



Quand les textes sont lemmatisés...

Étienne Brunet

► To cite this version:

Étienne Brunet. Quand les textes sont lemmatisés.... Inge Bartning. Mélanges publiés en hommage à Gunnel Engwall, Almqvist& Wiksell, pp.105-116, 2002, 91-22-01982-0. hal-01571423

HAL Id: hal-01571423

<https://hal.science/hal-01571423>

Submitted on 2 Aug 2017

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.

Étienne Brunet, Université de Nice

QUAND LES TEXTES SONT LEMMATISÉS ...

Le débat sur la lemmatisation¹ a commencé il y a trente ans, à l'époque où Gunnell Engwall entreprenait sa grande enquête sur le « Vocabulaire du roman français (1962–1968) » et où les chercheurs de Saint Cloud contestaient les recommandations de Charles Muller. La querelle est à ce jour apaisée et la guerre de trente ans ne sera pas une guerre de cent ans. Aussi bien des travaux estimables ont été publiés qui suivent l'une ou l'autre option (et parfois les deux). Ceux qui s'en tiennent à la graphie sont sans doute les plus nombreux, non seulement parce que la préparation et le traitement y sont plus aisés, mais aussi parce que les résultats permettent plus facilement la comparaison, l'intervention humaine dans les données étant réduite au minimum. Ceux qui veulent traiter un produit raffiné et s'attachent au lemme (Dominique Labbé est le chef de file de cette cohorte héroïque qui a compté Gunnell Engwall dans ses rangs) s'échelonnent sur le long chemin qui va du réel à l'idéal. Deux obstacles principaux se dressent sur leur chemin, dont l'un tient aux traitements des expressions ou mots composés, l'autre aux homographes.

Même ceux qui s'abstiennent de lemmatiser ont des difficultés de définition. On a dénoncé depuis longtemps l'ambiguïté du mot *mot*. Mais il en va ainsi de la graphie. Il ne suffit pas de décréter que la graphie est l'espace imprimé entre deux séparateurs. Encore faut-il dresser la liste des séparateurs et prendre position sur les cas ambigus où un même caractère peut ou non jouer ce rôle de séparateur (par exemple le point, l'apostrophe et le trait d'union). Mais le lemme est plus difficile encore à délimiter. Il ne suffit pas d'invoquer la garantie d'un dictionnaire pour régler le problème de la nomenclature. Une fois réglés les regroupements qui réduisent les variations du genre, du nombre, du temps, du mode et de la personne, que fera-t-on des syntagmes ou expressions? Les dictionnaires les traitent à l'intérieur de l'article consacré à l'un de leurs constituants mais leur accordent rarement une entrée indépendante. Le Petit Robert fait un sort particulier aux deux exemples canoniques qu'on cite toujours: le chemin de fer et la pomme de terre. Mais il refuse ce privilège à la pomme d'Adam et au chemin de croix. Et Alain Rey s'en explique franchement dans l'édition de 1977: « il nous a paru plus raisonnable de donner à l'ordre alphabétique ces vrais composés que sont pomme de terre, chemin de fer ou point de vue. De tels procédés ne peuvent être appliqués systématiquement dans un dictionnaire, sous peine d'augmenter excessivement le nombre des « mots » traités et le volume de l'ouvrage. Cependant, quand certaines expressions ou formes verbales, moins importantes, constituaient de véritables mots (on dit alors qu'elles sont lexicalisées), elles ont été présentées en capitales dans les

articles où elles sont incluses, pour attirer l'attention sur leur autonomie. Ainsi : EXCÈS DE POUVOIR (à l'article excès), la locution À L'EXCEPTION DE (sous *exception*), la locution COMME IL FAUT, qui s'emploie comme un adverbe ou comme un adjectif (à *falloir*). » On pourrait certes relever tous ces syntagmes pourvus de la majuscule et les ajouter à la nomenclature. Mais il en est tant d'autres qui n'ont pas cette marque de reconnaissance et dont on ne sait s'il s'agit d'une suite de mots intangible ou seulement d'une suite de mots fréquente mais modifiable. On a affaire en réalité à un continuum qui va de l'expression figée et lexicalisée à la variation libre. La frontière entre l'une et l'autre est mouvante selon les époques, selon les écrivains (beaucoup prennent plaisir à casser ou modifier les expressions consacrées) et selon les dictionnaires. Et même à l'intérieur d'un même dictionnaire les décisions sont parfois peu cohérentes. Même les exemples relevés par A. Rey ne sont pas à l'abri de la contestation: il reste un peu de liberté dans les syntagmes les plus fermés: on dira parfois *un excès manifeste de pouvoir* au lieu d'un *excès de pouvoir manifeste*, et l'adjectif *notable* peut s'introduire dans l'expression *à l'exception de*. Et que dire de *comme il faut*, dont l'emploi adjectival ou adverbial ne peut être reconnu que dans le contexte? Et c'est ici que réside la principale difficulté de la lemmatisation. C'est le contexte qui fixe la nature et les limites de la chaîne graphique. Le recours aux dictionnaires ne résout pas le problème de la nomenclature. Encore moins celui des homographes. S'il est difficile de regrouper ce qui doit l'être, il est plus délicat encore de séparer ce qui doit l'être. Ici aussi les dictionnaires électroniques sont utiles, mais insuffisants. Ils signalent l'ambiguïté d'une forme mais se gardent bien de décider laquelle des analyses possibles est exacte dans le cas précis qui leur est soumis. Même si le dictionnaire consulté est doté de pondérations ou de pourcentages, on ne peut toujours trancher en faveur du sens majoritaire ou, solution pire encore, donner raison à la première des analyses proposées. Ceux qui se dispensent de toute analyse et ne considèrent que les formes brutes peuvent invoquer là aussi la difficile standardisation dans cette opération de décantation, lors même qu'elle est menée dans le texte. Car la filtration peut être plus ou moins fine, et porter sur divers critères: la catégorie du mot, sa fonction et son sens, ce qui exige qu'on sache reconnaître ces mêmes critères dans les mots voisins. Comme un simple codage grammatical ne suffit pas à distinguer *voler 1* (dérober) de *voler 2* (dans les airs), l'analyse ne peut se contenter de la seule syntaxe et doit s'engager (mais jusqu'où?) dans la voie sémantique et pragmatique. On court le risque de la babélisation, ce que reconnaît Benoît Habert quand il évoque « l'inévitable éparpillement des étiquetages »².

Pour qui veut toucher du doigt la complexité des opérations nécessaires au toilettage d'un texte qu'on veut véritablement désambiguïser et lemmatiser, le meilleur modèle est fourni par Gunnel Engwall. Voici comment nous rendons compte de son ouvrage, il y a quinze ans³ :

« L'homographie est la terreur des lexicomètres. La plupart prennent la fuite, sans gloire, ou cherchent des faux-fuyants moins glorieux encore. Gunnel Engwall a osé affronter l'hydre aux mille têtes bourgeonnantes. Manuellement. Méthodiquement. En examinant, dans leur contexte (dans la concordance), toutes les occurrences suspectes où le verbe et le substantif ont le même vêtement unisexe (la danse et il danse), où l'adjectif se pare des plumes du nom (le bon, la bonne), et où, plus généralement, la fonction (et donc la nature) diffère d'un emploi à l'autre—ce qui se produit précisément dans beaucoup des mots fonctionnels. Si les grammairiens étaient équitables ou pitoyables, ils devraient interdire aux mots le cumul abusif des fonctions. Si encore G. Engwall s'était contentée de catégories grossières, si en particulier elle avait admis un grand réceptacle pour recueillir indifféremment tous les mots grammaticaux, la tâche eût été moins rude. Mais, héroïque jusqu'au bout, l'auteur a passé au tamis les 10146 occurrences de *le*, pour distinguer l'article du pronom, les 7987 emplois de *les*, les 5067 *est* (pour y relever quatre fois le point cardinal), les 5031 *pas* (où la négation cède 146 fois le pas au substantif), les 3235 *ce*, les 2942 *s'*, les 1547 *si*, et les 2583 *sur* (en n'y trouvant aucun adjectif, aucun fruit *sur*). Aucune enquête de cette taille (le corpus rassemble 500000 mots) n'a jamais atteint tant de minutie et de précision philologique. Ici il faut s'arrêter pour admirer, et s'agenouiller. Voilà enfin ouverts des domaines jusqu'ici interdits par l'homographie: celui de l'article défini, qui compte 32708 occurrences (dont 5995 *l'*, 10912 *la*, 8623 *le*, 7129 *les*, 1 *el*, 13 *L'*, 16 *La*, 13 *Le* et 6 *Les*), et celui de la troisième personne. Voilà enfin des bases à peu près sûres pour la statistique des parties du discours – mais les épigones auront-ils plus de courage que les devanciers et suivront-ils la pente raide où Gunnel Engwall s'est engagée? »

Dans les travaux de linguistique quantitative qui ont vu le jour depuis lors, on s'est contenté d'applaudir le courage de Gunnel Engwall, sans suivre son exemple, et la prudence a souvent choisi le même camp que la paresse. En s'abstenant de lemmatiser les données, on 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? 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 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.

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. Ce programme étant automatique n'est pas exempt d'erreurs. Mais il échappe à la fatigue, à l'inconstance, à la subjectivité et finalement au renoncement qui accompagnent les entreprises de désambiguïsation manuelle.

Comme le programme de Cordial est très rapide (quelques secondes suffisent pour un texte de 100 000 mots), nous n'avons eu aucun scrupule – ni aucun mérite – à lui proposer un corpus de grande ampleur, homogène comme celui de Gunnell Engwall, mais quatre fois plus étendu. Comme François Rastier s'est donné la tâche de cerner, à travers Cordial, les limites et les propriétés du genre littéraire, nous avons écarté d'emblée cette variable en réunissant des textes qui appartiennent tous au genre narratif. Les variables retenues concernent donc l'époque et l'auteur. Vingt-six textes ont été choisis qui illustrent le genre romanesque du XVIIIe siècle à nos jours (là aussi l'empan chronologique est plus large que celui de Gunnell Engwall).

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	Une Vie	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 côté 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	Ravissement	Ra
25 Le Procès, LE CLEZIO 91027		.0416	.9584	Procès	Pr
26 Hasard, LE CLEZIO 67650		.0309	.9691	Hasard	Ha
TOTAL	2186928				

TABLEAU 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).

Le corpus – dont la composition est détaillée ci-dessus (figure 2) – est soumis alors au programme de lemmatisation. 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° Sens	§	Phrase	Forme	lemme	ambig.	Typegra	CodeHexa	Codegram	Syntagme	Fonction	Num
=====	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 2. Un fichier lemmatisé par Cordial

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 3 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 5 (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.

Sommaire	Retour	N°	Mots	183	Lettres	959	Page	CLIC sur un mot pour voir les contextes	Ecart	Textes	Cherche	Notes	Code Syntaxe	page
		21	Swann			8415								
j' entendais le sifflement des trains qui , plus ou moins éloigné , comme le chant d' un oiseau dans une forêt , relevant les distances , me décrivait l' étendue de la campagne déserte où le voyageur se hâte vers la station prochaine ; et le petit chemin qu' il suit va être gravé dans son souvenir par l' excitation qu' il doit à des lieux nouveaux , à des actes inaccoutumés , à la causerie récente et aux adieux sous la lampe étrangère qu' le suivent encore dans le silence de la nuit , à la douceur prochaine du retour .									je 5 entendre 1 le 7 sifflement 2 de 7 train 2 qui 5 , plus 6 ou 8 moins 6 éloigné 3 , comme 8 le 7 chant 2 de 9 un 7 oiseau 2 dans 9 un 7 forêt 2 , relever 1 le 7 distance 2 , me 5 décrire 1 le 7 étendue 2 de 9 le 7 campagne 2 désert 3 où 5 le 7 voyageur 2 se 5 hâter 1 vers 9 le 7 station 2 prochain 3 ; et 8 le 7 petit 3 chemin 2 que 5 il 5 suivre 1 aller 1 être 1 graver 1 dans 9 son 7 souvenir 2 par 9 le 7 excitation 2 que 5 il 5 devoir 1 à 9 de 7 lieu 2 nouveau 3 , à 9 de 7 acte 2 inaccoutumé 3 , à 9 le 7 causerie 2 récent 3 et 8 au 7 adieux 2 sous 9 le 7 lampe 2 étranger 3 qui 5 le 5 suivre 1 encore 6 dans 9 le 7 silence 2 de 9 le 7 nuit 2 , à 9 le 7 douceur 2 prochain 3 du 7 retour 2 . je 5 appuyer 1 tendrement 6 mon 7 joue 2 contre 9 le 7 beau 3 joue 2 de 9 le 7 oreiller 2 qui 5					

Figure 3. 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, d'un code ou d'une structure, renvoie aux autres occurrences où le même objet est rencontré, les relations hypertextuelles s'appliquant aux quatre champs. Mais ces relations lient aussi entre eux ces champs entre eux, 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.

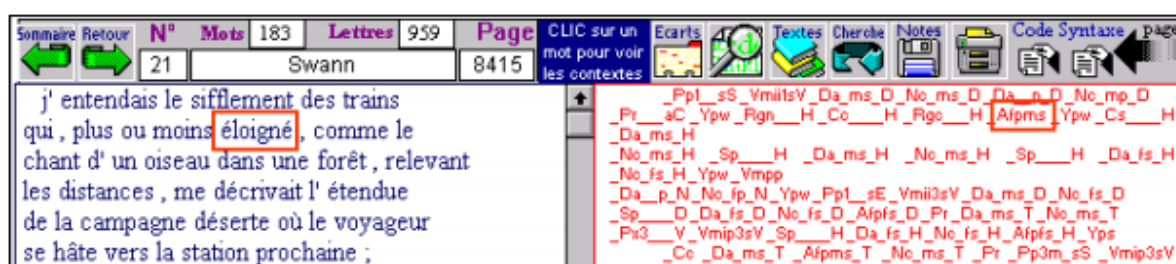


Figure 4. L'alignement forme-code

L'indexation et toutes les opérations subséquentes sont alors répétées quatre fois, au niveau structures syntaxiques (c'est-à-dire des séquences ordonnées des parties du discours, puis des codes grammaticaux, puis des lemmes, puis des formes. À l'issue de ce traitement, on obtient quatre index (figure 5) qui réagissent pareillement au clic de la souris. La forme, ou le lemme ou le code ou la structure 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 5, relatif à l'adjectif qualificatif, positif au masculin singulier. Lorsqu'on veut, non plus reconnaître un code particulier, mais rassembler les mots qui ont reçu le même codage (par exemple pour constituer une concordance ou un histogramme), on est renvoyé à une page spéciale (figure 6) qui dénombre toutes les combinaisons possibles.

Formes	Lemmes	Codes
N° 1 Marianne 33	N° 2 Paysan 52	2 1 2
N° 3 Zadig 57	N° 4 Candide 86	afpms 1 33 2 52 3 57
N° 5 Héloïse 246	N° 6 Emile 146	86 5 246 6 146 7 54 8
N° 7 Atala 54	N° 8 Rancé 117	17 9 145 10 221 11 169
N° 9 Chouans 145	N° 10 Pons 221	2 94 13 175 14 154 15
N° 11 Indiana 169	N° 12 Mare 94	38 16 116 17 97 18 300
N° 13 Bovary 175	N° 14 Bouvard 154	9 146 20 118 21 348 22
N° 15 Unevie 138	N° 16 Pierre 116	41
N° 17 Raquin 97	N° 18 Bête 300	afpms_1, 11 2 12 1 17 3
N° 19 Lune 146	N° 20 Stortiz 118	9 1 20 1
N° 21 Swann 348	N° 22 Temps 341	afpms_2, 2 1 4 3 5 1 6 3
TOUS LES TEXTES		2 9 3 10 2 11 1 12 2 13
		14 1 15 3 18 3 19 20 2
		1 12 22 8

_afpms fréquence totale: 3353

CLIQUEZ SUR UN TEXTE (ou sur TOUS) pour y repérer les contextes du code "_afpms"

Adjectif, qualificatif, positif, masculin, singulier.

OK

Structures
19 bvrplpv, 9 1 10 7 13 1 15
2 16 3 18 2 21 2 22 1
1 bvrplpvbdn, 16 1
1 bvrplpvbdnur, 2 1
1 bvrplpvbdncrbvpsppw, 22
1
1 bvrplpvbpw, 15 1

Figure 5. Les quatre index issus de *Cordial*

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.

Une fois que la sélection est faite, elle est transmise à la fonction appelante, qui délivre un contexte, une concordance, ou une liste, c'est à dire un tableau à deux dimensions dont chaque ligne dresse le profil d'une sélection grammaticale à travers le corpus et chaque colonne celui d'un texte du corpus à

travers les codes grammaticaux. Le choix des structures syntaxiques se fait de la même façon en choisissant dans l'ordre les catégories qui forment la séquence souhaitée, par exemple *déterminant* + *adverbe* + *adjectif* + *substantif* + *pronom relatif*, à chaque endroit de la séquence une constante (forme, lemme ou code grammatical détaillé) pouvant se substituer à la variable. Ainsi dans la figure 7 deux contraintes sont imposées à la sélection: l'adjectif doit appartenir au lemme beau et le verbe qui suit le relatif doit être au subjonctif (code *_V_s*). Si l'on ajoute que la fonction de chaque mot peut être isolée dans le code, et qu'un balisage sémantique est appliqué aussi au texte, on mesure combien l'étude, traditionnelle ou statistique, du texte s'enrichit de mille possibilités.

Catégorie 1	Sous-cat.2	Mode 3	Temps 4	Personne 5	Code choisi	1 2 3 4 5 6 7	Retour	Sommaire	
Verbe V	principal m	Infinitif n	Présent p	1re pers. 1	Adjectif, qualificatif, masculin, singulier,	Af_ms	→	←	
	auxiliaire a	Indicatif i	Imparfait i	2e pers. 2					
		Subjonctif s	Passé s	3e pers. 3					
	Conditionnel c	Subjonctif présent r	Futur f						
	Impératif f	Subjonctif imparfait m	Participe p	Participe passé a					
Substantif N	nom commun c				Effacer	Fonction 7			
	nom propre p								
Adjectif A	qualificatif f	Positif p	Genre 4	Masculin m	Continuer				
	ordinal o	Comparatif c							Féminin f
Déterminant D	article a		Nombre 5-6	Singulier s	Cliquez sur les critères souhaités puis sur le bouton CONTINUER				
	démonstratif d								
	interrogatif i								
Pronom P	indéfini t		Fonction 6	Pluriel p					
	possessif s	1re personne 1							
	pers. non réfléchi p	2e personne 2							
	possessif s	3e personne 3							
	démonstratif d								
	interrogatif t								
	indéfini i								
	relatif r								

Choisir la combinaison souhaitée. Un clic sur une option sert alternativement à activer ou désactiver la sélection. Les options inscrites dans la zone bleue sont réservées aux verbes. Certaines autres aux adjectifs ou aux pronoms. Les options 4 (genres), 5-6 (nombre) et 7 (fonction) concernent toutes les parties du discours, sauf les invariables. Le programme interdit les choix incohérents. Une fois réalisée la sélection, cliquer sur CONTINUER pour la transmettre au traitement en cours. (Le numéro des options indique la colonne intéressée dans le code).

Figure 6. Le choix d'un code grammatical

STRUCTURE SYNTAXIQUE				Retour	Sommaire
		1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0			
<i>Valeurs</i>		<i>Structure choisie</i> dranlv			
1	d	Déterminant Adverbe BEAU 3 Substantif Pronom_relatif _V_S			
2	r				
3	beau 3	Effacer	Verbe v	Adverbe r	
4	n	Continuer	Substantif n	Déterminant d	
5	l	Choisir dans l'ordre les éléments que l'on souhaite incorporer dans la combinaison syntaxique, en cliquant sur les boutons à droite de l'écran. Lorsqu'il s'agit d'une recherche de contexte ou d'une concordance, une valeur particulière (forme, lemme ou code grammatical détaillé) peut être donnée à un ou plusieurs éléments de la structure (cliquer sur les boutons à gauche de l'écran). Pour corriger une proposition, utiliser le bouton EFFACER. Quand le choix est fait, cliquer sur CONTINUER.	Adjectif a	Préposition s	
6	_V_s		Numéral m	Coordination c	
7			Pronom_relatif l	Subordination b	
8			Autre_pronom p	Interjection i	
9					
10					
11					

Figure 7. Le choix d'une structure syntaxique

La place nous manque ici pour entreprendre l'exploitation d'un tel corpus lemmatisé et nous renvoyons le lecteur à nos précédentes publications³. Là où l'étude des formes permet une première approche, celle des lemmes apporte le plus souvent une confirmation. Ainsi les 26 textes de notre corpus se répartissent de la même façon lorsqu'est mesurée la distance entre leurs vocabulaires, lemmatisés ou non. De même les listes de spécificités obtenues pour chaque texte à partir des formes et des lemmes ont beaucoup d'éléments communs, quoique les informations qu'elles donnent pour les verbes soient beaucoup plus sûres si la lemmatisation a opéré le regroupement des formes fléchies. Mais dès qu'on aborde la syntaxe et les faits de style, l'étude des formes montre vite ses limites et ses faiblesses: l'approche biaisée des mots grammaticaux ne permet qu'une approximation timide, car beaucoup de mots-outils sont homographes et servent à plusieurs usages, comme les couteaux suisses. Sans lemmatisation préalable, bien des aspects du texte restent inaccessibles: le genre grammatical, le nombre, la fonction dans la phrase, le temps verbal, le mode, la personne, les parties du discours et leurs multiples combinaisons réalisées dans les syntagmes et les structures syntaxiques, des plus simples (digrammes ou trigrammes) aux plus complexes, tout cela échappe aux formes brutes.

C'est le verbe qui gagne le plus à l'étiquetage: si l'on s'intéresse à son contenu sémantique, le regroupement des formes d'un même verbe offre un avantage indéniable, et si l'on étudie un temps verbal et qu'on regroupe les

verbes qui partagent le même codage et le même temps, l'avantage est encore plus net. Et faute de pouvoir aborder les points de vue qu'on vient de passer en revue, c'est le système verbal qu'on voudrait étudier dans notre corpus, à titre d'exemple. 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'étagant 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.

Une première réponse est dans le graphique 8. 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 passe composé, rejoignent l'imparfait. Les trois critères du verbes ne sont pas vraiment indépendantes. 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.

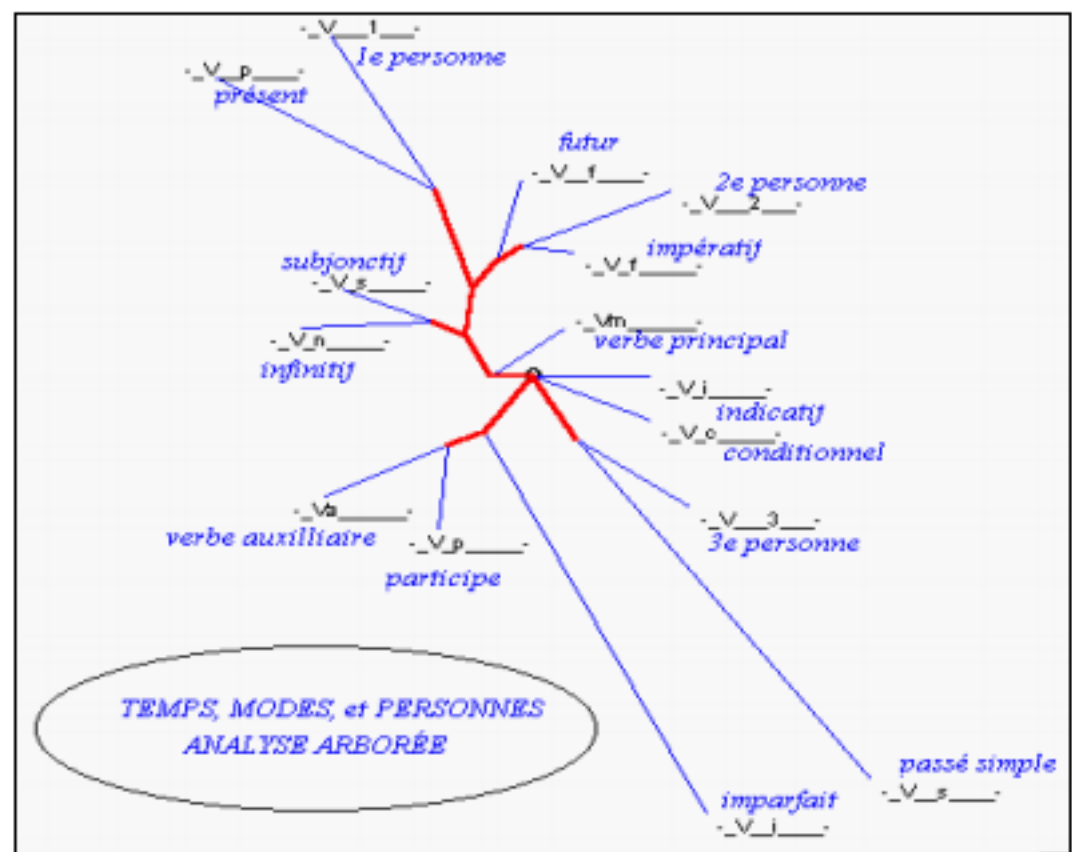


Figure 8. Analyse arborée des temps, des modes et des personnes

Les choix étant ainsi offerts, comment réagissent les textes? Comme nous l'avons observé à d'autres points de vue (distance lexicale, distance grammaticale, distance sémantique), 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 9, 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.

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.

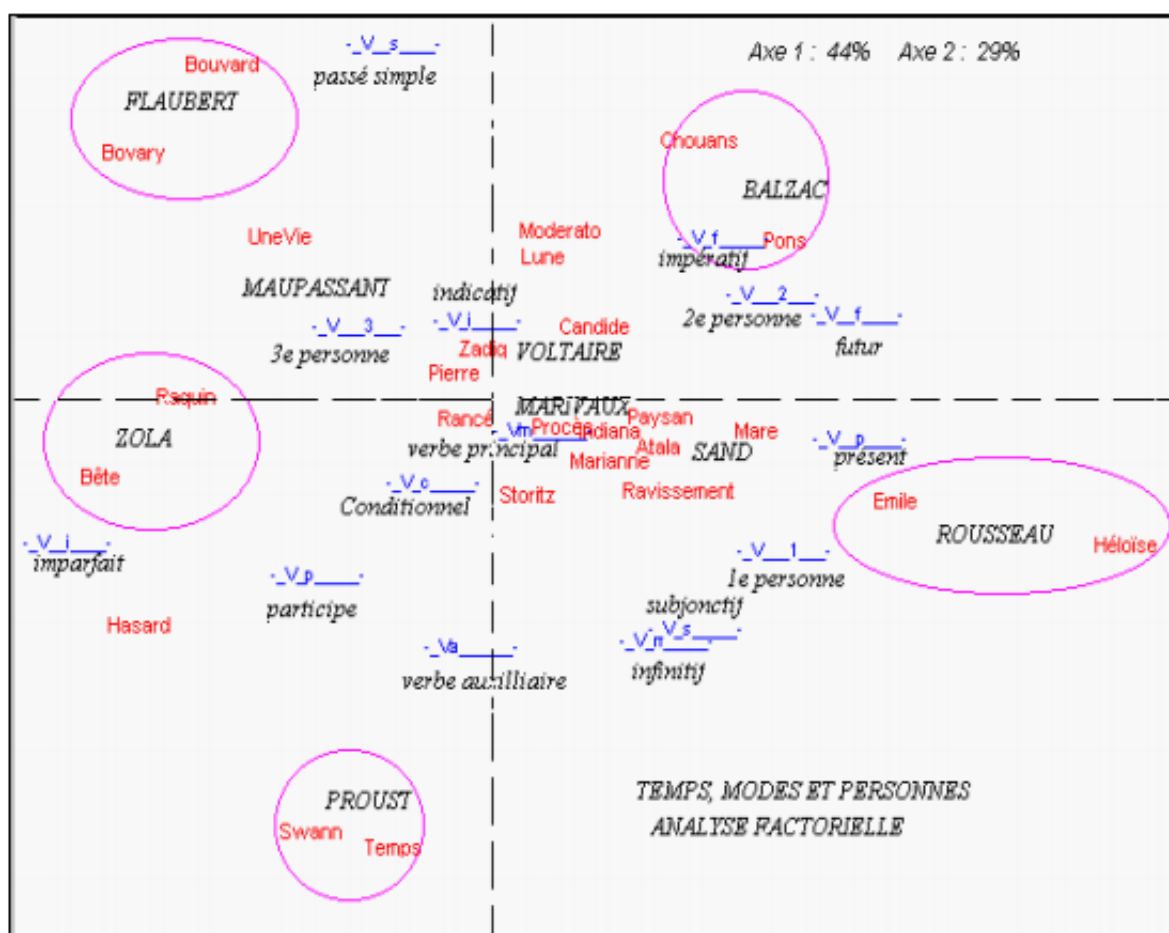


Figure 9. Analyse factorielle des modes, temps et personnes

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. En réalité cette frustration, qui est bien réelle dans les sciences humaines (le sociologue qui dépouille une enquête n'a guère le moyen d'approcher les gens), n'a pas lieu lorsqu'on a affaire à des textes. Le texte est toujours présent dans sa matérialité originelle et tous les chemins de traverse sont des chemins de retour. En fin de compte les clignotants statistiques ramènent au texte en soulignant discrètement les faits saillants. On pouvait s'inquiéter de la pertinence des résultats, quand seules les formes graphiques étaient soumises au traitement. Mais lorsqu'on met en parallèle la forme, le lemme, l'analyse grammaticale, la structure syntaxique et

le code sémantique, et qu'on obtient la convergence (c'est le cas du présent corpus et de beaucoup d'autres), on se rend compte que la voie empruntée jadis par la prudente audace de Gunnar Engwall est appelée à devenir la voie royale dans l'étude des textes. Il reste seulement à élargir la route, à adoucir les angles, à supprimer les péages, à humaniser la technique, et... à convaincre les usagers.

Notes

1 E. Brunet, « Qui lemmatise dilemme attise », *Scolia* n° 13 (Strasbourg 2), 2000, p. 7–32 et E. Brunet, « Le lemme comme on l'aime », *JADT* 2002, Rennes (sous presse).

2 B. Habert, A. Nazarenko, A. Salem, *Les linguistiques de corpus*, Armand Colin, 1997, p. 23.

3 In *Literary and linguistic Computing*, Vol. 2, N° 4, 1987, p. 251–253.