HAL open science

Le lemme comme on l'aime

Étienne Brunet

▶ To cite this version:

Étienne Brunet. Le lemme comme on l'aime. JADT 2002, A. Morin, P. Sébillot, Mar 2002, Saint-Malo, France. pp.221-232. hal-01790696

HAL Id: hal-01790696
https://hal.science/hal-01790696v1

Submitted on 28 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Le lemme comme on l'aime

Etienne Brunet, BCL(CNRS), Université de Nice Côte d'Azur

1. Lemmatisation

1 – Dans les travaux de linguistique quantitative, la prudence a souvent choisi le même camp que la paresse. En adoptant un profil bas, elle avouait l'impureté des données et faisait 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. On pourra s'en faire une idée à l'aide du tableau 1 qui fournit la liste, détaillée et simplifiée, des codes rencontrés.

Tableau 1. Les codes grammaticaux distingués dans *Winbrill*
(les codes de regroupement sont en colonne 2)

C:\HYPERBAS\HYPERLEM.EXE

étiquette d'origine	étiquette simplifiée	signification	étiquette d'origine	étiquette simplifiée	signification
ABR	_ABR	abréviation	----------	_S	substantif ----------
---------	_AV	avoir----------	SBC		nom commun
ACJ		AVOIR conjugué	SBP		nom propre
ANCFF		AVOIR infinitif	SBP?		nom propre probable
ANCNT		AVOIR forme en -ant	---------	_DTN	déterminant----------
APAR		AVOIR p.passé après AVOIR	DTN		déterminant
---------	_E	être ----------	DTC		déterminant contracté
ECJ		ETRE conjugué	---------	_PR	pronom_adj.----------
ENCFF		ETRE infinitif	PRV		pronom supporté par le verbe
ENCNT		ETRE forme en -ant	PRO		autre pronom (ou adj)
EPAR		ETRE p.passé après AVOIR	PRV_$$		pronom indéterminé (en, y, se)
--------	_V	verbe ----------	----------	_REL	----------
VCJ		verbe conjugué	REL		relatif (adj., pron. ou adv.)
VNCFF		verbe infinitif	---------	_PREP	----------
VNCT		verbe forme en -ant	PREP		préposition
VPAR		verbe p. passé après AVOIR	----------	_SUB	----------
--------	_ADV	---------- ----------	SUB		subordonnant
ADV		adverbe	SUB$		code par défaut pour QUE
----------	_A	adjectif----------	----------		----------
ADJ		adjectif (participe passé exclu)	INJ		interjection ou onomatopée
ADJ1PAR		part.passé adjectif derrière ETRE	PUL		particule
ADJ2PAR		par.passé adjectif NON derrière ÊTRE	SYM		symbole ou signe mathém.
		CAR	cardinal (chiffres ou lettres)		
CLIQUER DANS LE CHAMP POUR			FGW		mot étranger
LE FAIRE DISPARAITRE			sg		singulier
			pl		pluriel

CLIQUER...dans le champ pour le faire disparaître

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.

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 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 *a 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 (tableau 2).

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).

Tableau 2. La composition du corpus

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				

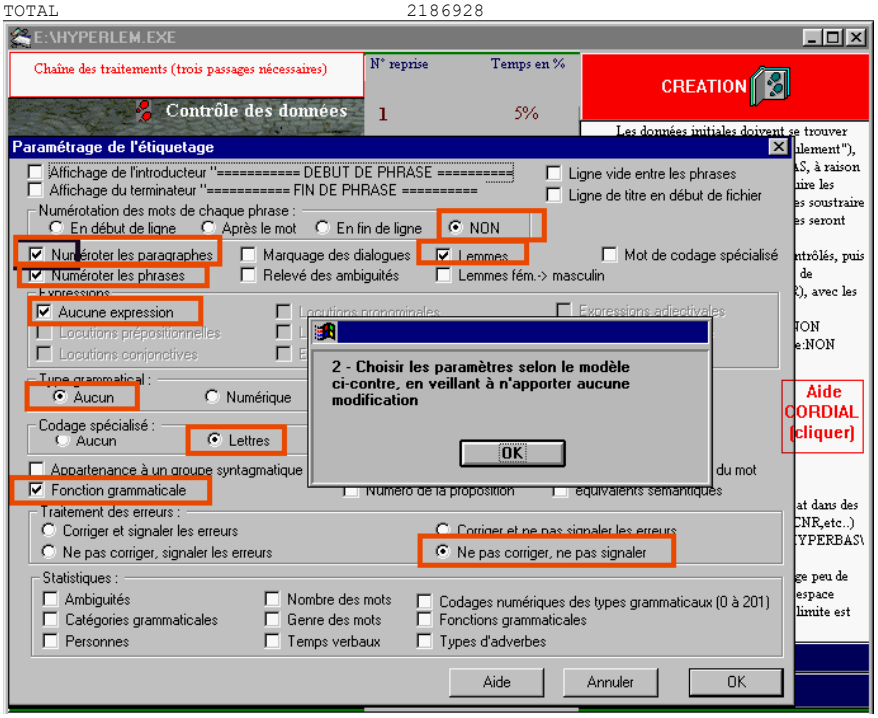

Figure 3. Les options de l'analyse dans *Cordial*

3 – La composition du corpus est détaillée ci-dessus (tableau 2). 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 3.

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 4, 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	Syntag.	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 4. Un fichier lemmatisé par *Cordial*

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 5 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 CABINET) distingue l'article du pronom codé 5 (dans ON L'ABANDONNAIT). 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*.

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 _AFP__P_N, soit *adjectif qualificatif, positif, au pluriel, dans un groupe en apposition*) on obtient successivement en vidéo inverse tous les adjectifs qui répondent à ce codage dans le champ de gauche.

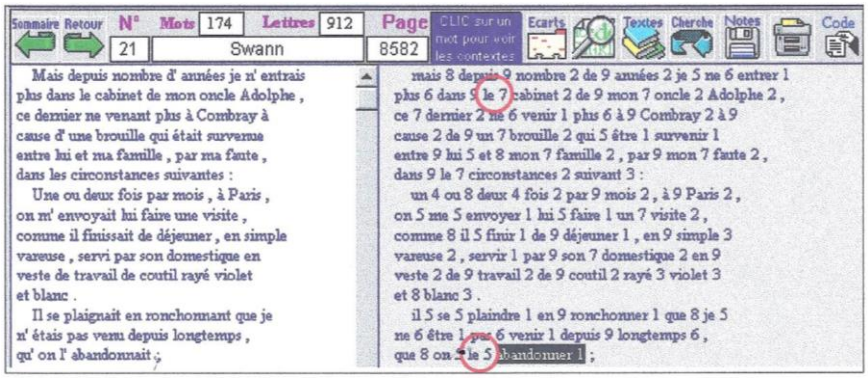

Figure 5. L'alignement forme-lemme

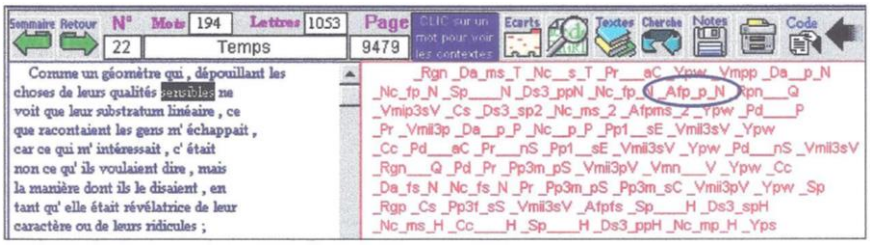

Figure 6. L'alignement forme-code

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 7) 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.

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 7, relatif à l'adjectif qualificatif, au pluriel, dans un groupe en apposition (c'est le même exemple que celui de la figure 6). Pour faciliter les recherches on a joint la fonction au code grammatical en dernière position (ici la lettre N pour le groupe en apposition).

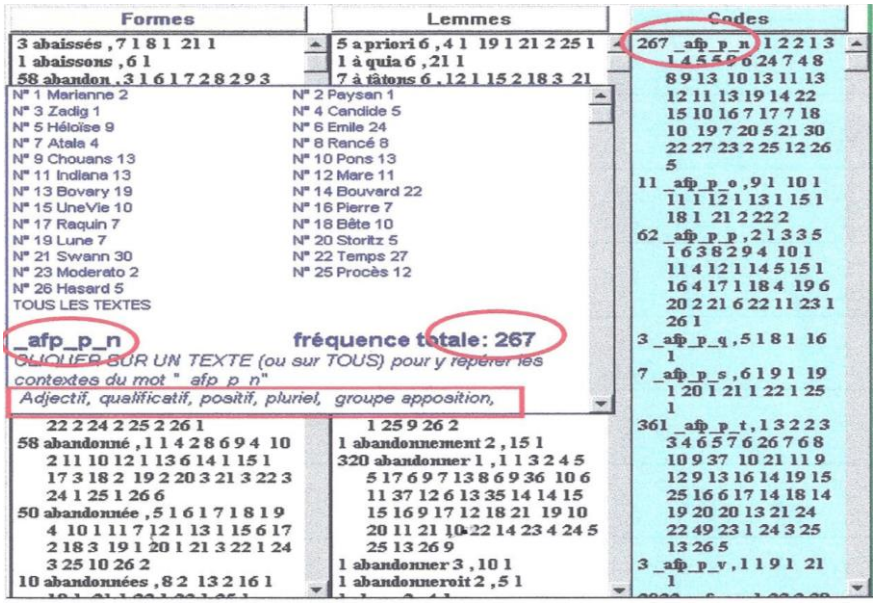

Figure 7. Les trois index issus de *Cordial*

2. Exploitation

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 8 et dont le résultat est consigné dans la figure 9.

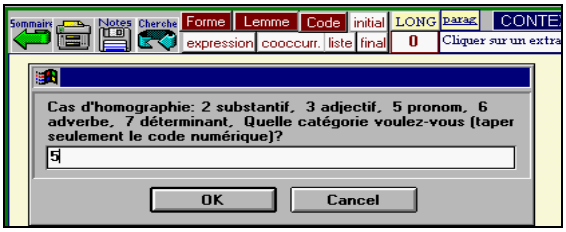

Figure 8. La désignation des homographes (ici TOUT)

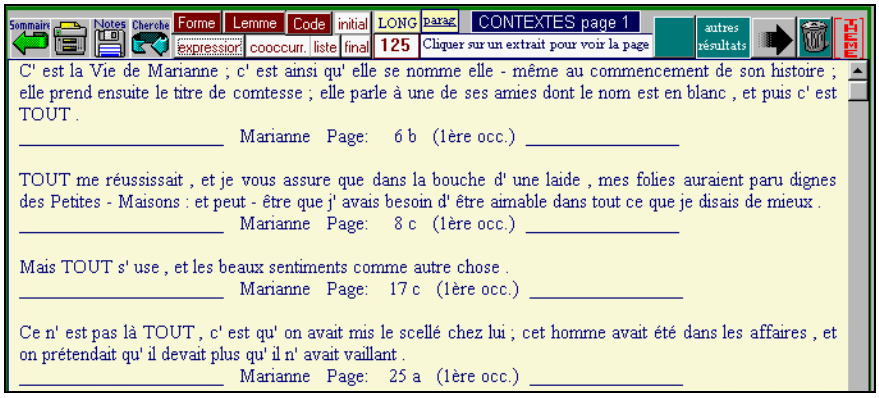

Figure 9. Les contextes de TOUT pronom

2 – Lorsqu'on a affaire aux codes grammaticaux, on est renvoyé à une page spéciale (figure 10) 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.

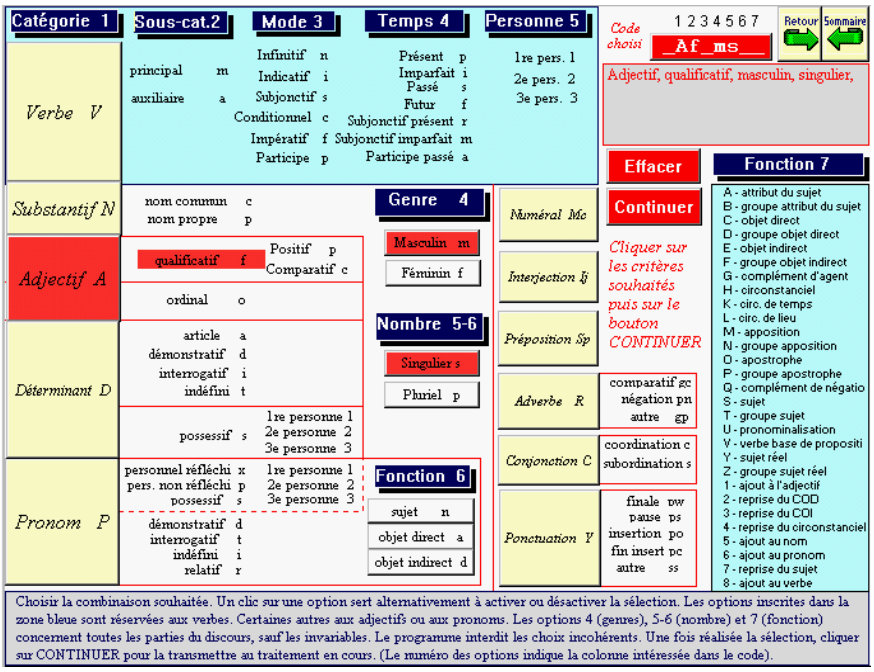

Figure 10. Le choix d'un code grammatical

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 11, après un clic sur les cases « subjonctif imparfait » et « pluriel »), ou une LISTE (figure 12).

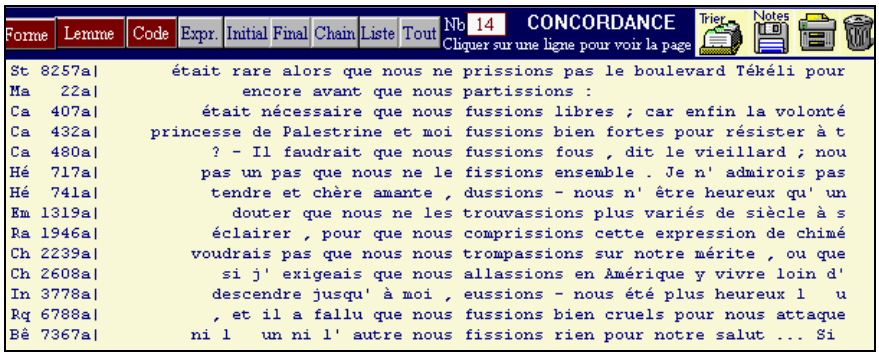

Figure 11. CONCORDANCE du subjonctif imparfait, au pluriel (extrait)

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 12, 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).

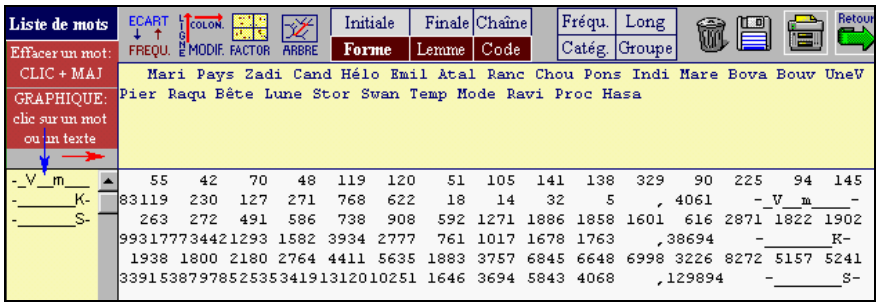

Figure 12. Relevé de quelques codes grammaticaux dans la page LISTE

3. Les fonctions grammaticales

1 – Pour donner une idée des possibilités offertes par la lemmatisation, nous nous attacherons à la seconde ligne du même tableau 12 où le symbole κ 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 13 ? On a rarement eu l'occasion jusqu'ici d'appliquer le nombre à de telles structures et c'est un champ nouveau qui s'ouvre à la statistique linguistique. Avant de s'engager dans l'interprétation, il est prudent de faire le relevé des autres fonctions grammaticales et par exemple de comparer à la courbe précédente celle des circonstanciels de lieu.

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 14).

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.

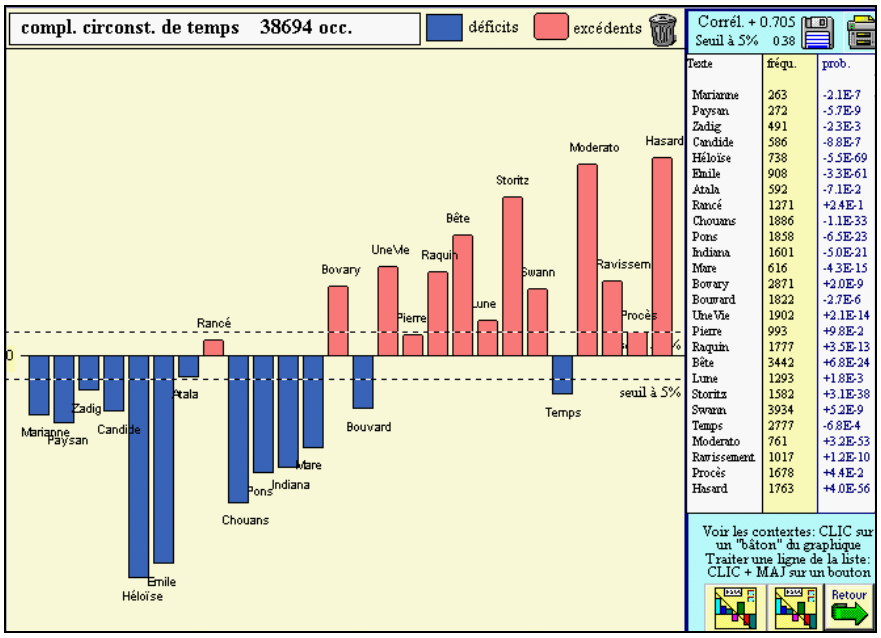

Figure 13. Le progrès des compléments circonstanciels de temps

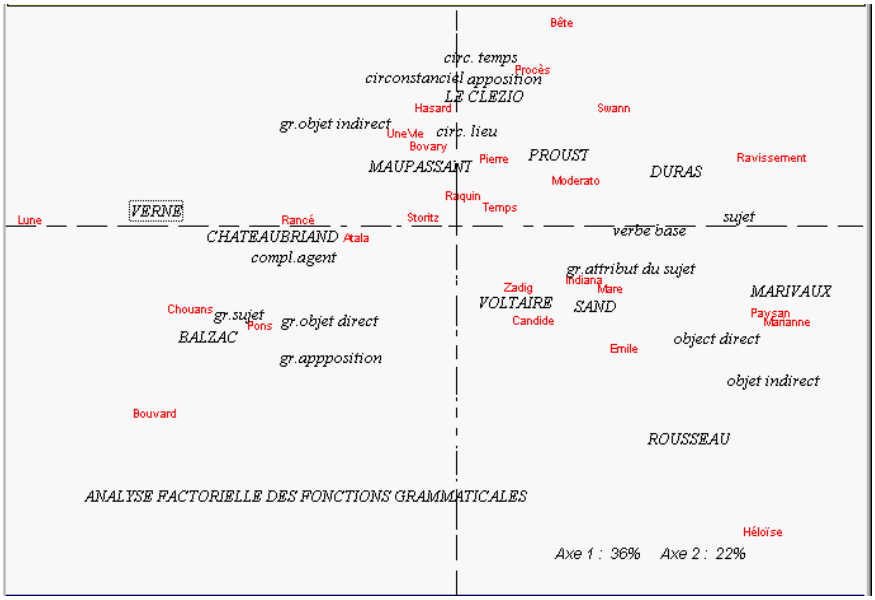

Figure 14. Analyse factorielle des fonctions grammaticales

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.

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 12 000 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.

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.

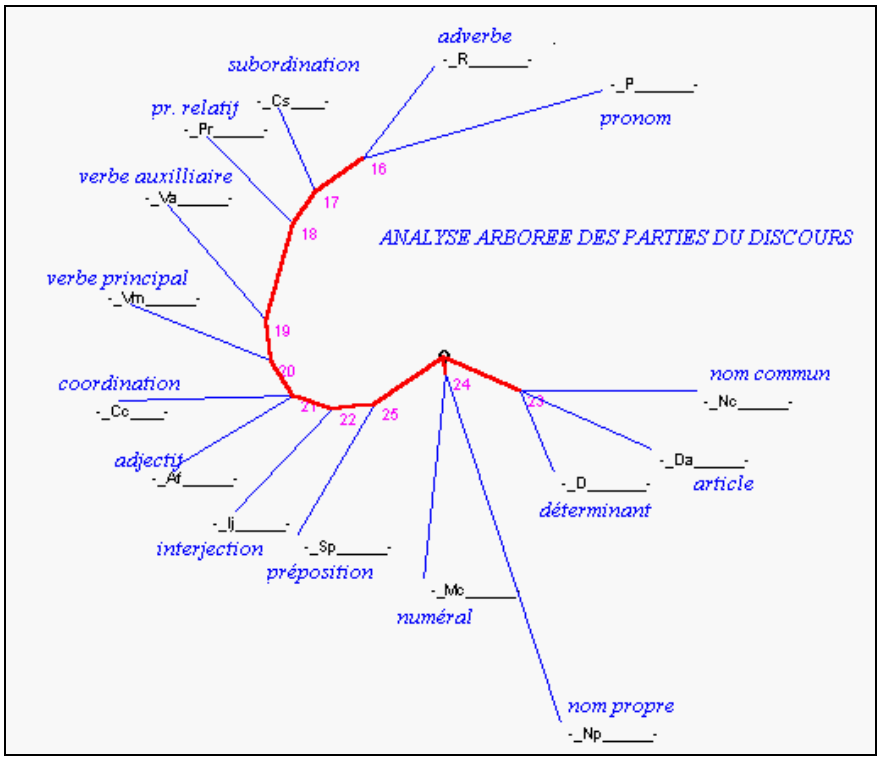

Figure 15. Analyse arborée des parties du discours

On se contentera des catégories principales, telles qu'elles apparaissent, en jaune, dans la figure 10. 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 15, 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.

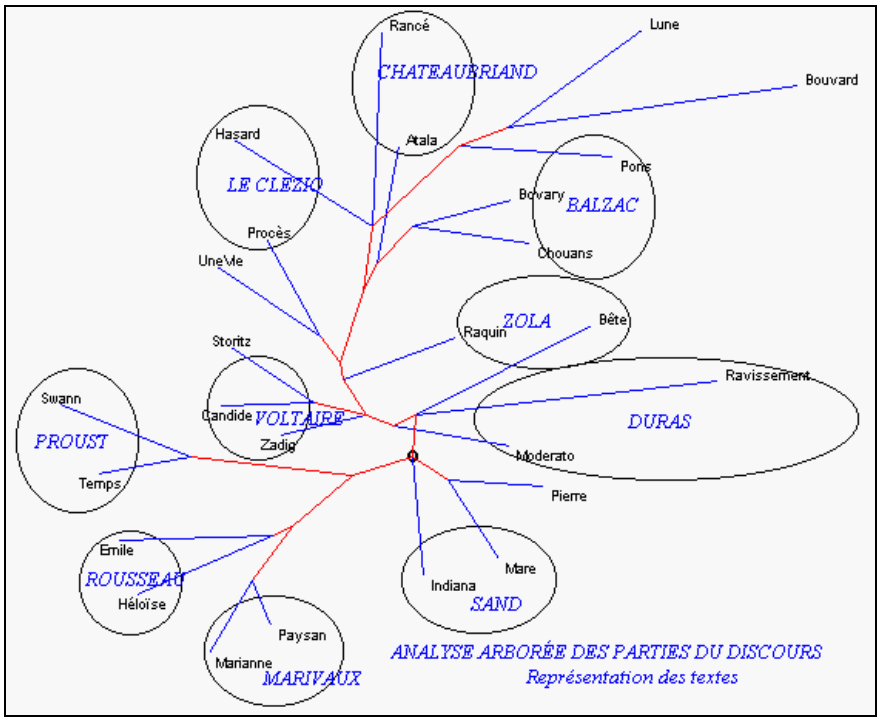

Figure 16. Analyse arborée des parties du discours. Représentation des textes

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 16, 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).

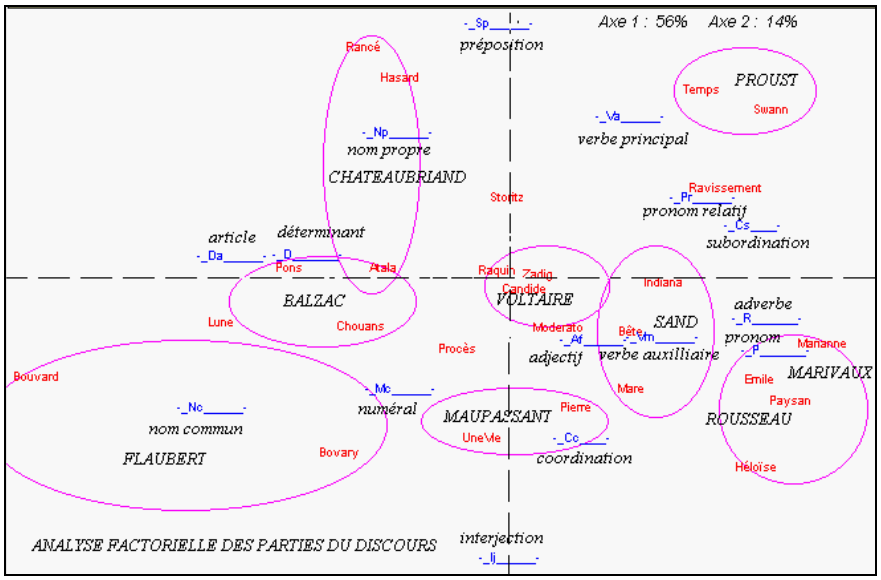

Figure 17. Analyse factorielle des parties du discours

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 17. 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, Proust 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.

5. Les temps, les modes et les personnes

1 – Il n'est pas certain que la préférence donnée par un écrivain à une catégorie grammaticale soit un choix volontaire et conscient. On a vu des manifestes littéraires s'en prendre aux principes de la composition, aux lois des genres, aux règles de la ponctuation, à l'impureté du vocabulaire (comme Malherbe), ou à son étroitesse (comme Hugo). Il est plus rare qu'on vise la syntaxe ou le dosage des parties du discours. « Paix à la syntaxe », disait Hugo. Mais lorsqu'il s'agit de l'emploi des verbes, on a tout lieu de penser qu'un écrivain y porte attention. Le choix qu'il fait du passé ou du présent, de la première ou de la troisième personne, a des conséquences importantes pour la conduite du récit et une telle décision ne saurait être inconsciente. Le système verbal s'étageant sur plusieurs plans : le mode, le temps et la personne (sans compter le nombre, l'aspect et d'autres paramètres), on pourrait isoler successivement les trois plans principaux (les seuls que relève *Cordial*), ou bien les croiser et, par exemple, consacrer une ligne du tableau à la troisième personne du pluriel du présent de l'indicatif des verbes auxiliaires (croisement de 5 variables). Pour une première approche il nous a paru prudent de s'en tenir aux grandes divisions, sans croisement, mais sans exclusive. En étudiant ensemble, comme variables indépendantes, les modes, les temps et les personnes, on se donne le moyen de repérer lequel de ces trois paramètres est le plus discriminant, mais aussi quelle interaction s'exerce entre les uns et les autres.

2 – Une première réponse est dans le graphique 18. On y voit 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 au centre du graphe, à peu de distance les uns des autres, à l'exception de l'impératif qui s'écarte vers le haut, ayant partie liée avec la deuxième personne, et du participe qui s'éloigne vers le bas en s'associant aux auxiliaires pour constituer les

temps composés. 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, au bas du graphe, contre laquelle s'unissent les deux autres, au haut du graphe. 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. La tension la plus intense (sur le graphe la distance la plus longue) est en effet celle qui oppose le présent (en haut) à l'imparfait et au passé simple (en bas). Le futur accompagne le présent, tandis que les temps composés, principalement le passé composé, rejoignent l'imparfait. Les trois critères du verbe ne sont pas vraiment indépendants. Si le mode maintient sa neutralité dans la partie engagée autour des temps (mis à part l'impératif et le participe), la personne exprime clairement ses préférences : la première pour le présent, la troisième pour le passé.

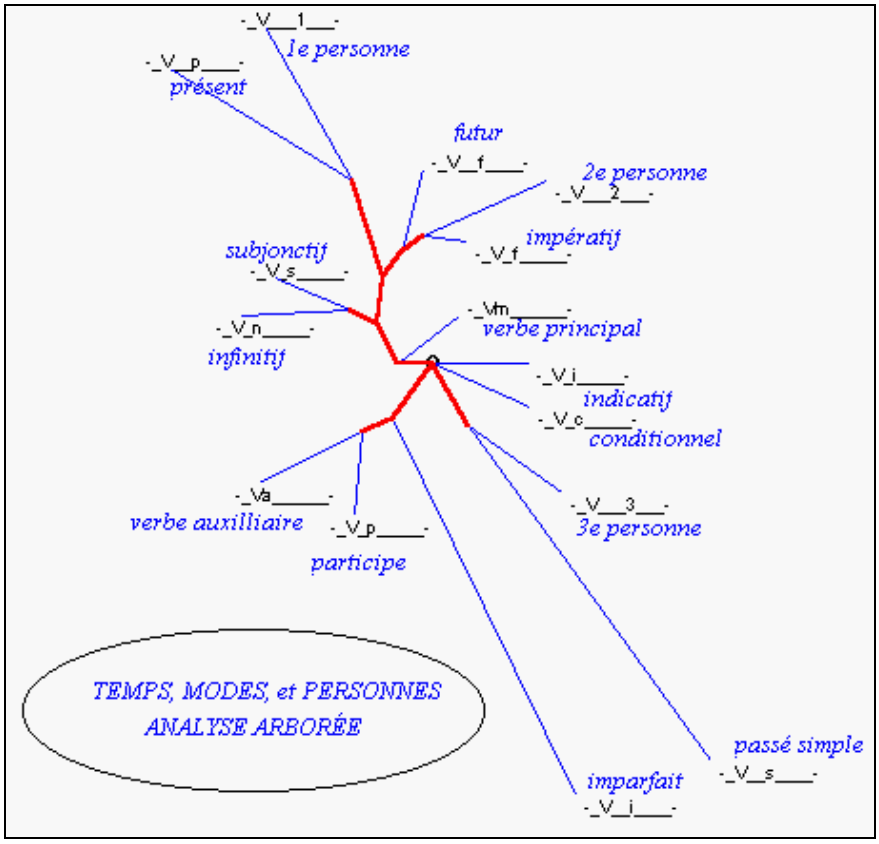

Figure 18. Analyse arborée des temps, des modes et des personnes

3 – Les choix étant ainsi offerts, comment réagissent les textes ? Comme précédemment, ils vont par couples, les textes ayant des choix solidaires s'ils ont le même père. Cette fraternité est particulièrement étroite s'il s'agit de Marivaux, Rousseau, Voltaire, Sand, Balzac, Flaubert, Maupassant, Zola ou Proust. Mais on distinguera, en les enveloppant dans un cercle sur le graphique 19, les couples qui affirment hautement leurs préférences et leurs exclusives communes, et ceux que le vote laisse indifférents et qui se rapprochent de l'origine des axes. Comme précédemment, Voltaire est parmi les abstentionnistes, et Verne parmi les hésitants déchirés.

S'il n'y avait l'exception de Duras et l'indécision de Verne, on pourrait admettre que la chronologie polarise les résultats : tous les écrivains antérieurs à 1850 se situent à droite, dans le présent, en compagnie des deux premières personnes. Sans doute la part du dialogue y est-elle plus importante. Mais ce n'est sans doute pas la seule raison. À gauche, dans la zone opposée, c'est, à partir de Flaubert, le règne de la troisième personne et du passé, particulièrement de l'imparfait que Proust admirait tant chez Flaubert. On peut être sensible à cette continuité stylistique qui prend naissance chez Flaubert et se maintient jusqu'au nouveau roman.

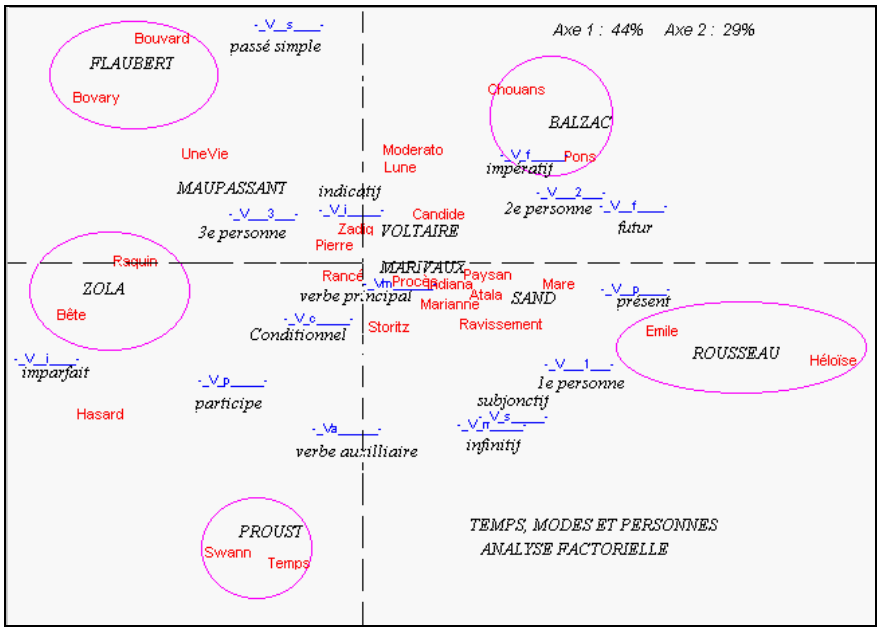

Figure 19. Analyse factorielle des modes, temps et personnes

Conclusion

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 présente. Reste une démarche plus ambitieuse (et qui dépasse le cadre du présent exposé) : mettre en parallèle non seulement la forme, le lemme et le code grammatical, mais aussi la structure syntaxique (ou combinaison de codes) et la variation thématique (à partir d'un thesaurus sémantique). En appliquant les mêmes méthodes à ces nouveaux objets, les résultats montrent – faut-il s'en désoler ou s'en féliciter ? – que tout cela converge[1].

1. NDÉ : voir en particulier dans le tome II le chapitre 9, « Qui lemmatise dilemme attise » (2000a), p. 165-184, contemporain du présent chapitre en termes de rédaction et présentant les premiers résultats mentionnés ici ; et, toujours dans le tome II, le chapitre 14, « Le corpus conçu comme une boule » (2007a), p. 279-292, qui, quelques années plus tard, se consacre complètement et explicitement au sujet.