Le lemme comme on l'aime

Étienne Brunet

UMR 6039 *Bases, corpus et langage* — Institut de Linguistique française – CNRS – Faculté des Lettres – 98 bd. Herriot – 06204 Nice Cedex – France – brunet@unice.fr

Abstract

The familiar debate concerning lemmatization of textual data may have lost virulence but still remains topical. Comparison made between studies based on non-lemmatized data, which are still valuable and the more sophisticated processing of normalized and tagged data, almost always confirms the results obtained from raw data. However, access to lemma and to grammatical tags opens up new dimensions to research. Whether considering phrase functions, parts of discourse, verbal tense, person or mood, the syntactical rather than the lexical aspects will be considered, exploring a new version of the HYPERBASE tool and involving an experimental corpus of 2 million tokens, including 26 texts and 13 authors.

Résumé

Le vieux débat sur la lemmatisation a perdu de sa violence, mais non de son actualité. On tente de montrer ici que les études lexicométriques qui se contentent des formes graphiques gardent leur valeur et que le traitement plus élaboré des données étiquetées et lemmatisées confirme le plus souvent les résultats acquis à partir des données brutes. Mais l'accès au lemme et aux codes grammaticaux ouvre des champs plus riches à la recherche.. Qu'il s'agisse des fonctions dans la phrase, des parties du discours, ou des temps, des personnes ou des modes verbaux, on explore ici ces perspectives, plus syntaxiques que lexicales, à travers un corpus expérimental de 2 millions de mots, regroupant 26 textes et 13 écrivains et soumis à une nouvelle version du logiciel HYPERBASE.

Mots-clés : lemmatisation, corpus étiquetés, Cordial, fonctions grammaticales, parties du discours, temps verbaux

1. Lemmatisation

1.1. Dans les travaux de linguistique quantitative, la prudence a souvent choisi le même camp que la paresse. En s'abstenant de lemmatiser les données, elle adoptait un profil bas, avouant l'impureté des données et faisant confiance à la statistique pour les dégager de l'entropie. Mais cette position attentiste peut-elle être indéfiniment prolongée? En trente années les industries de la langue ont fait des progrès et des outils de plus en plus performants sont disponibles sur le marché. Rares sont les rédacteurs qui méprisent l'usage du correcteur d'orthographe. On lui pardonne ses bévues eu égard aux services qu'il rend pour signaler les fautes de frappe et les accords négligés. Or il n'y a pas de correction possible sans analyse préalable. Et la lemmatisation entre nécessairement dans le processus. Les concepteurs de logiciels statistiques ont suivi cette tendance, parfois à moindres frais. En s'appuyant sur la troncature, ils ont pu isoler le radical et soumettre au calcul des effectifs regroupés, où la dispersion des formes fléchies était neutralisée. Et notre HYPERBASE a tenté de suivre dans cette voie l'exemple de TROPES, d'ALCESTE et de SPHINX (pour s'en tenir au domaine français).

Mais notre première tentative s'est révélée décevante et la version lemmatisée d'HYPERBASE n'a jamais été distribuée. Il y avait une raison juridique à cela: elle reposait sur le logiciel WINBRILL qui est certes gratuit mais dont la version française, fruit des efforts conjugués de deux chercheurs de l'INaLF, J. Lecomte et G. Souvay, ne nous appartenait pas. S'y ajoutait un embarras méthodologique: d'une part Winbrill n'opère qu'un étiquetage grammatical et l'on doit lui adjoindre des fonctions complémentaires pour accéder au lemme.

D'autre part les codes qu'on y distingue sont peu classiques et peu précis. La classe des déterminants n'est pas détaillée ; celle des pronoms manque de clarté et celle des verbes ignore les modes, les temps et les personnes. Un autre logiciel de lemmatisation a été mis au point dans le même laboratoire et a servi à constituer la nouvelle version de FRANTEXT où la catégorie grammaticale s'ajoute à la panoplie des critères de sélection (une forme, un vocable, une expression, une cooccurrence, une liste, une alternative, ou toute combinaison de ces objets). Mais ce produit interne n'était pas disponible à l'extérieur.

1.2. À qui donc s'adresser? On a songé d'abord à celui qui, sabre au clair, a maintenu sans faiblesse les exigences de la lemmatisation, à Dominique Labbé. Ses études lexicométriques, en particulier sur de Gaulle et Mitterrand, donnaient toutes les garanties souhaitables. Mais son logiciel, conçu pour une version ancienne du système Macintosh, exigeait une refonte préalable, rude tâche à laquelle ce chercheur a bien voulu s'atteler. En attendant que la nouvelle version soit disponible, un autre produit s'imposait, que beaucoup de gens utilisent sans le savoir et qui s'appellent Cordial. Le correcteur que Word Microsoft a parfois intégré à son traitement de texte est en effet emprunté à Cordial. Les nombreux prix glanés ici et là par ce logiciel s'accordent avec cette préférence enviée, qui en fait le correcteur le plus utilisé en France. Au reste les concepteurs de ce produit sont ouverts à la recherche universitaire et ont facilité l'expertise que mènent là-dessus François Rastier et son équipe. En particulier une version particulière du logiciel est destinée aux laboratoires spécialisés dans le traitement automatique de la langue, auxquels elle fournit un outil d'analyse et non plus seulement de correction. Cette version, anciennement dénommée "Cordial Université", est maintenant distribuée sous l'étiquette ANALYSEUR et correspond à la version 7 du produit standard. On pourrait penser à priori que ce produit se suffit à lui-même, puisqu'il est apte à délivrer des contextes pour tous les mots ou configurations qu'on lui propose et qu'il fournit à foison des statistiques d'ordre lexical, syntaxique et même sémantique. Cependant de telles statistiques sont toujours globales et s'appliquent au texte entier sans offrir aucune partition de l'ensemble, interdisant ainsi les comparaisons internes, même si une confrontation extérieure est fournie qui s'appuie sur un immense corpus de référence. Et d'autre part les conclusions restent incertaines car Cordial 7 s'en tient aux effectifs absolus et aux pourcentages sans jamais accéder aux véritables tests statistiques, encore moins aux méthodes multidimensionnelles. Il y avait donc place pour une expérimentation dont nous nous proposons de rendre compte.

Comme François Rastier s'est donné la tâche de cerner, à travers Cordial, les limites et les propriétés du genre littéraire, nous écartons d'emblée cette variable en réunissant des textes qui appartiennent tous au genre narratif. Le corpus est donc homogène à ce point de vue, les variables retenues concernant l'époque et l'auteur. Vingt-six textes ont été choisis qui illustrent le genre romanesque du XVIIIe siècle à nos jours.

N°	TITRE et AUTEUR	OCCURRENCES	Prob P	Prob Q	ABREGE	CODE
1	La vie de Marianne (L.1), MARIVAUX	19963	.0091	.9909	Marianne	Ma
2	Le Paysan Parvenu (L.1), MARIVAUX	21283	.0097	.9903	Paysan	Py
3	Zadig,VOLTAIRE	31435	.0144	.9856	Zadig	Za
4	Candide,VOLTAIRE	40009	.0183	.9817	Candide	Ca
5	La nouvelle Héloïse(L.1), ROUSSEAU	73820	.0338	.9662	Héloïse	Hé
6	Emile (L. 5), ROUSSEAU	83729	.0383	.9617	Emile	Em
7	Atala, CHATEAUBRIAND	35513	.0162	.9838	Atala	At
8	La vie de Rancé, CHATEAUBRIAND	70406	.0322	.9678	Rancé	Ra
9	Les Chouans, BALZAC	137474	.0629	.9371	Chouans	Ch
10	Le cousin Pons, BALZAC	129457	.0592	.9408	Pons	Po
11	Indiana, Georges SAND	112257	.0513	.9487	Indiana	In
12	La mare au diable, Georges SAND	46500	.0213	.9787	Mare	Ma
13	Madame Bovary, FLAUBERT	145798	.0667	.9333	Bovary	Bo
14	Bouvard et Pécuchet, FLAUBERT	113985	.0521	.9479	Bouvard	Bu
15	Une Vie, MAUPASSANT	90766	.0415	.9585	UneVie	Vi
16	Pierre et Jean, MAUPASSANT	53863	.0246	.9754	Pierre	Pi
17	Thérèse Raquin, ZOLA	84752	.0388	.9612	Raquin	Rq
18	La Bête humaine, ZOLA	164983	.0754	.9246	Bête	Bê
19	De la terre à la lune, VERNE	67440	.0308	.9692	Lune	Lu
20	Secret de Wilhelm Storitz, VERNE	64189	.0294	.9706	Storitz	St
21	Du coté de chez Swann, PROUST	203754	.0932	.9068	Swann	Sw

```
22 Le temps retrouvé, PROUST          166255   .076   .924   Temps      Tm
23 Moderato cantabile, DURAS           23624   .0108  .9892  Moderato   Mo
24 Ravissement, DURAS                  46996   .0215  .9785  Ravissemen Ra
25 Le Procès, LE CLEZIO                91027   .0416  .9584  Procès     Pr
26 Hasard, LE CLEZIO                   67650   .0309  .9691  Hasard     Ha
TOTAL                                2186928
```

Figure 1. La composition du corpus

On aurait pu étendre à 26 le nombre des auteurs, pour un échantillonnage plus varié et plus large. Mais on a préféré représenter le même écrivain par deux textes publiés par lui, si possible, aux deux extrémités de sa carrière. On voulait ainsi, à genre constant, accroître la distance entre les textes d'un même auteur (il y a par exemple plus de 40 ans dans la vie de Chateaubriand entre *Atala* et la *Vie de Rancé*) et voir si cette distance allait se maintenir ou se réduire quand la comparaison met en scène d'autres écrivains. Autrement dit on voulait mesurer conjointement la distance intra (qui oppose les textes d'un même auteur) et la distance inter (qui confronte les écrivains entre eux).

1.3. La composition du corpus est détaillée ci-dessus (figure 1). On y compte plus de deux millions de mots. Comme cela risque de dépasser les capacités de Cordial, le programme de lemmatisation est lancé pour chaque texte, en veillant à maintenir constants les paramètres de présentation, selon le modèle de la figure 2.

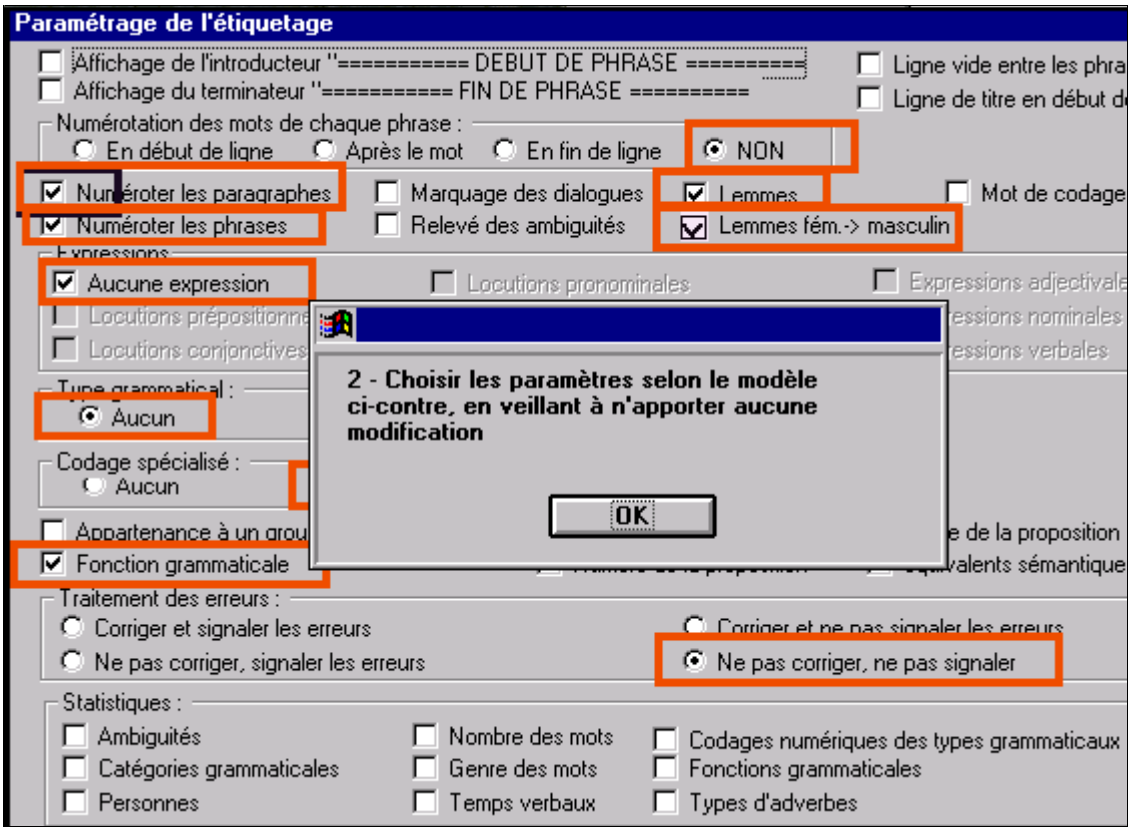

Figure 2. Les options de l'analyse dans Cordial

Outre le code grammatical qu'il propose de trois façons différentes, Cordial ajoute de nombreux renseignements relatifs au traitement des expressions, à la fonction dans la phrase, à la place hiérarchique du mot dans l'arbre syntaxique, et même à la classe sémantique à laquelle le mot se rattache. Nous n'avons retenu que ce qui était strictement nécessaire à l'analyse, soit la moitié des possibilités offertes dans la figure 3, à savoir: le numéro du paragraphe, le numéro de la phrase, la forme, le lemme, le code grammatical détaillé (codegram) et la fonction.

N°	§	Phrase	Forme	lemme	ambig.	Typegra	CodeHexa	Codegram	Syntagme	Fonction	Num Sens	
==== DEBUT DE PHRASE ==												
1	1	1	Je	je		36	0xE480	Pp1.sn	1	S	1	
2	1	1	crois	croire	A3	101	-	Vmip1s	2	V	1	
3	1	1	que	que	A3	21	0x0000	Cs	-	-	2	
4	1	1	la	le	A3	15	0x6000	Da-fs-d	5|5	T	2	
5	1	1	langue	langue		26	0x6080	Ncfs	5|5	T	2	forme
6	1	1	est	être	A3	103	-	Vmip3s	6	V	2	

Figure 3. Un fichier lemmatisé par Cordial

1.4. Hyperbase prend alors en compte les trois principaux éléments d'un tel fichier et les distribue séquentiellement dans trois champs parallèles, voués respectivement aux formes, aux lemmes et aux codes. La figure 4 met en correspondance les formes et les lemmes d'une même page de Proust. On notera que les lemmes, dans la partie droite de l'écran, sont pourvus d'un indice numérique, afin de séparer les uns des autres les homographes. Ainsi *le 7* (dans *le sifflement*) distingue l'article du pronom codé 4 (dans *qui le suivent*). Ces codes simplifiés qui reproduisent la classification de Muller et de Labbé (1 verbe, 2 substantif, 3 adjectif, 4 numéral, 5 pronom, 6 adverbe, 7 déterminant, 8 conjonction, 9 préposition) n'appartiennent pas en propre à Cordial, mais ont été dérivés de l'analyse complète fournie par Cordial.

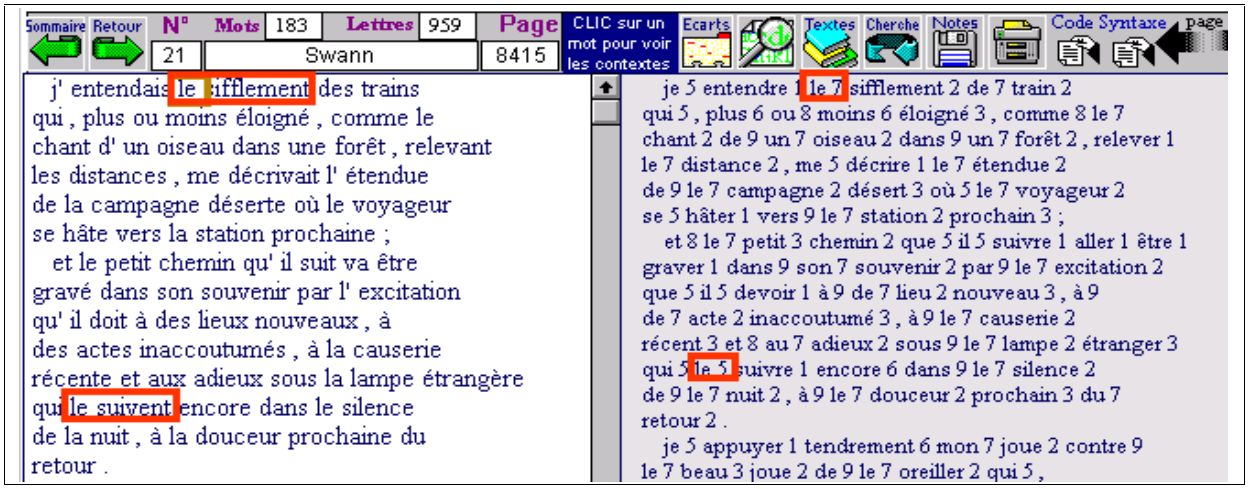

Figure 4. L'alignement forme-lemme

Cette analyse complète est rendue visible, quoique peu lisible, pour peu qu'on sollicite le bouton CODE situé à droite de la barre de menu. Là aussi l'alignement est rigoureux, en sorte que l'on sait précisément à quel mot correspond telle ou telle analyse. Ces trois champs sont sensibles au clic de la souris: tout objet que l'on désigne, qu'il s'agisse d'une forme, d'un lemme ou d'un code, renvoie aux autres occurrences où le même objet est rencontré, les relations hypertextuelles s'appliquant aux trois champs. Mais ces relations lient aussi entre eux ces trois champs, en sorte qu'en cliquant sur un code grammatical dans le champ de droite (par exemple *_Afpms*, soit *adjectif qualificatif, positif, masculin, singulier*) on obtient successivement en vidéo inverse tous les adjectifs qui répondent à ce codage dans le champ de gauche.

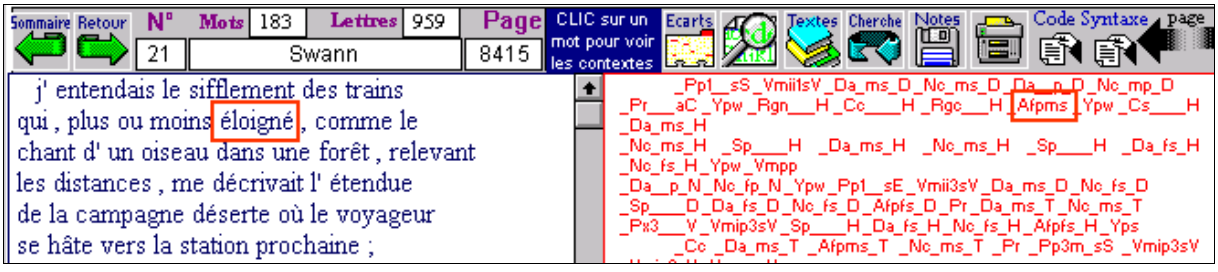

Figure 5. L'alignement forme-code

L'indexation et toutes les opérations subséquentes sont alors répétées trois fois, au niveau des codes, puis des lemmes, puis des formes. À l'issue de ce traitement, on obtient trois index (figure 6) qui réagissent pareillement au clic de la souris. La forme, ou le lemme ou le code qu'on désigne montre le détail de ses occurrences, parmi lesquelles l'utilisateur fait son choix pour se référer au texte.

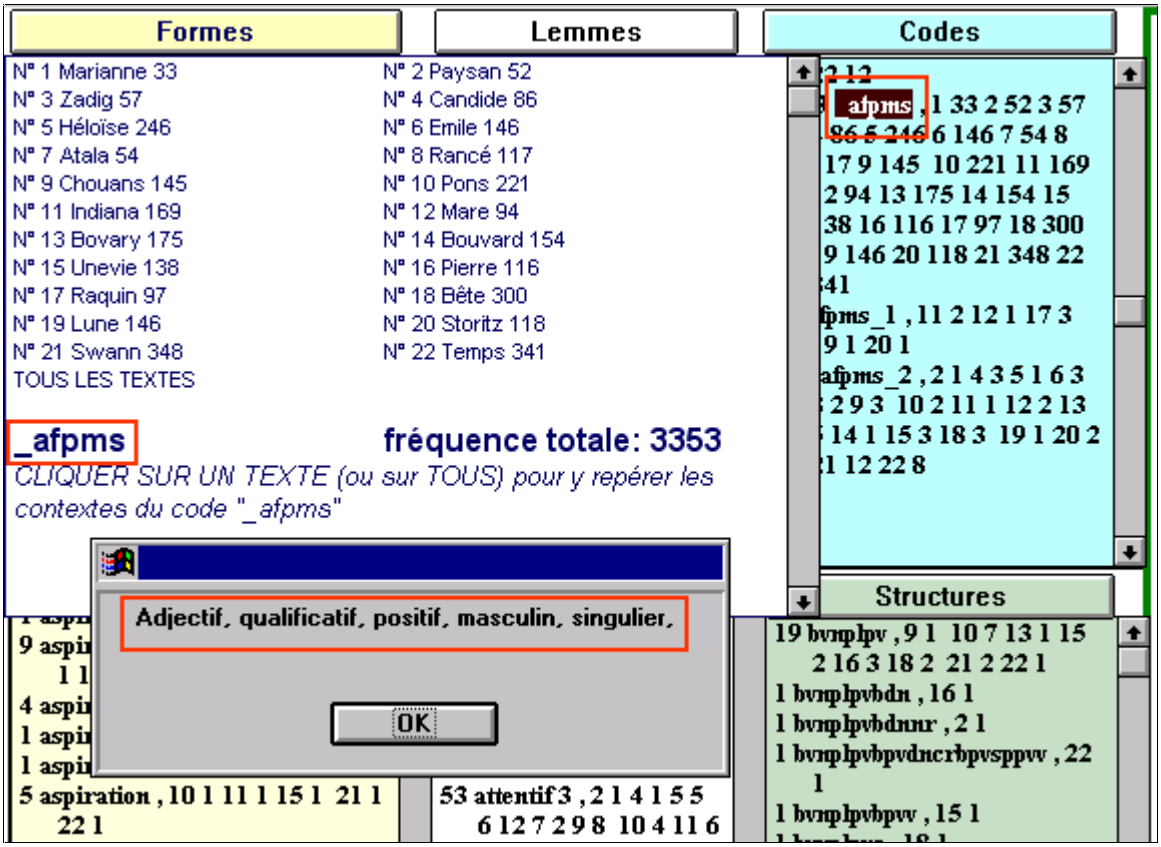

Figure 6. Les quatre index issus de Cordial

S'il s'agit d'un code, dont la signification peut être opaque, le décryptage est assuré et traduit en clair, comme dans l'exemple de la figure 6, relatif à l'adjectif qualificatif, au pluriel, dans un groupe en apposition (c'est le même exemple que celui de la figure 5). Pour faciliter les recherches on a joint la fonction au code grammatical en dernière position (ici la lettre N pour le groupe en apposition).

2. Exploitation

2.1. Qu'il s'agisse du texte ou du dictionnaire, la démarche qu'on vient de décrire est exploratoire: ayant un mot, un lemme ou un code sous les yeux, on s'interroge à son sujet et les fonctions hypertextuelles ou statistiques fournissent les informations relatives à l'objet relevé. Mais la démarche peut être inverse. Ayant en tête un mot, un lemme , une catégorie ou une hypothèse, on cherche à vérifier sa présence ou sa validité dans le corpus. S'il s'agit d'une forme, il suffira de l'inscrire dans la zone de dialogue que la fonction sollicitée (concordance, contexte, liste, graphique) présente à l'utilisateur. Si c'est un lemme, on ajoutera un blanc pour éviter que le lemme soit confondu avec la forme simple. Inutile d'ajouter le code numérique qui accompagne chaque lemme. Car le logiciel le récupère automatiquement, même si l'on a affaire à un homographe. Dans ce cas un dialogue supplémentaire apparaît, qui précise toutes les options possibles en demandant de faire un choix. Ce choix peut être large, comme dans le cas de l'homographe *tout*, qui fait l'objet de la figure 7 et dont le résultat est consigné dans la figure 8 .

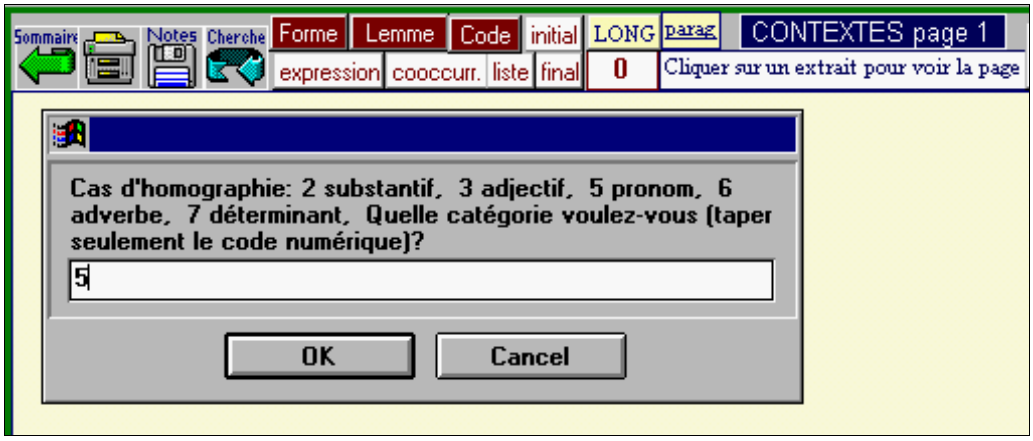

Figure 7 . La désignation des homographes (ici tout*)*

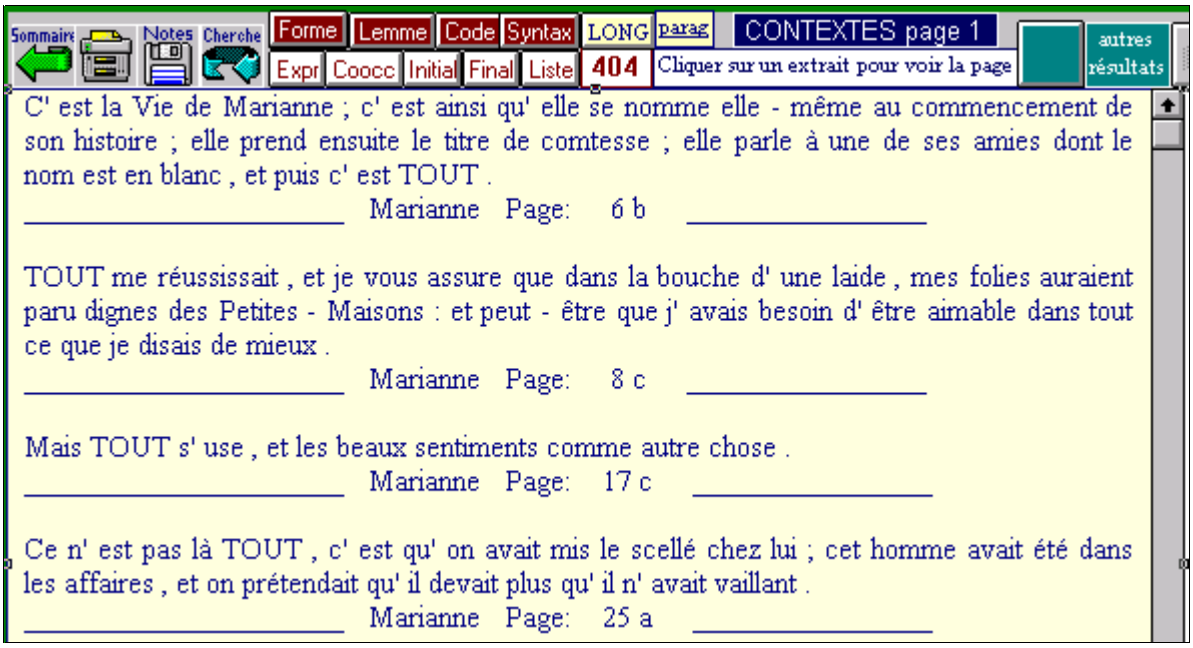

Figure 8. Les contextes de tout *pronom*

2.2. Lorsqu'on a affaire aux codes grammaticaux, on est renvoyé à une page spéciale (figure 9) qui dénombre toutes les combinaisons possibles. Car Cordial pousse loin l'analyse, en relevant pour chaque mot la catégorie, la sous-catégorie, le genre, le nombre, la fonction et s'il s'agit d'un verbe le temps, le mode et la personne. Un clic dans une option provoque alternativement l'activation ou la désactivation correspondante. Certaines options sont impliquées ou exclues automatiquement, dès qu'une autre est choisie, de telle façon qu'il y ait toujours cohérence. Car il serait absurde de sélectionner le futur d'un substantif ou le féminin d'un verbe à l'infinitif. Chaque clic modifie le filtre dont l'affichage apparaît dans une fenêtre, en haut et à droite de l'écran, avec sa traduction en clair. Toute colonne non intéressée par la sélection est remplie par défaut par un joker, dont l'effet est d'admettre tout code qu'on rencontre à cet endroit. Ainsi dans l'exemple choisi la colonne 3 n'ayant pas été sélectionnée, tous les adjectifs seront retenus, quel que soit le degré, positif ou comparatif. De même le vide rencontré dans la colonne 7 laissera la sélection indifférente à la fonction dans la phrase.

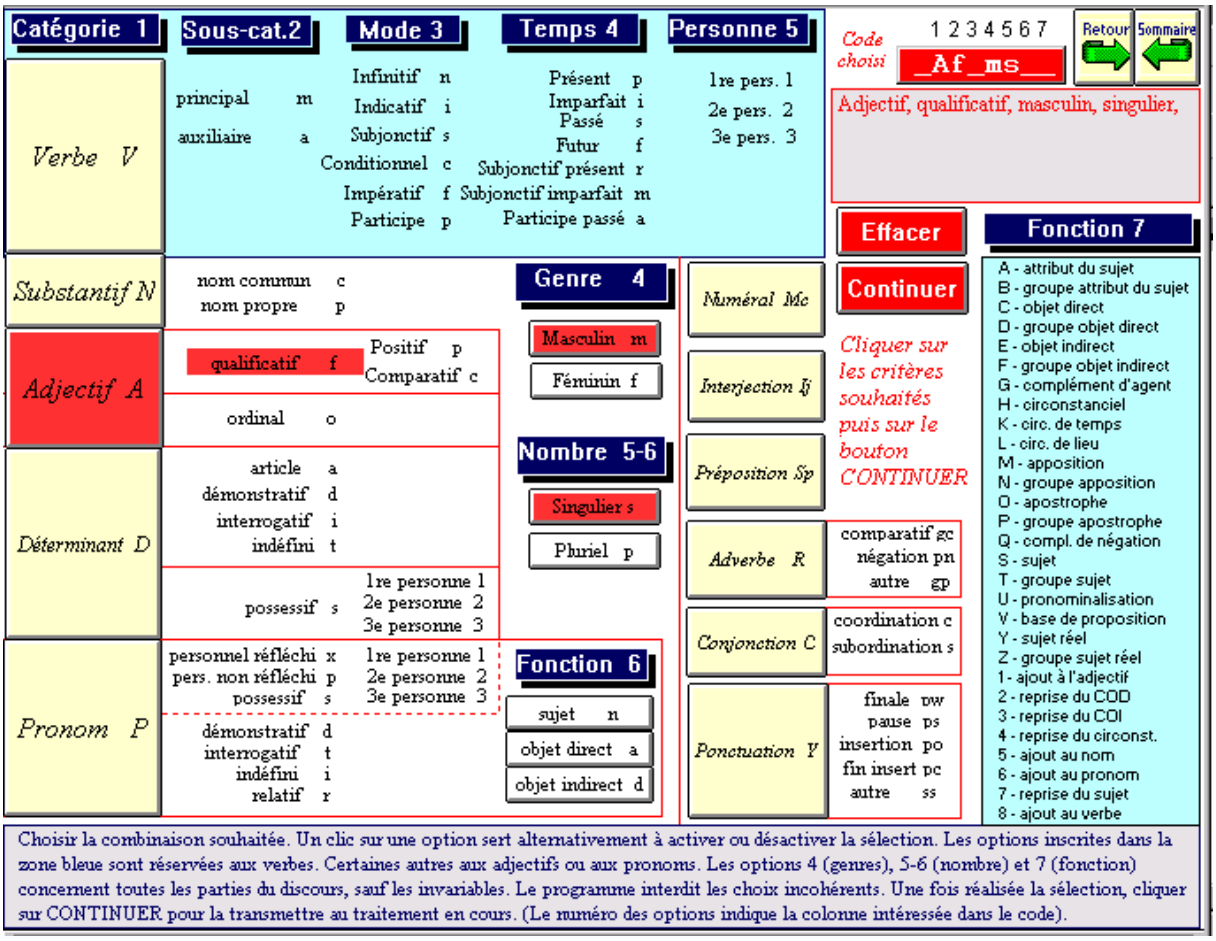

Figure 9. Le choix d'un code grammatical

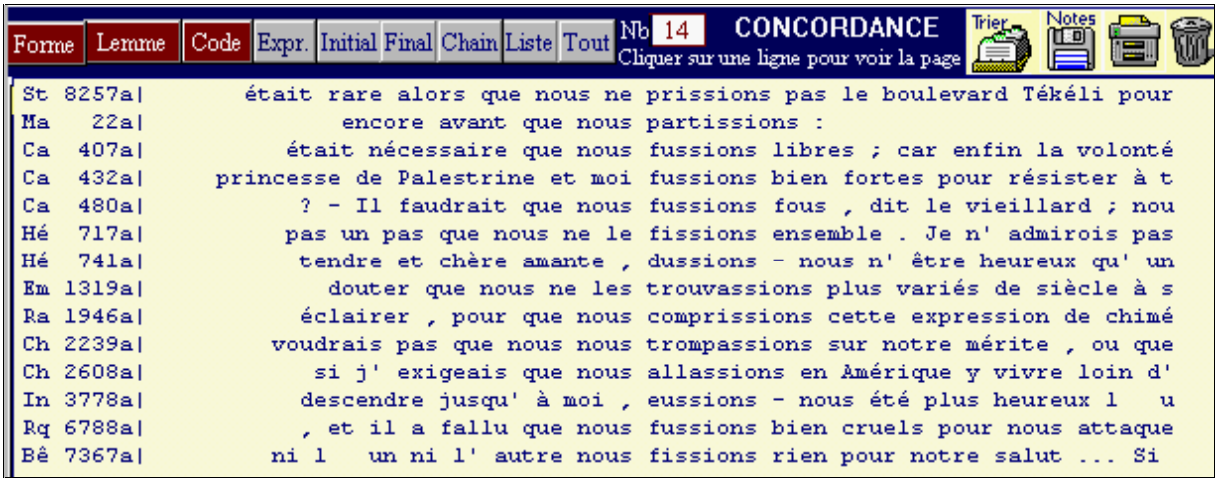

Figure 10. Concordance du subjonctif imparfait, au pluriel (extrait)

2.3. Une fois que la sélection est faite, elle est communiquée (par le bouton CONTINUER ou RETOUR) à la fonction appelante, qui délivre un contexte (comme dans la figure 9), une concordance (figure 10, après un clic sur les cases "subjonctif imparfait" et "pluriel"), ou une liste (figure 11).

La fonction LISTE est pourvue également d'un bouton CODE qui renvoie à la page grammaticale et reçoit d'elle le code sélectionné. Avec le même exemple du subjonctif imparfait, on obtient la première ligne du tableau 11, où les effectifs les plus élevés sont le fait

de Proust (768 + 622 sur un total de 4061). La conversion en courbe mettrait en relief cette particularité stylistique que partagent aussi Marivaux, Georges Sand et Jules Verne (lequel, sur le tard, rêve de l'Académie et surveille sa plume).

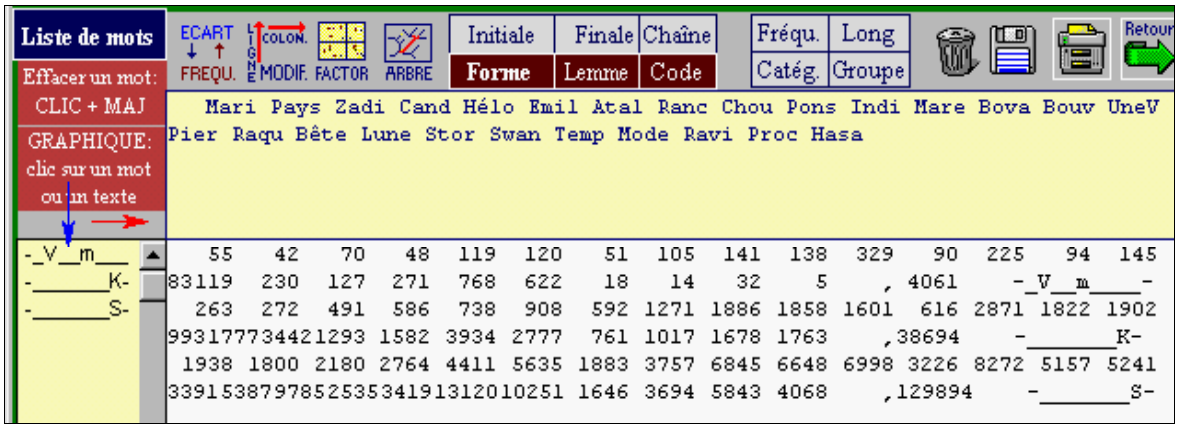

Figure 11. Relevé de quelques codes grammaticaux dans la page LISTE

3. Les fonctions grammaticales

3.1. Pour donner une idée des possibilités offertes par la lemmatisation, nous nous attacherons à la seconde ligne du même tableau 11 où le symbole K désigne les compléments circonstanciels de temps. Alors que les 13 auteurs du corpus sont répartis selon la chaîne chronologique, comment n'être pas frappé par la diagonale ascendante qui rend compte du progrès de cette structure dans la figure 12 ?

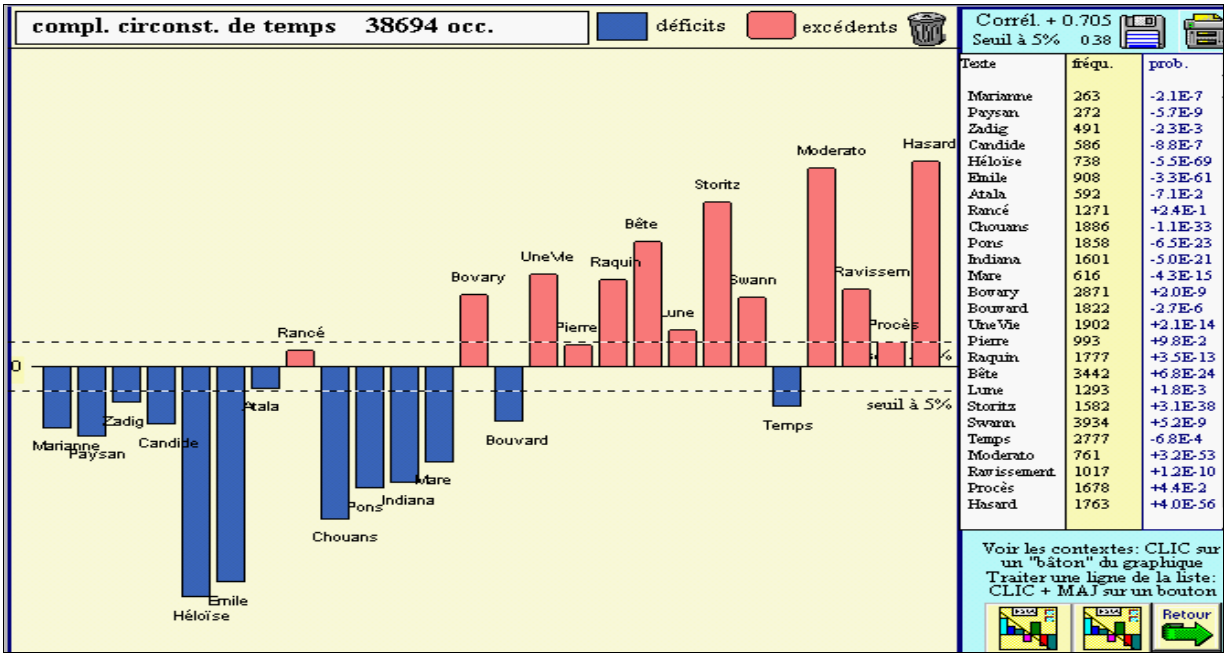

Figure 12. Le progrès des compléments circonstanciels de temps

3.2. Pour faire bonne mesure, mettons dans le même panier toutes les fonctions que distingue Cordial (il y en a 29), ou du moins toutes celles qui sont largement représentées dans notre corpus (il en reste 14). Reste à soumettre ce tableau à l'analyse factorielle (figure 13).

La fonction de base, assurée par le verbe, se situe à droite, avec ses acolytes immédiats que sont le sujet, le complément d'objet direct et le complément indirect. C'est là qu'on trouve les auteurs du XVIIIe siècle, mais aussi des représentants du XXe, Proust et Duras. Les

auteurs du XIXe se répartissent dans la moitié gauche, les premiers Chateaubriand et Balzac dans la partie basse, les seconds (Flaubert, Verne, Maupassant et Zola) dans la partie haute où Le Clézio les rejoint. Or les fonctions privilégiées durant cette période sont moins les fonctions de base que les extensions de ces fonctions, ce que la terminologie de Cordial désigne sous l'appellation "groupe sujet", "groupe objet direct", "groupe objet indirect". Cela signifie que la proposition s'étoffe ou s'alourdit d'épaisseurs adipeuses où les catégories nominales jouent un rôle majeur et qu'elle perd la simplicité nerveuse de la proposition classique. Les compléments circonstanciels prennent le relais au haut du graphique et participent pareillement à l'embonpoint de la proposition au moment où le réalisme succède au romantisme. Ces conclusions n'ont rien qui puisse surprendre. Elles demandent néanmoins à être étayées par d'autres études, qu'on souhaite plus larges et plus représentatives.

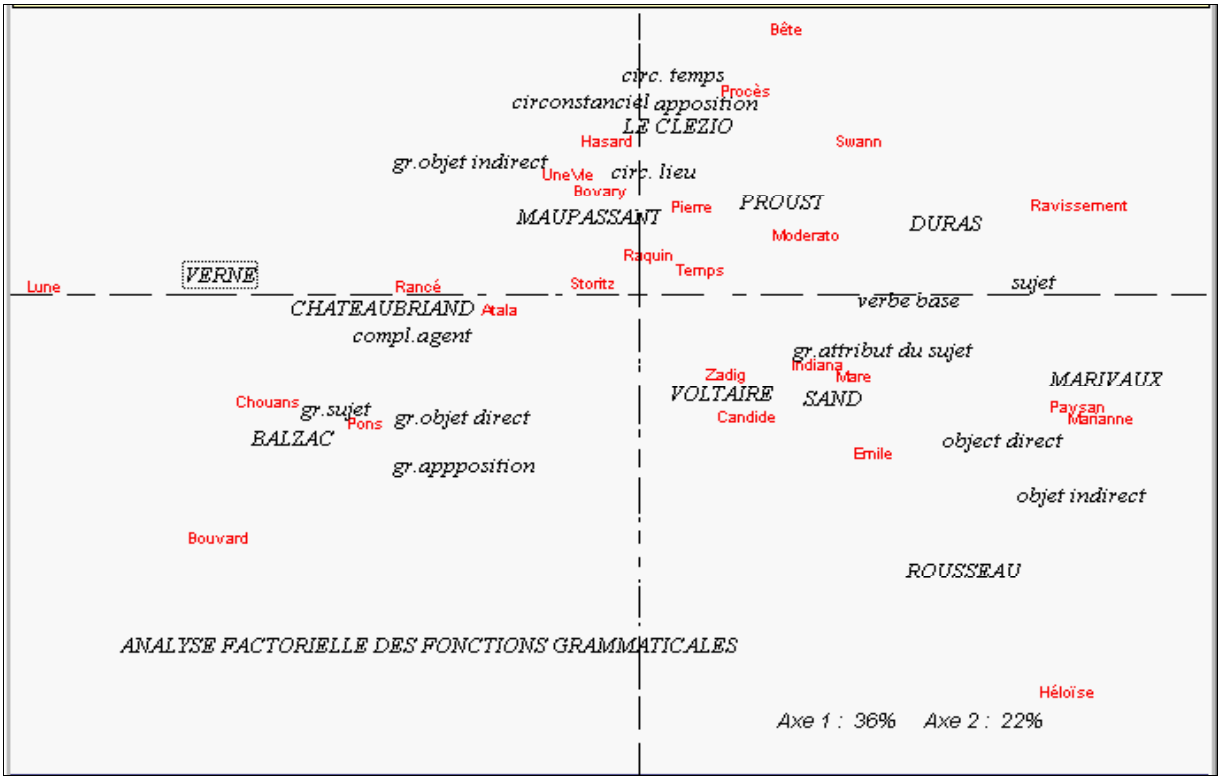

Figure 13. Analyse factorielle des fonctions grammaticales

4. Les parties du discours

Les fonctions grammaticales qu'on vient de relever dans Cordial n'offrent pas toutes les garanties de sécurité et de précision qu'on pourrait souhaiter. Il suffit de se plonger dans le détail d'une phrase un peu longue pour se rendre compte que l'analyse est souvent approximative, et qu'elle se laisse facilement abuser par le piège des incidentes, des incises, des emboîtements, et des constructions complexes, inhérentes à tout discours littéraire. Aussi bien Cordial se propose de corriger, non de traduire, et son analyse est soumise aux contraintes et aux limites de sa vocation première. Pour un produit qui sert tous les jours à un grand nombre d'utilisateurs, la contrainte la plus forte est d'aller vite (le logiciel peut traiter jusqu'à 12000 mots par seconde sur un Pentium 700). La structure profonde du discours y est donc abordée sans insistance et cela suffit le plus souvent pour le but proposé. La statistique n'a pas d'exigences fortes et peut aussi se contenter de résultats que la traduction automatique refuserait. Mais sa préférence va cependant aux données plus pures et plus sûres.

4.1. Or les parties du discours distinguées par Cordial sont nettement plus fiables que les fonctions grammaticales. C'est que leur relevé est plus facile. Pour beaucoup de mots qui ne souffrent pas de l'homographie, le codage est automatique et indépendant du contexte: la proposition du dictionnaire, étant unique, est immédiatement acceptée. Et là où deux catégories concurrentes ont à se partager les homographes (par exemple les cas très nombreux, du type *la marche/il marche,* où un substantif peut se confondre avec un verbe), une analyse de surface emporte souvent la décision. Là-dessus il est rare qu'on prenne en défaut le codage de Cordial, même dans les cas innombrables où *le, la l', les* articles doivent être distingués des pronoms personnels. On accordera donc plus de crédit au relevé des parties du discours qu'à celui des fonctions grammaticales.

On se contentera des catégories principales, telles qu'elles apparaissent, en jaune, dans la figure 9. En tenant compte des sous-catégories, du genre et du nombre, le tableau pourrait s'agrandir et se préciser à loisir. Cette première approche suffit à confirmer l'existence de lignes de force qui s'exercent dans le discours et qui opposent le substantif et le verbe comme les deux pôles d'un aimant. En prenant appui sur le relevé complet (dont le total atteint 2 millions d'observations), l'analyse arborée rend compte des alliances et apparentements qui lient entre elles les parties du discours. Visiblement, dans la figure 14, deux clans se sont formés: d'un côté le verbe (principal ou auxiliaire) tient sous sa coupe les pronoms, les adverbes, les conjonctions de subordination et les relatifs; de l'autre les substantifs (noms communs ou noms propres) règnent sans partage sur la valetaille des articles et des déterminants, et, avec moins de force, sur les prépositions, les numéraux et les adjectifs. Entre les deux camps hésitent les coordinations et les interjections.

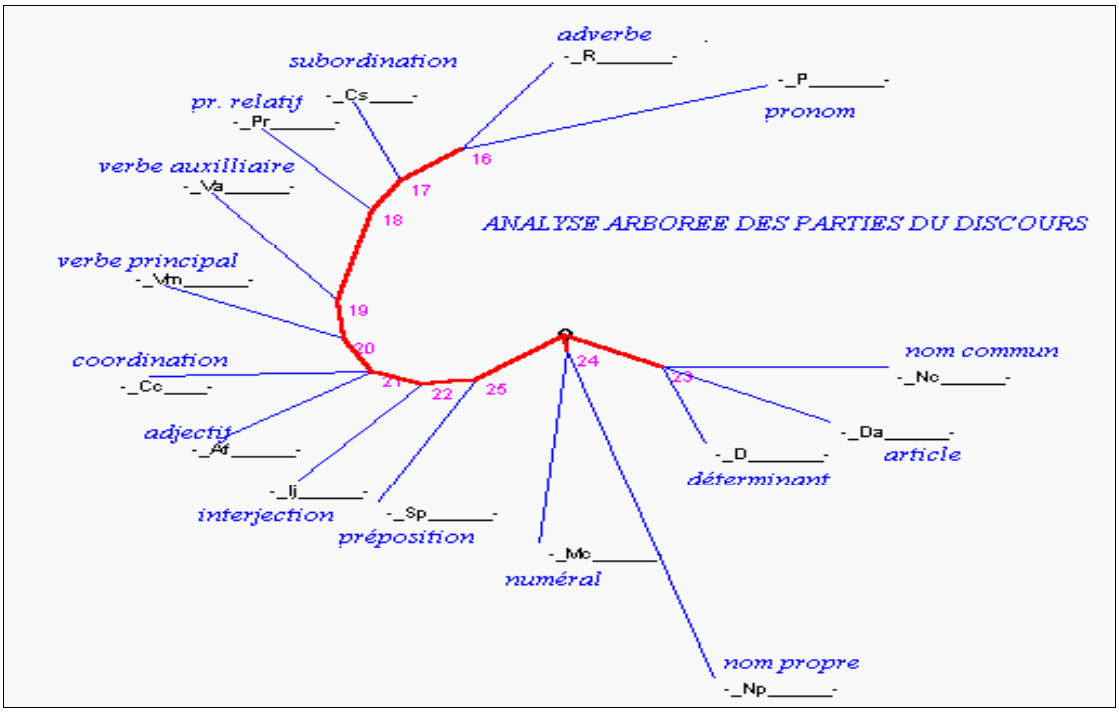

Figure 14. Analyse arborée des parties du discours

4.2. Ce n'est pas la première fois que nous rencontrons cette bipolarisation du discours et nous avons pu l'observer dans de nombreux corpus. La puissance et la précision de Cordial permettent cependant d'affiner et de confirmer les observations antérieures. Bien entendu, dans chaque phrase la cohabitation du verbe et du substantif est inévitable. Mais au niveau d'un texte tout entier, la préférence statistique peut être donnée à l'un ou à l'autre, ou à quelque autre catégorie. La question se pose de savoir si deux textes d'un même auteur font les mêmes

choix et si, au niveau des auteurs, il y a la même cohérence et la même lisibilité qu'on vient de constater dans les catégories. À partir du même tableau des données, le programme d'analyse arborée, orienté différemment (sur les colonnes et non plus sur les lignes), propose le graphe 15, dont l'interprétation est aisée si l'on adopte le principe: qui se ressemble s'assemble. On constate que généralement les deux textes d'un même auteur sont voisins sur le graphe, ce qui signifie que le dosage des parties du discours y est semblable. C'est le cas de Marivaux, de Rousseau, de Sand et de Proust qui partagent la même branche du graphe. C'est le cas aussi , mais avec moins de cohésion, de la branche opposée où se rejoignent Chateaubriand, Balzac et Flaubert. Dans l'entredeux flottent certains écrivains qui semblent n'avoir pas de parti pris dans cette affaire (Voltaire, Zola, Duras, Le Clézio) ou qui manifestent des tendances contradictoires (Verne).

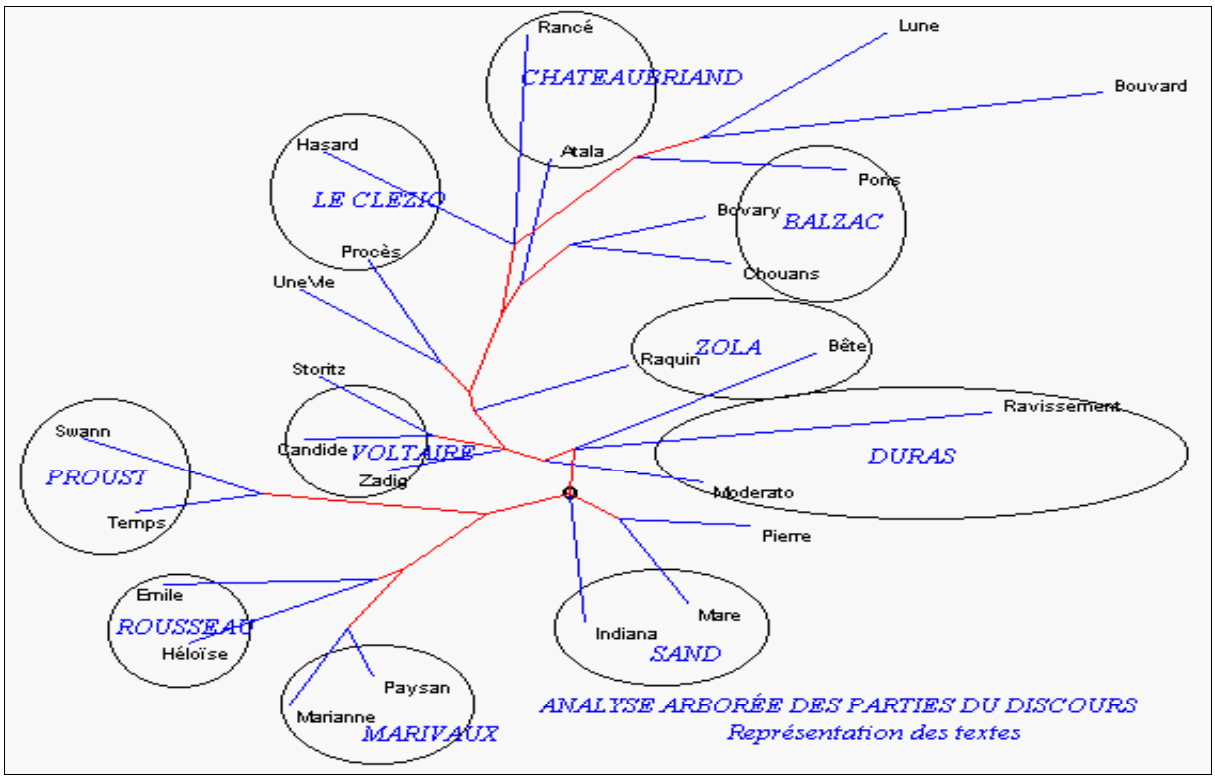

Figure 15. Analyse arborée des parties du discours. Représentation des textes

4.3. Reste à superposer ces deux graphes, pour comprendre pleinement non seulement le jeu des alliances et des oppositions entre catégories ou entre écrivains, mais celui qui ordonne tout ensemble les catégories et les écrivains. On voudrait que l'analyse nous dise les relations de préférence ou de réticence que tel ou tel écrivain peut avoir avec telle ou telle partie du discours. C'est le rôle de l'analyse factorielle (de correspondance), dont le résultat est reproduit dans la figure 16. L'échiquier répartit comme on s'y attendait les deux clans: à droite le verbe et ses auxiliaires, à gauche le nom et sa suite, l'adjectif hésitant à prendre parti. Sur cet échiquier les textes sont invités à prendre place. Ceux de Marivaux, Rousseau, Sand, Prost et Duras choisissent le verbe, ceux de Chateaubriand, Balzac et Flaubert se portent du côté du nom, et, moins nettement aussi (car la position diffère d'un texte à l'autre) ,Verne et Le Clézio. Voltaire est indifférent au centre et Maupassant à cheval sur la ligne de partage. À quel effet de style ou à quelle propriété du genre faut-il attribuer ces choix? Le dialogue sollicite plus souvent les verbes, la description fait plutôt appel aux catégories nominales et le récit peut mêler diversement ces ingrédients.

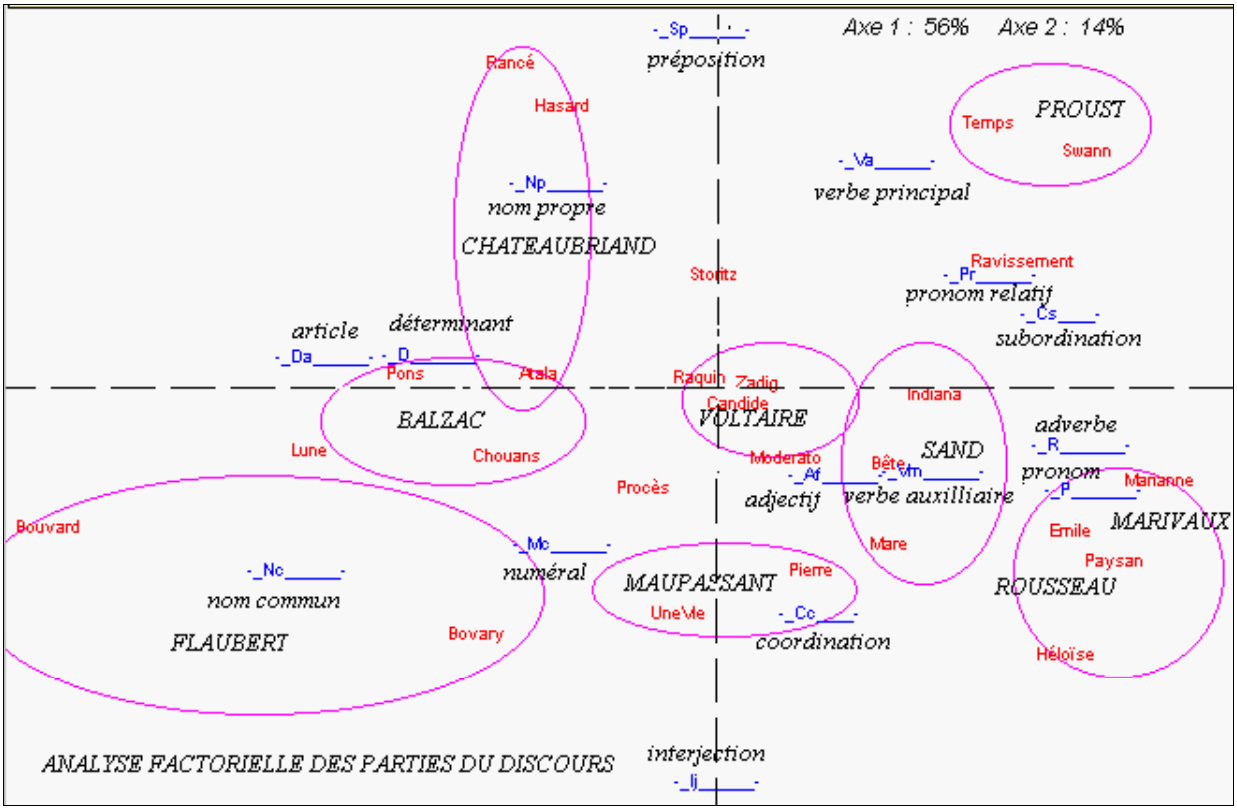

Figure 16. Analyse factorielle des parties du discours

L'étiquetage des données permet d'étendre l'enquête aux temps, aux personnes et aux modes verbaux. Mais la place nous manque pour en rendre compte. On constate que le mode n'est pas la pierre de touche qui puisse servir à classer les textes et les styles. Tous les modes restent groupés, à peu de distance les uns des autres. Les personnes sont plus excentriques, et, comme elles sont trois, leur constellation prend la forme d'un y, la branche la plus longue étant le fait de la troisième, contre laquelle s'unissent les deux autres. Mais la voix la plus forte appartient au temps; c'est elle qui impose sa loi au récit, en le sommant de choisir entre le présent et le passé.

En conclusion, une enquête sur la population des mots jouit de gros avantages si l'on a affaire à un état policé où les individus ont été recensés et possèdent une carte d'identité. C'est le cas des lemmes. L'étude prend l'aspect alors d'une recherche sociologique. En croisant la fonction, la catégorie, le temps , le genre, le nombre, etc., on peut suivre la même démarche que les autres sciences humaines, qui mettent en relation, à partir de leurs observations, la catégorie socioprofessionnelle, l'âge, le salaire, les opinions politiques, le niveau culturel, la mortalité, la fécondité, etc. Certains pourront regretter les formes brutes, dont la matérialité opaque pouvait receler quelque mystère, et renâcler devant un lemme blême, vidé de son sang, et réduit à un ensemble de traits abstraits, que nous nous sommes efforcé de circonscrire dans l'étude qui précède. Celle qui lui succède - elle dépasse le cadre du présent exposé et sera publiée ailleurs - est plus ambitieuse encore : elle met en parallèle non seulement la forme, le lemme et le code grammatical, mais aussi la structure syntaxique et le code sémantique. Et – faut-il s'en désoler ou s'en féliciter ? – tout cela converge.