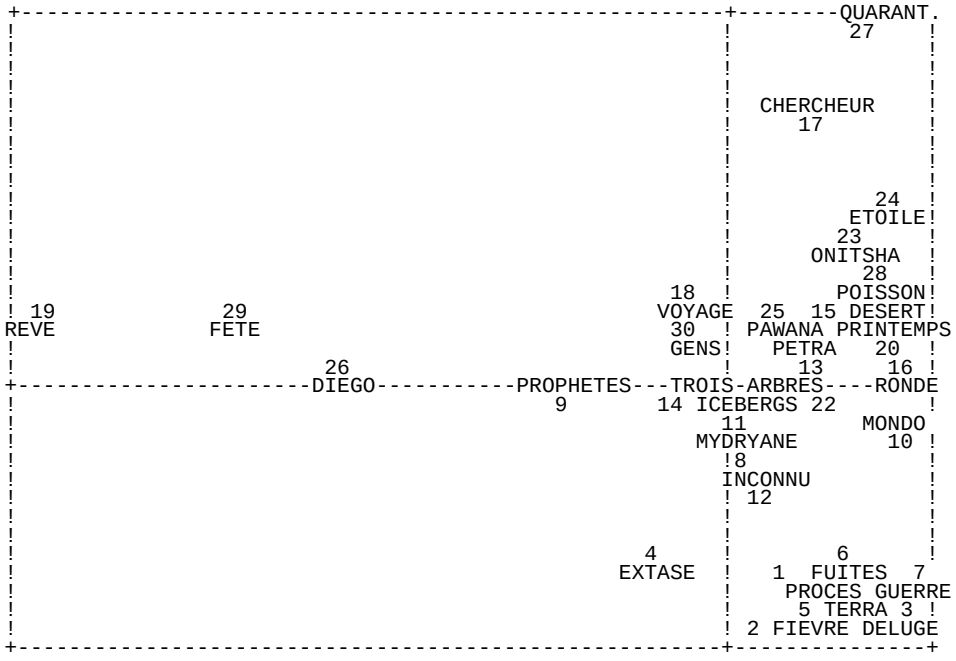
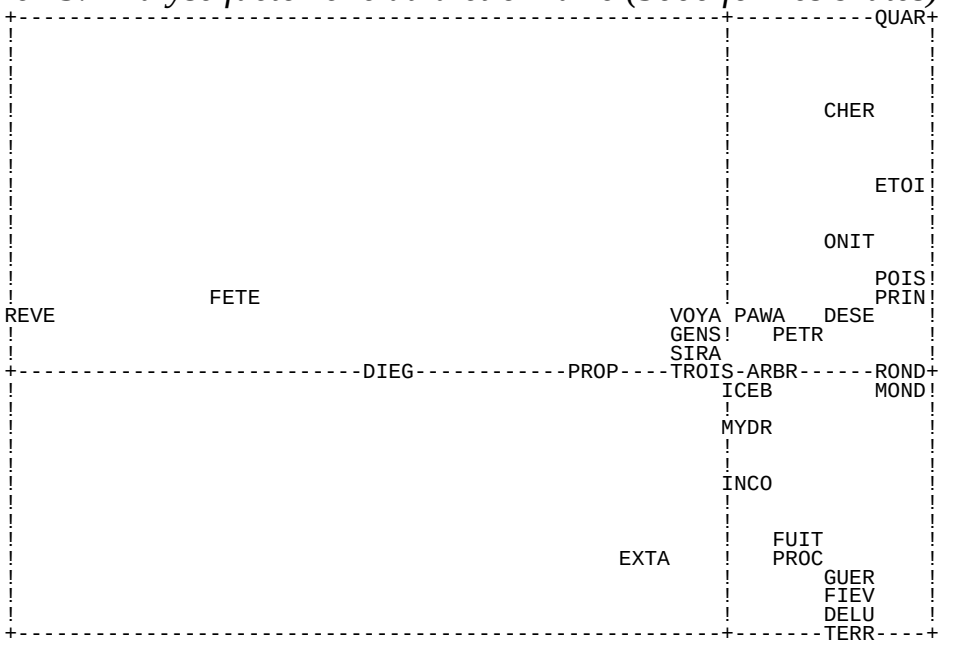


Figure 13. Analyse factorielle du dictionnaire (3000 formes brutes)



4 L. Lebart et A. Salem, *La statistique textuelle,* p. 226.

Il s'agit ici des 3000 formes les plus fréquentes relevées dans les deux corpus. Cette fois la fréquence des mots est considérée - et non plus leur seule présence (ou absence), sur laquelle s'appuyait le calcul de la distance lexicale. Ce tableau est immense et contient 3000 * 30 = 90 000 éléments, qui diffèrent dans les deux corpus. Le parallélisme est pourtant scrupuleusement respecté, comme si l'étiquetage était neutre et sans influence. Certains remarqueront que les facteurs sont les mêmes qu'on avait observés dans l'étude de la connexion (ou distance) lexicale. L'axe principal est pareillement lié au genre, et le second à l'évolution. Or la connexion lexicale est surtout sensible aux basses fréquences, et donc à la thématique, alors que le TLE donne la part belle aux fréquences hautes, et donc aux variables stylistiques. Variables stylistiques et thématiques semblent donc aller de pair, dans un ensemble surdéterminé où la même image est obtenue quel que soit l'angle de vue, comme s'il s'agissait d'une boule.

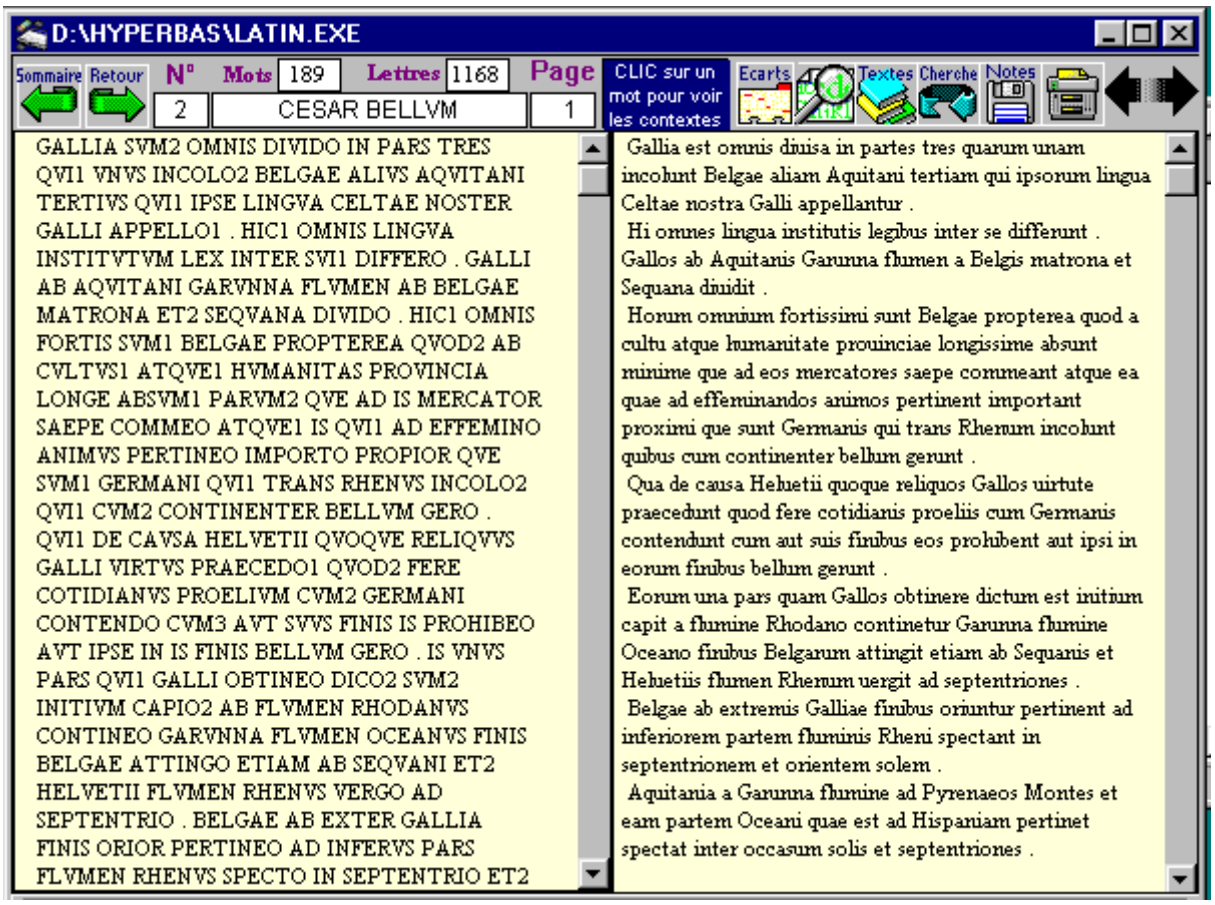
IV - Les lemmes

L'expérience précédente laisse subsister un doute, à cause des faiblesses du logiciel d'étiquetage: imaginons un étiqueteur défaillant qui distribuerait les codes au hasard. Quel serait le résultat sinon précisément celui que proposent les données brutes? On aurait pu choisir des outils plus sûrs, comme le lemmatiseur développé par Maucourt et Papin à l'INaLF (sur lequel est fondée la nouvelle base de *Frantext*), ou celui de *Cordial* (que le traitement de texte *Microsoft Word* a emprunté pour le français) , ou encore celui de *Sylex* intégré au logiciel *Sphinx Lexica*. Nous avons visé plus haut encore, en traitant les données lemmatisées du *LASLA* de Liège, aimablement fournies par Sylvie Mellet. Le latin en effet, du fait de la déclinaison, est difficilement exploitable si l'on se contente des données brutes. Le *LASLA* a donc pourvu à l'analyse semi-automatique de toutes les formes rencontrées dans les textes dépouillés. Cette analyse est très précise et très fiable puisqu'elle a été partout dirigée et contrôlée par une équipe de chercheurs spécialisés qui ont fourni la bonne réponse chaque fois que l'automate hésitait.

1 - Cette fois le processus de transformation est complet. Le texte soumis à notre programme n'est plus celui des formes de l'original, mais la séquence des lemmes auxquels les formes renvoient. L'effectif que l'on voyait croître avec le dégroupement de l'étiquetage s'amenuise considérablement, maintenant qu'on n'envisage que la vedette de regroupement, et l'on passe de 71 657

formes brutes à 16 656 vocables, soit moins du quart. Comme en perdant ses désinences le texte latin perdait toute lisibilité, on a restitué le texte original dans une présentation synoptique, dont la figure 14 donne un exemple.

Figure 14. La première page de la Guerre des Gaules
Lemmes (à gauche) et formes (à droite)



Mais une autre base symétrique a été constituée à partir du texte original, sans aucune préparation. Là encore on cherche à appliquer le même calcul au même texte, en prenant en compte tantôt les formes brutes, tantôt les lemmes. Et pour profiter des explications données précédemment, nous choisirons là aussi le calcul de la distance entre deux textes. Ceux-ci sont nombreux (69) même s'ils ne recouvrent pas la totalité de la littérature latine (il manque en particulier Tite-Live). Les mesures sont abondantes et précises, soit (69 * 68) / 2 = 2346 au total. L'analyse factorielle appliquée à ce tableau est sensible à l'air de famille qui unit les textes appartenant au même auteur, sauf s'il s'agit de Sénèque, qui partage sa production dans deux genres opposés. Si les tragédies de Sénèque font corps avec la poésie à droite du graphique, sa prose

philosophique et morale fait la transition au haut du graphique, à cheval sur l'axe vertical. De l'autre côté de l'axe, c'est le territoire de la prose que se partagent les historiens (César en bas de la carte, puis Salluste, Quinte-Curce et Tacite au centre), les orateurs et les moralistes (dans la partie supérieure). Ce qui nous importe ici est moins l'influence attendue du genre, que le parallélisme des deux analyses. La seconde (figure 16) constituée à partir des lemmes, confirme en tous points celle des formes brutes[5] (figure 15).

*Figure 15. Analyse de la connexion lexicale dans la littérature latine (**Formes**)*

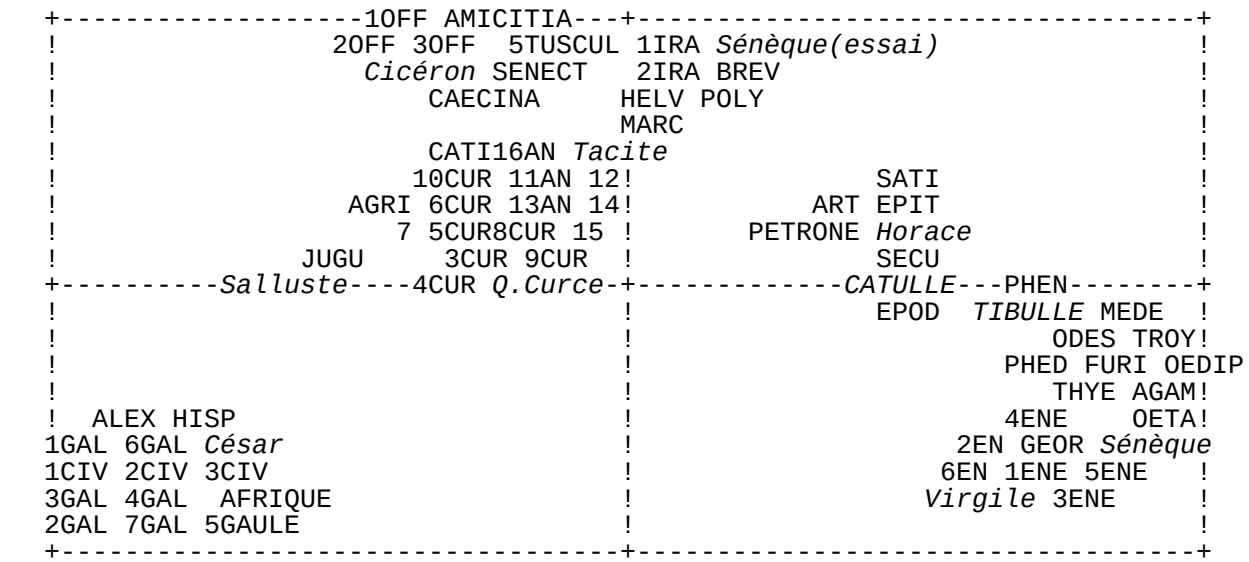
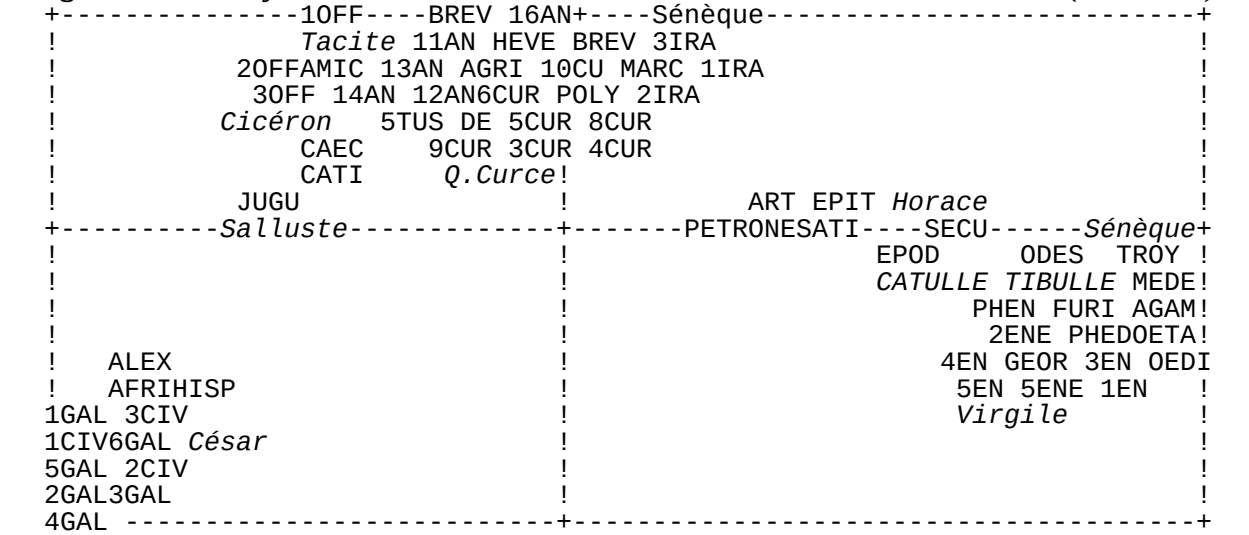


*Figure 16. Analyse de la connexion lexicale dans la littérature latine (**Lemmes**)*

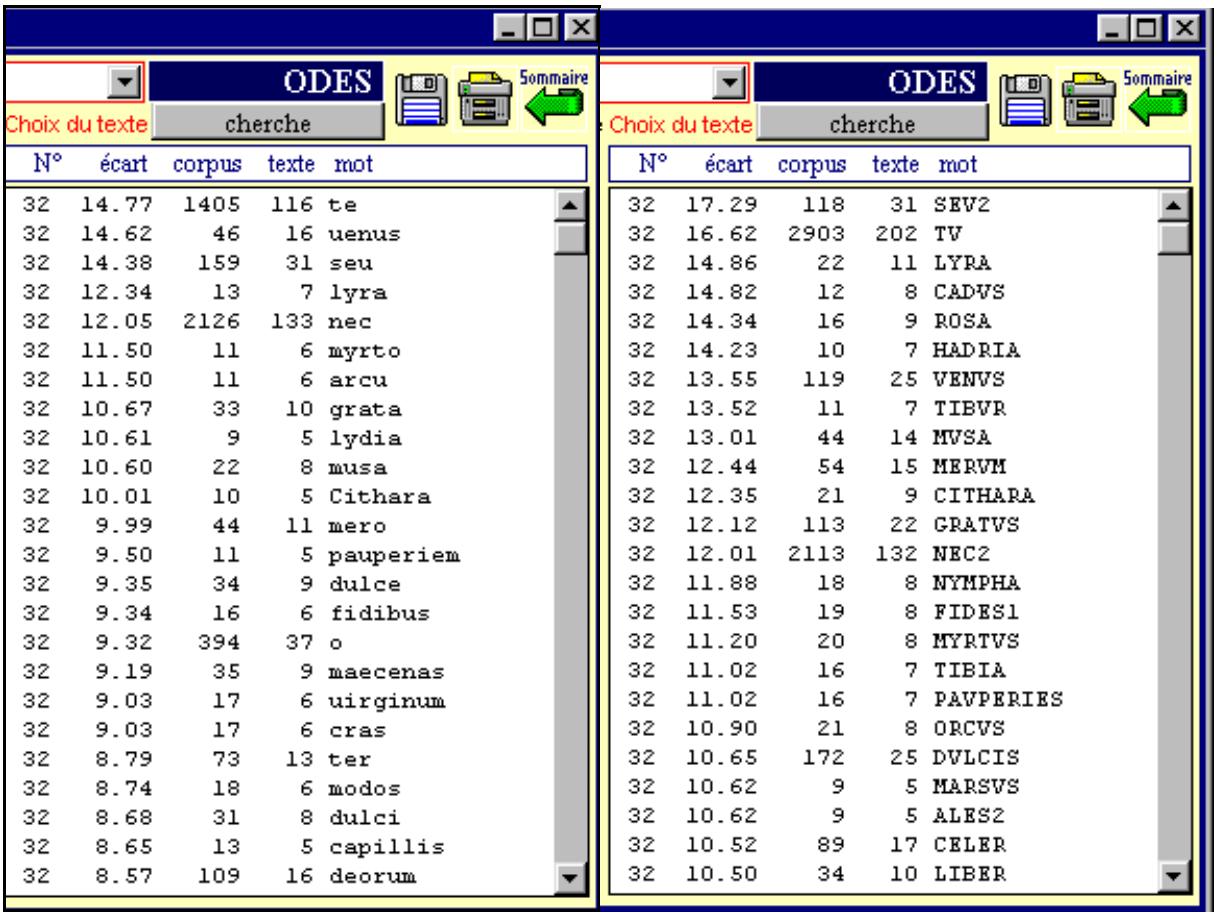


[5] La place nous manque pour illustrer plus avant la convergence des deux approches et reproduire ici les deux analyses réalisées sur le tableau lexical entier, à l'instar des figures 12 et 13. Là encore le parallélisme est très étroit.

2 - Il va sans dire que nous préférons l'analyse fondée sur les lemmes. Les homographes latins, qui sont plus nombreux encore que ceux du français, engendrent une pollution diffuse, qui, de loin, peut échapper à l'observation. Mais dans le détail des mots individuels, la différence est sensible entre formes brutes et lemmes. La sûreté de ces derniers vient certes de leur pureté mais aussi d'une assise plus large livrée à l'observation. L'émiettement des formes diminue en effet l'effectif de chacune. Elle affaiblit aussi leur témoignage que d'autres formes, appartenant au même lemme, peuvent contester ou contredire, et le jugement reste en suspens quand un même vocable distribue ses représentants parmi les excédents et les déficits. Ainsi les deux listes de spécificités qu'on obtient pour le même texte ne se recouvrent que partiellement dans l'exemple des *Odes* d'Horace (figure 17). La moitié des vingt premiers éléments de la liste des lemmes n'a pas de correspondant, du moins au même rang, dans celle des formes. Certes on a quelque chance de retrouver les absents, éparpillés dans le reste de la liste. Mais l'ordre de préséance est bousculé. Et les observations, portant sur des effectifs partiels (*te*, en tête de liste, n'a que la moitié des occurrences du vocable *TV)*, n'ont pas même fiabilité, ce dont témoigne aussi la valeur amoindrie des écarts réduits.

Figure 17. Les spécificités des Odes *d'Horace*

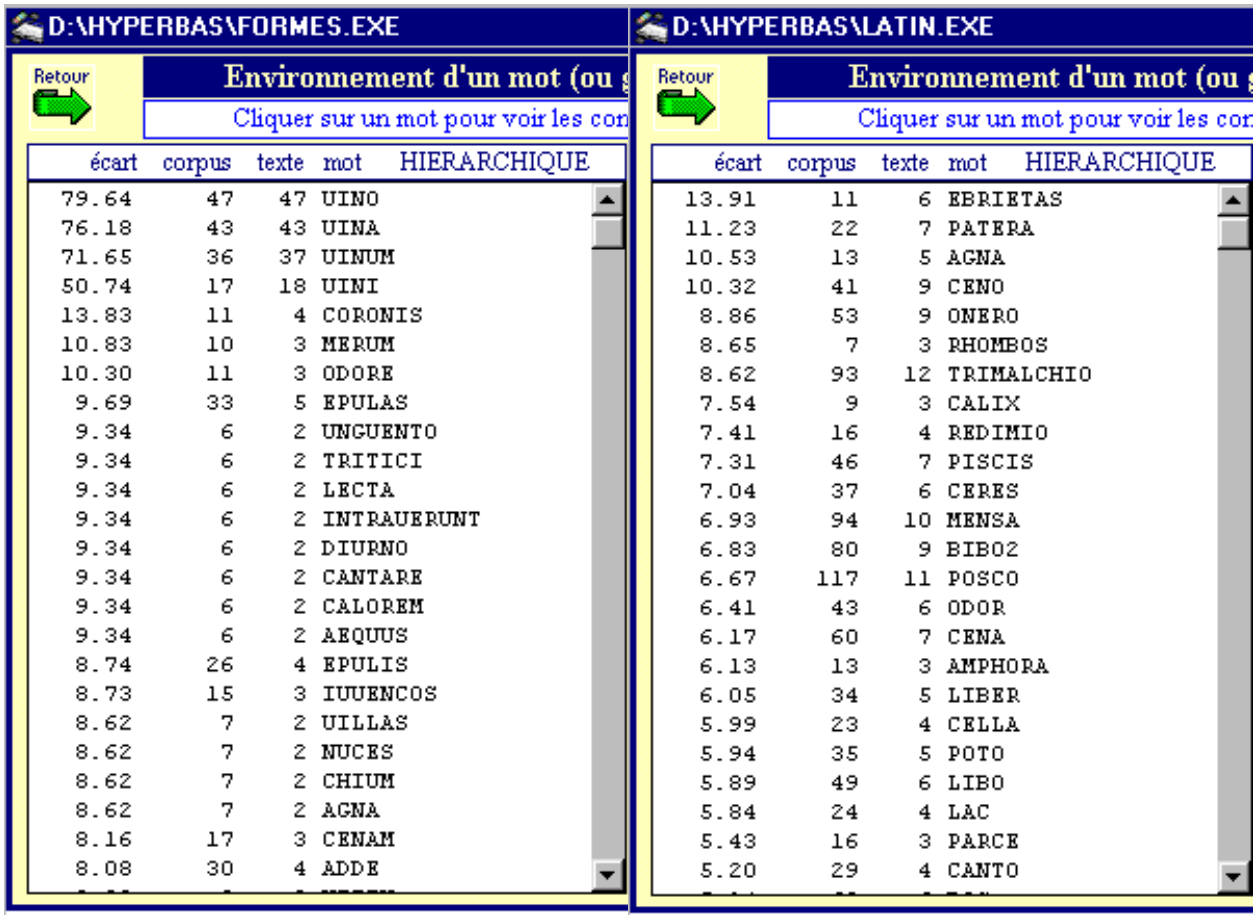
Formes brutes ***Lemmes***



3 - Si la focalisation porte sur un mot dont on souhaite cerner l'environnement, les résultats diffèrent considérablement selon que l'on a affaire à des lemmes ou à des formes fléchies, même si le relevé des contextes est identique (ce qu'on peut faire en réunissant les différentes flexions du mot-pôle, comme les formes *vinum, vini, vino* et *vina*, dans l'exemple ci-dessous, figure 18). La liste des corrélats associés au vocable *VINVM* est beaucoup plus claire et significative, s'il s'agit de lemmes (partie droite de la figure) et les rencontres avec le mot-pôle s'y produisent plus souvent, ce qui donne à la statistique une assise plus solide. On gagne ainsi en quantité comme en qualité.

Figure 18. Les corrélats du vin

Formes brutes ***Lemmes***



****

Faut-il donc lemmatiser? "La décision , conclut André Salem, est d'ordre économique. Il est dans l'absolu toujours préférable de disposer d'un double réseau de décomptes (en formes graphiques et en lemmes). Une lemmatisation complète, sur un corpus important, reste une opération coûteuse. Indispensable

dans un travail de recherche, elle est beaucoup moins justifiée s'il s'agit d'obtenir rapidement des visualisations et des typologies […] "[6]. On voit que la querelle ancienne a beaucoup perdu de son alacrité. Comme celle qui opposait les partisans de la loi hypergéométrique et ceux de l'écart réduit. Comme celle, plus radicale encore, qui confrontait jadis les tenants et les adversaires du schéma d'urne. Trente ans après les travaux des pionniers, des Guiraud et des Muller, la discipline s'est assagie. Elle dispose maintenant d'une panoplie d'outils spécialisés, dont l'utilité est reconnue suivant l'usage qu'on veut en faire. Les outils ne sont mauvais que dans la main des mauvais ouvriers.

[6] L. Lebart et A. Salem, *La statistique textuelle*, p 226.