Apport des technologies modernes à l'histoire littéraire

Etienne Brunet

***Chapitre 2 (numérique) de : Étienne BRUNET,* Tous comptes faits.**
Écrits choisis, tome III. Questions linguistiques*, Bénédicte PINCEMIN (éd.),*
Paris : Éditions Champion, sous presse (publication prévue en 2016).
Publié en ligne par la revue Texto ! Textes & Cultures, http://www.revue-texto.net
Volume XXI – n°1 (2016). Coordonné par Christophe GÉRARD.

Apport des technologies modernes à l'histoire littéraire[1]

Quand l'histoire littéraire s'adresse à l'informatique pour résoudre ses incertitudes – ce qui reste exceptionnel –, elle a tendance à poser naïvement les questions les plus perfides. Comment dater sûrement un texte ? Comment décider sûrement de l'attribution d'un texte douteux ? Puisque l'ordinateur est supposé doté d'une mémoire infaillible, d'une attention imperturbable et d'une impartialité irréprochable, c'est à lui, pense-t-on parfois, de donner les réponses scientifiques, le critique se réservant pour sa part les jugements de valeur, les appréciations stylistiques, esthétiques, morales, littéraires. En somme, on demande à l'ordinateur ce qu'on obtient du carbone 14 pour la datation d'un gisement préhistorique, ou ce qu'on obtient de l'empreinte digitale pour l'attribution d'un meurtre. Mais où se trouve l'empreinte digitale d'un auteur ? Où dénicher, dans le lacis des mots, la signature, invisible et indélébile, de l'écrivain ? Comment déceler la substance ou la propriété que le temps imprimerait à un texte et que l'ordinateur n'aurait qu'à déchiffrer comme s'il s'agissait du cachet de la poste ?

Et de partir en quête d'indices, de témoignages, de preuves. Leur multiplicité est déjà suspecte, leur diversité plus encore. Et que dire de leur incohérence et de leurs contradictions ? Comme on risquait de décevoir ainsi les littéraires, on s'est plus volontiers adressé aux linguistes, en leur fournissant des aides pour la description des faits de langue. Ainsi mille comptages ont été réalisés, dont certains n'avaient qu'un but pédagogique et utilitaire[2]. D'autres ne cachaient pas des visées plus fondamentales sur le fonctionnement des unités linguistiques ; d'autres enfin avaient apparemment pour seule justification le plaisir de

1. NDÉ : Article publié dans Henri Béhar et Roger Fayolle (dir.), *L'histoire littéraire aujourd'hui*, Paris : Armand Colin, 1990, p. 94-117 (1990a).

2. L'exemple le plus connu en France est fourni par *L'Élaboration du français fondamental*, Didier, 1964.

compter et le désir de voir. Mais, dans la plupart des cas, on privilégiait la structure lexicale, plutôt que le contenu du texte. La loi de Zipf avait donné le branle à de multiples recherches sur la distribution des fréquences, les unes s'appuyant sur la statistique classique, les autres s'aventurant dans l'empirisme, et toutes poursuivant la formule magique, le nombre d'or, l'équivalent moderne de la pierre philosophale. Ainsi était née la linguistique quantitative[3]. Personne n'eut l'idée d'appeler cela littérature ou critique quantitative[4].

L'outil informatique

Les sciences de la littérature – l'histoire littéraire se range sous cette appellation officielle – n'ont-elles donc rien à espérer de l'ordinateur ? Ce serait dommage que trop d'exigences en ce domaine fassent renoncer au bénéfice immédiat que l'informatique offre aux disciplines littéraires comme à toutes les autres. À un premier niveau on observera en effet que certains matériaux de l'histoire littéraire relèvent des techniques documentaires et que dans ce domaine les vertus ménagères de l'ordinateur peuvent rendre de précieux services. S'il s'agit de gérer des informations bibliographiques, dates, titres, éditions, tirages, traductions, etc., l'informatique offre depuis longtemps les outils appropriés de classement et de gestion. Et la Sorbonne abrite précisément un projet de base de données de ce type, qui d'ailleurs n'est pas seulement bibliographique puisque l'indexation des données porte aussi sur les éléments biographiques des auteurs, les mouvements littéraires, les médias de diffusion et les institutions littéraires (Béhar, 1986). Je me contenterai de signaler un projet du même genre, auquel Raymond Ortali travaille aux États-Unis et qui tend à mettre à la disposition du public un disque optique contenant toute l'œuvre de Ronsard, mais aussi toute la critique consacrée à cet écrivain et l'iconographie qui lui est propre. La technologie du laser offre en effet, sous le faible volume d'un disque compact (ou CD-ROM) et pour un prix modéré, la possibilité d'engranger – et de restituer sélectivement – d'énormes masses d'information, soit un demi-milliard de caractères[5] accumulés sur une surface à peine plus grande que les deux mains d'un enfant.

---

3. Ou encore *lexicométrie* ou *lexicologie statistique*.

4. Il existe toutefois une « histoire quantitative », dont l'histoire littéraire peut se réclamer.

5. Pour ceux dont l'imagination saisit mieux des unités moins menues que les

On voit que ce projet, comme le précédent, dépasse le niveau bibliographique et qu'il fournit non pas seulement les références (ou informations tertiaires), mais aussi des informations factuelles : le texte même (c'est le matériau primaire) et les commentaires qu'il a suscités (niveau secondaire). Ce qu'on appelle bibliométrie – et dont Alain Vaillant est ici spécialiste – se situe pareillement au-delà de la bibliographie. Quand on a, comme lui, accès aux documents de la Bibliothèque nationale et aux archives du Dépôt légal, on dispose d'une banque de données considérable dont la gestion et l'exploitation ont nécessairement recours à l'informatique. Même le simple catalogue des grandes bibliothèques devient si encombrant et si vite périmé que le papier n'est plus un support adapté, ni même la microfiche. Et l'ordinateur est depuis longtemps l'outil indispensable des systèmes documentaires, non seulement pour la gestion locale, mais aussi pour la consultation à distance par la voie télématique.

On peut seulement regretter que les bases de données bibliographiques n'aient pas, dans le domaine littéraire, l'ampleur qu'elles ont dans les sciences exactes ou en matière économique, médicale ou industrielle. La base *Francis-H* qui a été constituée par le CNRS pour les sciences de l'homme est encore trop pauvre pour être vraiment utile. Sachant qu'à Washington le catalogue de la Bibliothèque du Congrès se trouve d'ores et déjà sur disque optique, on ne peut que constater le retard français, sinon dans la maîtrise technologique, du moins dans la production et la diffusion des bases de données. Quand la bibliothèque du Centre Georges Pompidou a voulu montrer au public les possibilités du disque optique en matière de recherche documentaire, faute d'exemple français, elle a dû recourir à une illustration américaine : l'encyclopédie Grolier intégralement reproduite sur un CD-ROM et consultable à partir d'un micro-ordinateur[6]. Espérons que la Bibliothèque

caractères, disons que cela représente l'équivalent de 1000 textes complets ou de 200 000 pages.

6. Il est possible que les grandes encyclopédies et les grands dictionnaires choisissent désormais un support informatique, et principalement le CD-ROM, pour leur diffusion et leur mise à jour. En gardant le support papier, ces entreprises éditoriales deviendraient trop lourdes, trop coûteuses et trop lentes, et leur produit serait dépassé le jour même de la publication. Au moment où, quinze ans après son lancement, le *Trésor de la langue française* s'approche des dernières lettres de l'alphabet, et où des révisions sont déjà nécessaires, plutôt que de se lancer dans la spirale des compléments, on envisage pour l'avenir un dictionnaire électronique qui intégrerait au jour le jour les modifications et les ajouts et pourrait être distribué par abonnement.

nationale suivra sans tarder l'exemple américain et que son catalogue sera disponible dans les centres documentaires sous la forme d'un disque laser régulièrement mis à jour[7].

La recherche documentaire

Dans le domaine littéraire, les techniques documentaires peuvent prendre un aspect particulièrement intéressant, du fait que la matière première est textuelle. Or l'ordinateur est très à l'aise lorsqu'on lui donne à gérer des fichiers dits ASCII qui ne contiennent que les caractères alphanumériques. C'est le cas des textes littéraires, comme de tous les textes, pour peu qu'on fasse le sacrifice des codes typographiques. Quoi de plus facile en effet que de distinguer le code binaire d'un *a* et le code d'un *b*, et conséquemment un mot d'un autre ? Les caractères étant discrets au sens saussurien et leur nombre étant limité, plus d'ambiguïté à redouter, et peu de complexité dans la recherche, le tri et le traitement de l'information. Le son et l'image opposent une autre résistance. La recherche documentaire peut donc s'exercer non pas seulement sur les références bibliographiques, sur les champs, rubriques ou descripteurs que contient la base, mais sur le texte même, l'indexation prenant en compte tous les mots qu'on y trouve et non ceux du titre ou du *thesaurus*[8]. L'indexation n'a rien de bien nouveau pour les chercheurs littéraires, habitués depuis toujours à adjoindre à leur thèse et à leurs publications savantes un *Index nominum* et souvent aussi un *Index rerum*. Glossaires, dictionnaires, gloses et citations relèvent de la même démarche qui a produit les index et les concordances à l'ère électronique. Ces outils de la recherche ont vu le jour, il y a quelques décennies, au temps où les traitements n'étaient encore que mécanographiques. Ils se sont banalisés aujourd'hui, et les chercheurs littéraires répugnent de moins en moins à s'en servir. On a fini par reconnaître les vertus de ces instruments qui, étant automatiques, exhaustifs, impersonnels, ordonnés,

7. Signalons que les vertus ménagères de l'ordinateur peuvent être mises à profit pour mettre de l'ordre dans un manuscrit. Ce n'est pas que l'ordinateur puisse accomplir une analyse graphologique, mais si la transcription de l'apparat critique est suffisamment codée, non seulement la mise en page peut être assurée par la machine avec la souplesse et la sûreté souhaitables, mais la textologie y trouve le moyen de systématiser ses recherches et d'opérer des tris dans la poussière des mots biffés, ajoutés, changés, à la façon d'un aimant rassembleur de clous. Le laboratoire que Louis Hay dirige sur les manuscrits modernes est l'exemple auquel chacun songe.

8. Il y a parfois des compromis, par exemple quand la recherche opère sur les résumés.

multipliables et bon marché[9], remplacent avec profit la boîte à fiches. À titre d'illustration on trouvera, dans le tableau l, l'exemple d'un index complet et synoptique de *À la recherche du temps perdu* et un extrait de la *Concordance* du *Projet de constitution pour la Corse*, de J.-J. Rousseau[10]. On peut cependant leur adresser deux ou trois reproches : d'une part leur maniement est lourd, surtout s'il s'agit de la concordance d'un texte long[11] ; il arrive alors que le volume des contextes restitués s'étende sur des milliers de pages et que l'adoption de la microfiche soit imposée par les contraintes économiques[12]. D'autre part, s'il s'agit d'un index, le va-et-vient entre les références et le texte dépouillé ne va pas sans fatigue pour l'utilisateur. De toute façon, la disponibilité de tels documents reste faible : rares encore sont les textes qui ont bénéficié du traitement automatique et, quand la chose se produit, les résultats, faute d'éditeur intéressé, ont souvent une diffusion confidentielle et l'on voit parfois le chercheur qui les a produits rester le seul dépositaire d'un exemplaire unique. S'ajoute à cela la disparité des normes de dépouillement, de lemmatisation, de présentation et de codage.

L'Institut national de la langue française échappe à ces griefs, par l'étendue du dépouillement, qui recouvre quatre siècles de littérature et enveloppe plus de 3000 textes, par l'immédiateté et la sélectivité des résultats, que permet le mode conversationnel, par l'homogénéité des dépouillements et des traitements, les principes de saisie et de regroupement n'ayant guère varié depuis l'origine, par la facilité et la souplesse de l'interrogation et, il faut le dire aussi, par le caractère assez économique des prestations. Ces avantages s'expliquent en partie par les progrès des équipements. Les capacités de stockage et la puissance des unités de traitement sont devenues telles que de grandes masses de texte sont désormais disponibles, au moins dans le domaine français. La télématique donne un accès instantané à n'importe lequel des 160 millions de mots qu'on a recensés à Nancy.

---

9. Pour l'utilisateur, sinon pour le producteur.

10. Ces deux publications sont au catalogue des éditions Slatkine-Champion, Genève-Paris (1983a, 1986a).

11. Quand le contexte est celui de la phrase, le volume de la concordance représente jusqu'à quinze fois la taille du texte original.

12. La publication la plus imposante, la plus luxueuse et la plus coûteuse en ce domaine est certainement celle de Roberto Busa qui a publié, en une vingtaine de volumes, la concordance des neuf millions d'occurrences relevées dans la *Somme* de saint Thomas d'Aquin.

Tableau 1.

Tableau 1. — Index de A la recherche du temps perdu
(édition de la Pléiade).

	tom.1 SWANN	tom.1 FLEUR	tom.2 GUERM	tom.2 SODOM	tom.3 PRISO	tom.3 FUGIT	tom.3 TEMPS
P							
	. .	542 c	41 f	. .	. .	456 c	. .
4	. .	542 c	. .	. .	. .	. .	. .
-7.53							
-9.34							
Pacifié							
1	. .	. .	334 f	. .	. .	. .	. .
Pacifier							
. pacifie							
1	. .	. .	. .	. .	253 e	. .	. .
Pacifique							
1	. .	. .	412 f	. .	. .	. .	. .
. pacifiques							
2	. .	. .	401 d	. .	. .	. .	770 f
3	0	0	2	0	0	0	1
-2.91							
-2.78							
Pacifisme							
	. .	757 c	. .	. .	. .	. .	826 a
3	. .	. .	. .	. .	. .	. .	844 e
Pacifiste							
2	. .	484 a	261 b	. .	. .	. .	. .
. pacifistes							
1	. .	. .	. .	. .	. .	. .	777 g
3	0	1	1	0	0	0	1
Pacotille							
1	. .	833 g	. .	. .	. .	. .	. .
Pacte							
1	. .	. .	. .	623 d	. .	. .	. .
. pactes							
2	36 b	. .	437 b	. .	. .	. .	. .
3	1	0	1	1	0	0	0
-2.33							
-2.46							
Paddock							
1	. .	898 a	. .	. .	. .	. .	. .
Padouan							
1	324 c	. .	. .	. .	. .	. .	. .
Paganisme							
1	. .	. .	. .	891 c	. .	. .	. .

	tom.1 SWANN	tom.1 FLEUR	tom.2 GUERM	tom.2 SODOM	tom.3 PRISO	tom.3 FUGIT	tom.3 TEMPS
Paillasson							
	325 a	. .	30 e	. .	. .	. .	1010 g
	326 b	. .	31 f	. .	. .	. .	1019 g
	. .	. .	542 d	. .	. .	. .	. .
7	2	0	3	0	0	0	2
Paille							
	13 b	496 b	195 f	791 a	49 d	623 f	859 d
	60 f	637 d	. .	867 f	203 e	625 c	943 b
	117 e	650 g	. .	1014 a	210 a	. .	948 d
	167 b	752 b	. .	. .	. .	. .	. .
	198 a	817 c	. .	. .	. .	. .	. .
	. .	848 b	. .	. .	. .	. .	. .
	. .	861 b	. .	. .	. .	. .	. .
24	5	7	1	3	3	2	3
-3.19							
-4.52							
Pailleronisme							
1	. .	. .	496 b	. .	. .	. .	. .
Pailleté							
	. .	. .	57 f	. .	. .	. .	. .
2	. .	. .	57 g	. .	. .	. .	. .
Paillette							
. paillettes							
	. .	. .	53 g	. .	. .	. .	. .
	. .	. .	97 a	. .	. .	. .	. .
3	. .	. .	583 d	. .	. .	. .	. .
Pain							
	49 f	445 f	24 e	737 f	10 b	443 g	782 d
	71 a	486 e	27 a	753 a	138 f	568 b	794 c
	168 d	591 f	27 b	843 d	139 b	568 c	. .
	170 e	616 d	74 f	940 c	139 d	657 g	. .
	193 b	700 a	165 f	972 g	312 f	. .	. .
	402 e	712 a	411 b	. .	. .	. .	. .
	421 e	860 b	. .	. .	. .	. .	. .
36	7	7	6	5	5	4	2
. pains							
	. .	699 g	425 g	1106 d	. .	. .	. .
4	. .	811 c	. .	. .	. .	. .	. .
40	7	9	7	6	5	4	2
-7.43	0.6	0.8	-0.2	-0.3	0.0	0.2	-1.2

Concordance du Projet de constitution pour la Corse *de J.-J. Rousseau*
(édition de la Pléiade).

montré	1	p.905 l.35	fois mais elles sont incompatibles comme il sera montré
	2	p.929 l. 1	J'ai montré jusqu'ici comment le peuple Corse pouvoit
	3	p.932 l.43	occupé d'autres soins, montre que ceux là ne sont pas au
montrent	1	p.927 l. 3	sans régle; maintenant ces forets immenses ne montrent
montrer	1	p.937 l.19	émulation au travail il ne faut pas le leur montrer
	2	p.937 l.33	elle consiste plus à la montrer ou à la décrire qu'à
	3	p.937 l.40	qu'il leur donne; lui montrer ce qu'il doit estimer, c'est
moral	1	p.933 l.40	dépends du trésor moral; c'est ce dernier qui nous met
morale	1	p.948 l. 5	Je ne leur prêcherai pas la morale, je ne leur ordonnerai
mot	1	p.904 l.28	mot l'art de rafiner sur l'agriculture, d'établir des
	2	p.911 l.33	mot les villes et leurs habitans non plus que les *fiefs*
	3	p.912 l.21	et que chaque chose reste à sa place. En un mot il
	4	p.929 l.20	modernes. Ce mot de finance n'étoit pas plus connu des
	5	p.929 l.21	anciens que ceux de taille et de capitation. Le mot vectigal
	6	p.931 l.18	Je veux en un mot que la proprieté de l'état soit aussi
	7	p.932 l.19	Que ce mot de corvée n'effarouche point des Républicains !
	8	p.948 l. 8	le mot; et qu'ils seront bons et justes sans trop savoir
motifs	1	p.937 l.26	motifs positifs d'agir. Les grands mobiles qui font
mots	1	p.942 l.14	partisans. Heureusement les mots ne sont pas les choses.
moulins	1	p.914 l.26	de ses voisins; les scies, les forges, les moulins se
mourir	1	p.915 l.12	combats, resolus de mourir ou de vaincre et n'ayant pas
	2	p.943 l.19	et tout ce qui dépend de moi. Je jure de vivre et mourir
mous	1	p.905 l.16	Ceux qu'on tire des villes sont mutins et mous, ils ne
mouvement	1	p.949 l.20	conduit pour ainsi dire les peuples avec un mouvement
mouvements	1	p.924 l.42	ordonnés. Parmi tous ces mouvements de trafic et

On peut ainsi suivre à la trace, à travers les textes, les emplois d'un mot, d'une expression, d'un thème, et analyser dans le contexte les faits de variation stylistique, de spécialisation individuelle ou d'évolution collective qui s'inscrivent dans le discours et alimentent l'histoire littéraire.

Comme cette réalisation est assez récente et que des embarras de copyright ont gêné jusqu'à présent sa pleine diffusion[13], il n'est pas inutile de dire un mot de *Stella* – c'est le nom que lui a donné son créateur, J. Dendien. Une fois que la liaison télématique est établie avec le serveur de Nancy[14], le chercheur définit son corpus, en précisant à sa guise les dates, le genre, le ou les auteur(s), le ou les texte(s) désiré(s), ou toute combinaison de ces critères. Puis il choisit les mots ou les expressions qui l'intéressent. Et enfin il indique la présentation qu'il souhaite et qui peut prendre la forme d'un index, d'une liste de fréquences, ou d'une concordance avec un contexte de la longueur souhaitée. Les résultats sont fournis en quelques secondes. On peut les imprimer parallèlement ou en différé, les enregistrer, les transmettre à un autre logiciel. Comme un terminal d'interrogation a été installé dans une bibliothèque universitaire de Paris, j'imagine que certains chercheurs ont fait l'expérience de cet outil qui n'a pas son équivalent à l'étranger et qui devrait se révéler fort utile à l'histoire littéraire. Car on a là le moyen de contrôler l'évolution d'un mot, l'usage spécifique qu'en fait tel écrivain ou tel mouvement, et derrière l'histoire des mots on peut espérer lire en filigrane l'histoire des idées et des sensibilités. L'antenne niçoise de l'INaLF a ainsi procédé à des interrogations portant sur la trace de Confucius dans la pensée française, la place réservée à « l'imagination » ou à « l'esprit »[15] dans notre littérature, le bestiaire des écrivains[16], le jeu des couleurs, la représentation des classes sociales ou de la parenté, les modalités du temps ou de l'espace, etc.

Ainsi, ayant à parler du chat chez Colette, qu'on trouve tapi au coin de chaque page, nous n'avions guère besoin de l'ordinateur. Mais la

13. Pour les textes tombés dans le domaine public, la restitution du contexte ne soulève pas de problème. Pour les autres, les prestations ne vont pas au-delà des index, en attendant l'accord des éditeurs.

14. Ce serveur – c'est un ordinateur assez puissant – porte un nom de serviteur : *Ciril*.

15. NDÉ : voir tome III, chapitre 14, « Peut-on mesurer l'esprit ? Un essai statistique d'après les données du *Trésor de la langue française* » (1984b).

16. NDÉ : voir tome I, chapitre 13, « L'exploitation des grands corpus : le bestiaire de la littérature française » (1989a), p. 301-326.

machine nous a été bien utile pour déceler la présence féline dans tous les romans français de l'époque, afin de mesurer la préférence exorbitante de Colette[17]. Cette notion de mesure nous introduit dans l'univers des mathématiques. Et les esprits littéraires peuvent refuser de franchir ce pas, sans faire injure à la machine. Les rédacteurs du *Trésor de la langue française* utilisent peu la statistique, même si une rubrique porte ce nom à la fin de chaque article du dictionnaire. Ce qui leur importe, ce n'est pas le nombre des contextes fournis par la machine pour chaque mot, mais la diversité des acceptions et l'évolution des sens qui accompagne le cours de l'histoire[18]. Les historiens de la littérature peuvent tirer du même matériau le même profit que les historiens de la langue, sans s'engager, non plus qu'eux, dans la voie des chiffres.

Les nouvelles perspectives

C'est pourtant du côté des chiffres que nous souhaitons nous diriger. Cette démarche peut apparaître incongrue à certains, ambitieuse à d'autres et dangereuse à tous. Il y a pourtant des projets plus téméraires encore, qui prétendent non seulement analyser la structure morphosyntaxique du discours, mais aussi rendre compte de la substance sémantique. Si la traduction automatique arrivait à forcer les obstacles qui obstruent le chemin depuis vingt ans, plus rien n'arrêterait l'intelligence artificielle, pas même le sanctuaire de la création littéraire. Loin de toute intention sacrilège, nous bornerons notre audace à mêler les chiffres aux lettres. Et le mot « lettres » doit être entendu ici dans son sens le plus menu, le plus « littéral » si l'on peut dire. Les méthodes quantitatives sont en effet appliquées à la reconnaissance et au dénombrement des lettres, la plus importante étant évidemment le blanc, qui délimite les mots. On en reste donc à la surface du discours, au signifiant lexical. Et on reconnaîtra d'emblée les limites des tentatives actuelles, qui n'envisagent guère que les mots individuels, coupés de leur contexte et donc de leur sens, et qui se dégagent difficilement des procédures probabilistes et des préalables méthodologiques que suppose cette démarche[19]. Il ne faut guère attendre

17. Si l'on classe quelque 200 romans selon le rôle qu'y joue le chat, les cinq textes de Colette recensés occupent les rangs 1, 3, 12, 13 et 27.

18. Ce dictionnaire est conçu comme synchronique. Mais le corpus est si vaste que l'usage n'est pas lisse, de la Révolution à nos jours, et que l'écorce éclate au fil des ans.

19. Ce n'est pas le lieu de développer les formules employées, qui sont fondées sur les lois classiques (normale, binominale, hypergéométrique, Poisson) et dont l'exposé est clairement établi dans les ouvrages de Charles Muller.

de certitudes de détail, surtout quand ces méthodes sont appliquées aux problèmes complexes de datation ou d'attribution. Mais quand des corpus très amples sont mis en œuvre et que la loi des grands nombres produit un recul suffisant, des perspectives intéressantes s'ouvrent à l'histoire littéraire. On donnera quelques exemples des résultats qu'on a obtenus par ces méthodes et qui intéressent la *Recherche du temps perdu*, les *Rougon-Macquart*, les *Mémoires d'outre-tombe*, l'œuvre de Hugo, de Giraudoux, de Rousseau[20], et plus largement le vocabulaire de la littérature française, de 1789 à nos jours. Dans tous les cas, les traitements s'appuient sur un dictionnaire des fréquences, réalisé à partir du corpus étudié et analogue à ceux que nous présentons dans le tableau 2 et qui concernent Zola, Hugo et l'ensemble du corpus des XIX^{e} et XX^{e} siècles.

1 – De Giraudoux, à qui nous avons consacré une thèse, nous ne retiendrons qu'une seule illustration, représentée ci-dessous (figure 1) : c'est la relation entre l'homme et la nature, entendons par là le rapport mathématique entre les occurrences des deux mots, ou plus largement des deux thèmes. On avait en effet donné un code sémantique aux substantifs, à côté d'un code grammatical. Le graphique, qui suit la chronologie de la gauche à la droite, rend compte du progrès des préoccupations « humaines » et du déclin parallèle des thèmes liés à la nature[21]. Cette évolution est d'ailleurs amplifiée par le passage de Giraudoux au théâtre après une production romanesque, mais elle ne se confond pas avec le genre littéraire, comme le prouve la situation des romans tardifs (*Combat avec l'ange* et *Choix des élues*). D'autres courbes montrent pareillement que Giraudoux – qui déclarait détester les abstractions, ces « éponges molles » – perd pied devant la marée montante des mots abstraits et qu'au reste le rapport concret/abstrait se superpose au rapport substantifs/verbes et à l'alternative entre richesse et pauvreté lexicale. Cette empreinte du temps est assez singulière chez ce professionnel de la jeunesse que fut Giraudoux.

20. NDÉ : la plupart de ces études monographiques sont représentées dans le tome I.

21. Ces deux mouvements contrariés se trouvent résumés dans l'évolution contrastée des trois substantifs les plus fréquents chez Giraudoux : quand le JOUR décline (r = - 0,46), l'HOMME et la FEMME se multiplient (r = + 0,58 et + 0,56 respectivement). Rappelons que r est un coefficient de corrélation chronologique et que le seuil de confiance est ici largement dépassé.

Tableau 2.

-ABI- 2 ligne 1 : fréquences absolues
ligne 2 : écarts réduits

FREQUENCE	PROSE	P.POETIQUE	VERS	TECHNIQUE	MONOLOGUE	DIALOGUE	RESTE
TENDANCE	1800 1825	1835 1845	1855 1865	1875 1885	1900 1913	1922 1930	1935 1942 1955
abîme							
4173	2953	108	736	376	802	1070	2301
	-21.13	10.29	48.25	-6.48	0.76	12.44	-10.56
-0.62	437 295	559 264	355 287	263 218	247 168	213 106	208 329 224
-0.64	4.9 -0.4	15.5 1.4	6.8 1.8	1.5 -4.3	-3.1 -5.4	-4.4 -9.0	-3.0 2.7 -5.7
abîmer							
575	496	8	29	42	110	89	376
	2.05	0.94	1.79	-3.65	0.22	-1.69	1.17
+0.38	20 15	26 45	22 36	38 78	53 51	54 23	20 46 48
+0.40	-4.2 -4.2	-2.4 2.0	-2.1 0.0	0.8 6.2	1.8 2.8	2.4 -1.8	-2.6 1.1 0.5
abominable							
1669	1595	9	24	41	359	314	996
	13.63	-1.89	-4.80	-12.25	2.87	0.64	-2.84
+0.28	52 80	77 44	60 103	235 357	206 80	75 65	78 103 54
+0.06	-7.7 -3.8	-4.0 -5.5	-4.1 0.0	14.6 23.2	8.1 -2.0	-3.8 -3.3	-2.4 -1.0 -6.8
abondance							
1383	981	21	35	346	198	147	1038
	-12.00	1.93	-2.20	14.38	-4.23	-7.30	9.26
+0.00	307 122	80 44	67 64	40 63	98 75	103 71	65 94 90
-0.42	18.6 2.2	-2.0 -4.1	-1.8 -2.4	-4.5 -3.4	-0.1 -0.9	0.8 -1.1	-2.2 0.0 -1.7

Dictionnaire des fréquences de Zola.

volume :	1	Fréquence I	I	Tendance I	Dispersion II	II	z : corpus époque II	II	III	III	III	IV	RAQUIN IV	IV	FERAT IV	V	V	V
titre :	FOR	CUR	VEN	CON	FAU	EXC	ASS	PAG	NAN	POT	BON	JOI	GER	OEU	TER	REV	BET	ARG	DEB	PAS
à		f:	51303	r:	-0.07	s:	4.104	z:	-24.85	-23.88						1371	4.06	1973	3.49	
	2364	2008	2046	2226	2209	2578	2775	1579	2258	2204	2468	2084	3009	2271	3224	1174	2621	2848	3836	2177
	4.11	1.42	0.04	2.45	0.81	4.98	-3.22	-7.57	-6.85	-6.03	-5.32	-2.66	-1.26	-4.06	3.05	-2.25	5.01	3.09	6.13	0.77
abaissant		f:	3					z:	-1.68	-2.07						0		0		
	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0
abaissé		f:	5					z:	-2.75	-2.59						0		0		
	0	0	0	0	0	1	0	1	1	0	0	0	0	1	1	0	0	0	0	0
abaissement		f:	5					z:	-3.30	-3.68						0		0		
	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	1	0
abaisser		f:	40	r:	0.19	s:	1.331	z:	-2.51	-2.46						1		0		
	2	1	0	0	4	2	3	2	3	1	1	0	6	2	1	4	1	1	0	5
abandon		f:	219	r:	-0.03	s:	1.421	z:	13.70	12.16						3	-0.99	20	4.45	
	6	10	3	7	10	12	6	6	13	10	11	14	11	18	9	8	11	13	11	7
	-1.10	0.59	-1.97	-0.68	0.24	0.64	-1.90	-0.75	0.58	-0.20	-0.20	1.52	-0.60	2.35	-1.15	1.16	0.26	0.45	-1.04	-0.72
abandonnant		f:	60	r:	0.21	s:	0.996	z:	7.57	6.28						1	-0.37	1	-0.79	
	1	5	1	0	4	4	3	3	2	1	3	4	3	5	1	3	5	3	8	1
	-0.98	1.84	-0.91	-1.60	0.93	0.77	-0.24	0.52	-0.61	-1.15	-0.11	0.90	-0.32	1.27	-1.40	1.28	1.35	-0.08	0.98	-0.97
abandonné		f:	183	r:	0.13	s:	1.329	z:	0.53	-0.20						5	0.29	5	-0.60	
	10	4	7	8	10	5	9	13	5	6	4	13	16	5	5	4	8	8	23	10
	0.82	-1.13	-0.10	0.16	0.82	-1.18	-0.47	2.42	-1.43	-0.99	-1.89	1.86	1.56	-1.31	-1.84	-0.22	-0.17	-0.53	3.10	0.87
abandonner		f:	424	r:	-0.24	s:	1.953	z:	9.35	11.47						8	-0.68	35	5.21	
	22	10	12	16	27	24	10	19	14	23	30	35	15	15	21	13	12	25	28	10
	0.96	-1.54	-1.21	-0.36	2.18	1.07	-2.99	0.83	-1.65	0.53	1.59	4.01	-2.13	-1.22	-0.87	0.83	-1.77	0.59	-0.15	-1.86

Dictionnaire des fréquences de Hugo.

	f= fréquence					r= tendance				s= dispersion			<--comparaison externe (corpus TLF)--> c=corpus		e=époque		p=prose	V=vers
compar.interne -->	p= poésie					tv= thé.vers				tp= thé.prose					ro= romans			co=corresp.
LET	HER	AUT	DAM	BOR	TUD	RUY	RAY	RHI	CO1	CON	LS1	MI1	MI2	MI3	FIN	RUE	CO2	MER	CO3	LS2	LS3
à				f:	41305		r:		0.961		s:		1.026	c:	0.5	e:	-3.8	p:	7.5	v:	-0.9
			5716	po	-33.5		tv		2.0		tp	-1.5		16850		ro	0.3		16763	co	28.1
1653	473	395	3498	415	442	639	411	4184	2856	1218	997	3725	3639	3182	596	432	4858	2602	3212	1182	485
-5.1	0.9	-4.5	-0.3	0.1	-2.2	1.8	-6.7	3.4	11.0	-13.0	-13.0	2.3	-1.7	-2.2	-11.9	-7.9	19.9	2.6	23.0	-15.8	-8.9
abaca				f:	1																
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
abaissant				f:	11									c:	2.4	e:	1.7	p:	1.7	v:	-0.7
0	0	0	0	0	0	0	1	1	0	1	0	0	1	4	0	0	0	2	0	1	0
abaissé				f:	11									c:	-0.1	e:	-0.4	p:	0.5	v:	-2.5
0	0	0	2	0	0	0	0	3	0	0	0	1	2	0	0	0	1	1	1	0	1
abaissement				f:	19									c:	1.4	e:	1.1	p:	1.2	v:	0.2
0	0	0	0	0	0	0	0	3	0	1	1	1	4	1	0	1	3	2	1	1	0
abaisser				f:	46		r:		-0.092		s:		0.588	c:	0.5	e:	-0.3	p:	0.2	v:	-3.3
			11	po	0.6		tv		-1.1		tp	-0.0		27		ro	2.5		7	co	-2.7
3	0	2	5	1	0	0	0		2	1	2	6	4	11	1	0	0	1	0	3	2
abandon				f:	33		r:		-0.203		s:		0.431	c:	-4.1	e:	-4.5	p:	-4.0	v:	-2.5
			5	po	-0.8		tv		0.2		tp	-0.9		19		ro	2.0		8	co	-1.2
1	1	1	6	0	0	0	0	2	3	1	1	4	4	1	0	1	1	4	1	0	1
abandonnant				f:	2									c:	-3.7	e:	-4.0	p:	-3.7	v:	-2.0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
abandonné				f:	79		r:		-0.383		s:		0.632	c:	-4.3	e:	-4.7	p:	-3.4	v:	-2.7
			8	po	-2.3		tv		-0.7		tp	3.3		46		ro	3.2		18	co	-2.1
5	0	1	13	1	1	1	0	5	1	2	0	13	2	12	2	0	3	3	0	3	0
0.8	-0.9	0.1	2.5	0.2	4.2	-0.1	-1.1	0	0.5	-1.2	-0.8	-1.7	-0.8	2.3	2.4	0.1	-1.1	-2.0	-1.0	-2.1	-0.3	-1.2
abandonner				f:	62		r:		-0.083		s:		0.445	c:	-9.8	e:	-9.4	p:	-7.8	v:	-6.4
			7	po	-1.8		tv		0.3		tp	-0.3		26		ro	0.2		26	co	1.3
3	1	0	4	0	0	1	0	2	1	1	0	4	4	11	2	1	0	2	3	0	0
0.1	0.4	-0.9	-0.6	0.5	-0.9	0.1	-0.9	-0.4	2.5	-1.0	0.5	-0.2	-0.7	2.8	0.5	0.1	0.6	-1.0	-0.1	-1.7	-1.1

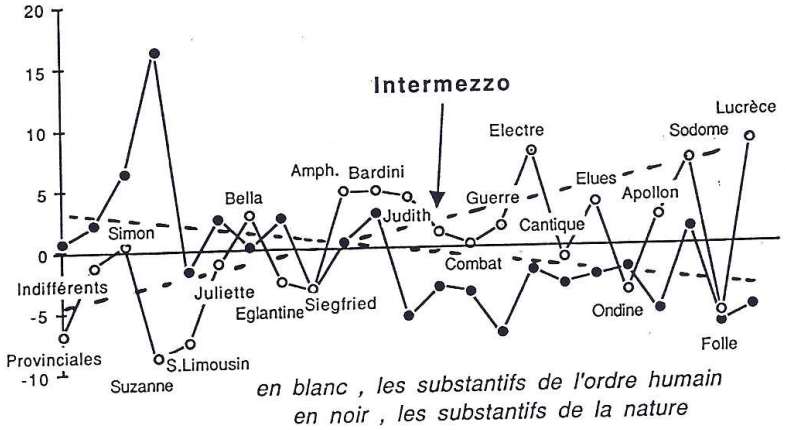

Figure 1. La nature et l'homme

Et on la retrouve chez Zola, qui a peu de points communs avec Giraudoux sinon cette aversion déclarée pour l'intellectualisme et les mots creux, et qui lâche prise lui aussi devant l'embonpoint lexical que l'âge amène avec lui. Les *Rougon-Macquart* constituant une série de vingt romans régulièrement échelonnés dans le temps, il est possible de calculer pour chaque mot un indice de corrélation chronologique, qui mesure un progrès ou une régression au cours de la rédaction. Un tri effectué sur ce coefficient permet de constituer deux lots : d'une part les mots dont l'emploi se fait plus rare appartiennent plutôt à la description d'un monde coloré et sensuel, où le corps et la femme jouent le premier rôle : FEMME, TÊTE, BRAS, COU, FACE, POITRINE, ÉPAULE, TEMPE, CARESSES, LÈVRE, SOURIRES, JOUISSANCE, EFFUSION, ARDEUR, COMÉDIE, VOLUPTÉ, BAISERS, MOLLESSE, RÊVERIE, MOUE, GRIMACE, ROUGEUR, IMPUDENCE, MIROIR, PORTRAIT, JUPE, GILET, TAPIS, NAPPE, FLACON, SENTEUR, TEINTE, etc. De l'autre côté les mots qui s'imposent de plus en plus à l'esprit de Zola sont ceux qui expriment le malheur et le désastre, et aussi la résistance nécessaire : CATASTROPHE, DÉCHÉANCE, DÉSASTRE, TEMPÊTE, MASSACRE, DESTRUCTION, CONFUSION, ANÉANTISSEMENT, DÉLUGE, EFFONDREMENT, DÉBÂCLE, DÉTRESSE, DÉMENCE, MENACE, ABOMINATION, EXÉCRATION, DOUTE, CHAGRIN, PLAINTE, PEINE, FUITE, ÉPUISEMENT, OBSTINATION, EFFORT, POURSUITE, TRAVAIL, ENRAGEMENT.

Chez Proust, il est plus facile d'étudier le thème du temps que l'effet du temps dans la genèse de l'œuvre, car il n'y a pas chez lui de relation linéaire entre la composition et la publication. On observe pourtant une évolution – sensible à un lecteur qui lirait d'une traite, comme l'ordinateur, *À la recherche du temps perdu*. La figure 2 rend compte de

cette évolution, grâce à une procédure mathématique complexe (une analyse factorielle) dont le profane peut retenir le principe approximatif : qui s'assemble se ressemble. On voit en effet sur le graphique un halo sombre entourer la *Prisonnière* et la *Fugitive* : AMOUR, DÉSIR, JALOUSIE, BESOIN, PEUR, DOULEUR, alors qu'une zone de lumière physique et morale enveloppe *Swann* et les *Jeunes Filles en fleur* : SOLEIL, MIDI, CIEL, JOUR, NUIT, MATIN, SOIR, HEURE, LUMIÈRE, MUSIQUE, NATURE, TERRE, MER, JARDIN, EAU, ROSE, CHEMIN, RUE, ÉGLISE, PROMENADE, VOYAGE, JOIE, PLAISIR, BONHEUR, CŒUR[22].

On pouvait espérer de Hugo – dont le corpus recouvre deux millions de mots et soixante ans de production – que l'effet du temps se déploierait largement, vu la longévité de l'écrivain. Mais deux facteurs viennent perturber l'observation : d'une part le trompe-l'œil des dates de publication qui, comme chez Proust, coïncident rarement avec celles de la rédaction, et d'autre part l'interférence des genres littéraires qui, malgré la *Préface de Cromwell*, maintiennent leurs spécificités et leurs prérogatives. En neutralisant l'influence du genre, on arrive à mettre en relief l'évolution thématique de Hugo qui s'accorde avec celle de Zola : moins de place donnée au sentiment et à la nature, une hantise croissante des images de mort et de désastre, une exigence plus intense de liberté, de justice et de résistance.

A-t-on le droit, dès lors, de parler d'une constante, liée à l'âge et s'imposant à tous les écrivains ? Ce serait imprudent de l'affirmer et Chateaubriand semble échapper à ce sort commun, mais il est vrai qu'enjambant la Révolution, son œuvre subit les influences contrariées de l'Ancien Régime et du nouveau et que l'évolution personnelle de l'écrivain a ressenti les contrecoups des grands séismes de l'époque. Au reste les changements s'observent à divers niveaux que nous n'avons pas le temps d'envisager ici : non seulement les thèmes changent (et les mots pour les exprimer), mais aussi les structures lexicales, les habitudes syntaxiques, les particularités stylistiques. Il est difficile d'apprécier par une seule constante un ensemble de sondages et de mesures dont la cohérence est parfois incertaine et c'est pourquoi les problèmes de datation sont si délicats à résoudre par les procédures lexicométriques, sauf à connaître la réponse donnée à l'avance par l'étude des manuscrits.

22. Le même graphique met en relief, dans la partie gauche, l'environnement mondain de *Guermantes* et de *Sodome*.

Figure 2. Analyse factorielle des substantifs de Proust

En trois ou quatre occasions, nous avons pu vérifier dans l'analyse interne les conclusions externes de l'histoire des textes. Ainsi les travaux de Maurice Levaillant ont établi que, dans le quart de siècle où furent composés les *Mémoires d'outre-tombe*, la quatrième partie fut rédigée avant la deuxième et la troisième. C'est ce que nous montrent diverses courbes, indépendantes et convergentes, établies sur l'abondance des noms propres, la proportion des catégories grammaticales, la variété

lexicale, l'importance du dialogue, le rapport point/virgule. De même, les données externes de l'*Émile* laissent entendre que la *Profession de foi du vicaire savoyard* constitue un morceau autonome dans le livre 4 et que la rédaction en est antérieure. En étudiant le rythme de la phrase et la répartition des signes de ponctuation dans l'ouvrage préalablement découpé en tranches, on observe effectivement l'irrédentisme des tranches qui appartiennent au *Vicaire savoyard* et qui s'éloignent du reste du livre 4[23].Une dernière confirmation peut être cherchée jusque dans l'incertitude chronologique de la *Recherche du temps perdu*. Chacun sait que le *Temps retrouvé* est un patchwork où des passages anciens sont intégrés à la rédaction dernière. Or, les nombreuses analyses factorielles que nous avons réalisées sur le texte de Proust montrent assez souvent trois pôles distincts : celui de Swann (les deux premiers textes), celui de Guermantes (les deux suivants), et celui d'Albertine (la *Prisonnière* et la *Fugitive*), et un îlot flottant, *Le Temps retrouvé*, qui, soumis à des contradictions internes et joignant les deux bouts de la chaîne, hésite à prendre parti[24] (figure 2).

2 – Les monographies quantitatives qu'on vient d'évoquer sont des entreprises indépendantes, chacune d'elles constituant un corpus qui sert de référence – ou de norme – aux textes particuliers qui forment l'ensemble considéré. Mais l'histoire littéraire ne saurait se contenter de ces jalons épars, même s'il s'agit de moments importants de notre littérature. L'histoire littéraire ne peut renoncer à établir un lien. ou au moins un chemin, entre ces jalons. Et précisément la lexicométrie fournit l'instrument de comparaison. Encore faut-il que les méthodes quantitatives aient une unité de mesure et respectent scrupuleusement les conventions qui définissent les éléments comptabilisés. La délimitation graphique du mot est une de ces conventions, comme aussi la volonté de s'en tenir à la surface du discours et d'éviter les objets seconds (par exemple les figures de rhétorique) qui peuvent être plus purs et plus pertinents, mais dont le relevé suppose l'interprétation humaine – et donc exclut l'unanimité. Or les études sur la littérature française disposent, à Nancy, d'une mine de données inépuisable, dont l'abondance n'est pas la

23. Voir notre *Index-Concordance* de l'*Émile*, Slatkine-Champion, 1980, t. l, p. 569-576.

24. C'est un peu ce qui se passe dans le dernier roman des *Rougon-Macquart*, *Le Docteur Pascal*. Zola pensait à ce texte depuis longtemps, comme à une conclusion et à un résumé, et non comme à une exploration nouvelle. Ce texte, qui est bien le dernier chronologiquement, occupe pourtant dans les graphiques une position centrale qui convient à sa fonction fédérative.

première vertu, mais plutôt l'homogénéité. Tous les textes ont en effet été enregistrés avec les mêmes normes de saisie et de lemmatisation. On peut donc établir les ponts non seulement entre les textes d'un même écrivain, mais aussi entre les écrivains, entre les genres littéraires, entre les époques. Et surtout on peut choisir l'ensemble de ce grand corpus, ou tel sous-ensemble que l'on voudra, comme toile de fond, pour mettre en relief les traits caractéristiques d'un texte, d'un écrivain ou d'un mouvement littéraire. Les méthodes ne changent pas[25], seul a changé en s'élargissant le terme de référence.

Là encore les éléments qui vont se détacher sur l'usage moyen[26] peuvent être de nature diverse : des classes de fréquence, ou de longueur, des catégories grammaticales, des champs sémantiques, des unités de segmentation, des expressions ou combinaisons de mots. Pour rester dans la ligne des illustrations précédentes, nous nous en tiendrons au contenu lexical et aux mots individuels. Le rôle de la machine se borne ici à calculer un écart réduit pour tous les mots où la fréquence le permet et à établir des listes triées selon les valeurs de cet écart. On obtient au moins deux lots : celui des excédents et celui des déficits. Les premiers de la liste positive sont fortement liés aux thèmes de l'écrivain, aux héros qu'il a choisis et au cadre qu'il s'est créé. La vertu heuristique est faible en de tels cas, où le calcul rejoint l'évidence. Ainsi les mots spécifiques de Proust sont pour l'ordinateur les mêmes que relèverait le lecteur le moins attentif : DUCHESSE, PRINCESSE, BARON, DUC, PLAISIR, HÔTEL, TANTE, GENRE, HABITUDE, MÈRE, MARQUISE, MOMENT, PLAGE, AMABILITÉ, AIR, JALOUSIE, PRINCE, VALET, RELATION, NOM, ALTESSE, AMI, SALON, SOIRÉE, MAÎTRESSE, PATRONNE, CLAN, CHARME, RÉALITÉ, SOUFFRANCE, DÉSIR, FEMME. Mais ces listes s'allongeant on découvre des faits inattendus qui éveillent l'attention et peuvent ouvrir des voies de recherche. H. Mitterand, devant de telles listes pointillistes, évoque l'image du portrait-robot. Il s'agit bien d'un profil décomposé, où les traits spécifiques sont restitués dans le désordre (c'est-à-dire à plat, en pièces détachées, dans une séquence linéaire qui peut être alphabétique, hiérarchique ou grammaticale).

25. Il s'agit principalement du Chi2, ou de l'écart réduit.

26. Ce point moyen ne peut être institué en norme fondamentale de la langue, parce que le corpus de Nancy ne représente qu'une variété de l'usage : le langage soutenu à vocation littéraire, mais surtout parce que tout corpus, si vaste soit-il, garde la trace des aléas de sa composition, si pondérée soit-elle.

Tableau 3. Liste des adjectifs caractéristiques de Hugo

mot	*fréquence*	*écart*	*mot*	*fréquence*	*écart*	*mot*	*fréquence*	*écart*
sombre	1198	46.50	inexprimable	93	13.20	pathétique	39	8.50
hideux	360	37.49	rayonnant	89	12.87	boiteux	46	8.24
obscur	573	33.05	ravissant	121	12.84	morose	45	8.26
charmant	1031	31.41	radieux	112	12.73	ivre	124	8.20
farouche	319	30.56	vénérable	112	12.69	horizontal	32	8.16
difforme	131	30.08	vermeil	114	12.41	tragique	95	8.13
lugubre	272	29.65	étrange	463	12.26	content	242	8.12
effrayant	354	28.97	vertigineux	39	12.25	infini	361	8.12
noir	1701	28.81	informe	69	12.20	grand	4801	7.99
sinistre	331	28.24	nocturne	109	12.20	vocal	18	7.95
formidable	284	27.73	basse	356	12.17	frémissant	82	7.89
cher	1379	27.16	visionnaire	35	12.02	austère	98	7.86
ténébreux	194	27.01	gothique	114	11.93	grave	375	7.86
pensif	251	26.96	crépusculaire	35	11.84	byzantin	33	7.84
profond	1108	26.75	éblouissant	106	11.80	immonde	61	7.80
cordial	163	26.07	infâme	160	11.71	interdit	47	7.73
monstrueux	250	24.01	magnifique	317	11.69	utile	255	7.73
auguste	196	24.12	livide	107	11.46	méchant	222	7.70
géant	289	23.50	inouï	123	11.44	hermétique	14	7.69
vaillant	159	21.82	invisible	179	11.15	superbe	235	7.49
haut	1710	20.86	indistinct	40	11.03	éloquent	73	7.35
hagard	109	20.27	sourd	286	10.90	clément	31	7.31
redoutable	211	20.01	abject	39	10.49	crénelé	23	7.08
excellent	499	19.63	flamboyant	61	10.41	terrible	461	7.08
serein	191	19.51	vaste	379	10.38	tremblant	195	7.05
doux	1210	19.04	funèbre	166	10.34	massif	84	7.04
mystérieux	403	18.87	rêveur	142	10.25	échevelé	40	7.01
insondable	59	18.27	béant	85	10.08	lumineux	126	6.98
altier	91	17.74	noble	505	9.95	sidéral	15	6.91
âpre	175	17.20	sévère	225	9.92	visible	134	6.91
immense	688	17.17	souterrain	97	9.90	doré	172	6.87
misérable	425	16.66	ineffable	76	9.73	chevelu	32	6.83
morne	231	16.39	prodigieux	141	9.70	chétif	66	6.78
joyeux	318	15.84	béni	120	9.66	respectueux	98	6.72
fier	392	15.41	idéal	242	9.56	inabordable	20	6.67
blême	138	15.40	poignant	68	9.52	stupéfait	93	6.67
bourru	51	15.14	grosse	400	9.47	sagace	17	6.55
splendide	164	15.05	fatal	211	9.27	humain	778	6.49
chère	677	14.76	plate	98	9.21	sublime	234	6.42
bon	2110	14.74	colossal	86	8.88	bossu	50	6.41
sépulcral	55	14.70	exquis	149	8.85	conventionnel	33	6.39
vil	167	14.61	épars	108	8.83	importun	37	6.35
admirable	368	14.49	absent	139	8.82	bleu	477	6.29
nain	85	14.47	double	261	8.82	digne	328	6.29
vieux	2054	14.47	puissant	318	8.81	barbu	29	6.28
affreux	399	14.24	ducal	30	8.53	inexorable	39	6.25
effaré	149	14.23	effroyable	126	8.51	inaccessible	49	6.19
vivant	588	14.19	magistral	28	8.46	innocent	112	6.17
plein	1280	14.17	célèbre	70	8.45	triomphant	97	6.17
énorme	427	14.16	éperdu	114	8.34	tortueux	35	6.15
nu	352	14.11	vertical	28	8.33	bête	467	6.14
horrible	332	13.54	hautain	73	8.31	épouvantable	104	6.09
chauve	86	13.43	frissonnant	78	8.30	latent	22	6.07
saint	1364	13.42	fourbe	30	8.28	architectural	16	6.04

Et le profil comporte des ombres, à côté des reliefs, des déficits aussi instructifs que les excédents. Ainsi les mots que Proust utilise moins souvent que ses contemporains sont parfois tenus à l'écart pour des raisons de style (le registre bas des mots COPAIN, BOUQUIN, GARS, MÔME, GAMIN), mais plus souvent par le faible intérêt porté par Proust au référent correspondant, par exemple la religion, la question ouvrière, l'univers du droit, les choses militaires (malgré la guerre de 14-18), la question scolaire, le débat scientifique, l'action politique (malgré Norpois), le monde du travail, qu'il s'agisse des champs, de l'usine, ou du bureau, et même le cadre immédiat de la vie quotidienne : l'habitat, le vêtement, l'outillage, le corps, le monde animal et le règne végétal (Brunet, 1983a,

139-172). On peut certes objecter que le silence n'équivaut pas toujours au désintérêt et que le jeu des pudeurs, des sous-entendus, des refoulements, des déguisements échappe à une machine trop naïve qui sous l'opacité des fréquences ne discerne pas les feintes du discours et les fentes du non-dit. Mais il est bien difficile au plus secret des écrivains de taire continuellement ce à quoi il pense et la fréquence des mots est rarement un simple leurre[27].

Il serait amusant de montrer que le vocabulaire spécifique de Zola est à l'opposé de celui de Proust, au point que la liste négative de l'un devient positive chez l'autre. Zola fait son régal de ce qui dégoûte Proust, les mots populaires (et pas seulement les expressions de Françoise), et le peuple lui-même (OUVRIER, MÉCANICIEN, CHAUFFEUR, VENDEUR, CHARCUTIER, BLANCHISSEUSE, POISSONNIER), le corps humain dans toutes ses parties, le corps social dans toutes ses classes, et le cadre concret dans tous ses détails. Tout cela s'accorde parfaitement avec la théorie naturaliste – dont Proust, qui est réaliste aussi, à sa façon, devait prendre le contre-pied. Chez Zola, comme chez Proust, ces listes apportent moins de surprises que de confirmations et nous renvoyons là-dessus le lecteur à notre ouvrage (Brunet, 1983a, 397-426). Pas de révélation extraordinaire non plus dans le tableau 3 qui détaille les adjectifs chéris par Hugo. Le premier de liste, SOMBRE, est l'écho et la rime de la tête de liste chez les substantifs : OMBRE. Ceux qui suivent : HIDEUX, OBSCUR, CHARMANT, FAROUCHE, DIFFORME, LUGUBRE, EFFRAYANT, NOIR, SINISTRE, FORMIDABLE, TÉNÉBREUX, PENSIF, PROFOND, MONSTRUEUX, AUGUSTE, GÉANT, etc., sont si manifestement hugoliens que cela décourage le commentaire. Une remarque s'impose toutefois à l'endroit de la méthode que certains puristes de la mathématique ont contestée, parce que le schéma d'urne n'était pas respecté dans l'exercice du langage et du choix des mots. Les écarts par rapport à une répartition aléatoire sont effectivement considérables, mais ces écarts ne servent qu'au tri et au classement, comme le feraient aussi de simples pourcentages. Si donc la méthode conduisait à l'impasse, pourquoi ses résultats seraient-ils aussi clairs et lisibles ?

---

27. La critique traditionnelle s'appuie volontiers sur les comptages implicites dont s'accompagne la lecture humaine des textes. Chaque fois qu'un jugement critique utilise les mots *caractéristique*, *typique*, *original*, *commun*, *rare*, *spécifique*, *particulier*, *fréquent*, *souvent*, *parfois*, *exceptionnel*, etc., il fait le relevé de ses compteurs, et beaucoup des appréciations dites qualitatives se fondent sur une somme d'expériences, et donc sur une base quantitative.

Ces résultats ne laissent pas parfois d'être étonnants. Suivant le sillage de Proust et hypnotisé comme lui par le chant de la grive, nous pensions que le temps donne la clé des *Mémoires d'outre-tombe* comme de la *Recherche du temps perdu*. Or si l'obsession du temps est bien présente chez Proust sous toutes ses formes, cette thématique prend un aspect particulier chez Chateaubriand, lequel est insensible au temps des horloges, au temps mécanique qui égrène les secondes, les minutes, les jours, les semaines et les mois. Toutes les divisions courtes du temps sont déficitaires dans les *Mémoires d'outre-tombe*, alors que s'y déploie largement le rythme des SAISONS, des ANNÉES, et des SIÈCLES[28]. Le regard de Chateaubriand porte au-delà de l'horizon et son temps est celui d'un presbyte, ou d'un prêtre, ou d'un prophète qui considère l'ÉTERNITÉ (c'est le dernier mot des *Mémoires*). Au déficit de certaines formes du temps s'ajoute, si l'on peut dire, un déficit aussi peu attendu : celui du sentiment. Alors que *René* et *Atala* ont donné le branle à la passion romantique, le mot PASSION est celui qui vient en tête parmi les déficits, en compagnie de AFFECTION, IMPRESSION, EXPRESSION, ATTENTION, SENSATION, ÉMOTION, PLAISIR. Chateaubriand semble avoir retenu de son siècle un peu de la retenue classique.

3 – L'histoire littéraire ne se réduit pas à enchaîner des monographies d'écrivains. Elle s'intéresse aux mouvements, aux écoles et aux genres littéraires. Or, les études quantitatives montrent que la loi du genre est souveraine et que son pouvoir discriminant est plus fort que celui du temps ou celui du tempérament propre de l'écrivain. Chez Giraudoux, le vocabulaire, la syntaxe et le rythme de la phrase diffèrent selon qu'il s'agit d'un roman ou d'une pièce de théâtre. Et si Hugo qui a cultivé tous les genres prétend supprimer les barrières entre eux, la ségrégation n'en reste pas moins entière, et chez lui la poésie ne se mêle pas à la prose, ni le roman au théâtre, ni la fiction à la correspondance. Et c'est précisément l'interférence du genre qui empêche de fonder de grands espoirs sur les méthodes quantitatives pour attribuer un texte à un écrivain plutôt qu'à un autre. La démarche se justifierait si les écarts entre les genres étaient moindres que les différences qui séparent l'écriture de deux écrivains. Or c'est le contraire qu'on observe. Les distances *intra* s'avèrent souvent plus importantes que les distances *inter*. On en donnera pour preuve l'analyse factorielle des pronoms personnels dans les

28. NDÉ : voir aussi tome III, chapitre 5, « Quand le temps change avec le temps » (1993c), (sous forme numérique).

98 textes que nous avons recensés sous six signatures différentes et dont l'ensemble représente 9 millions d'occurrences (figure 3).

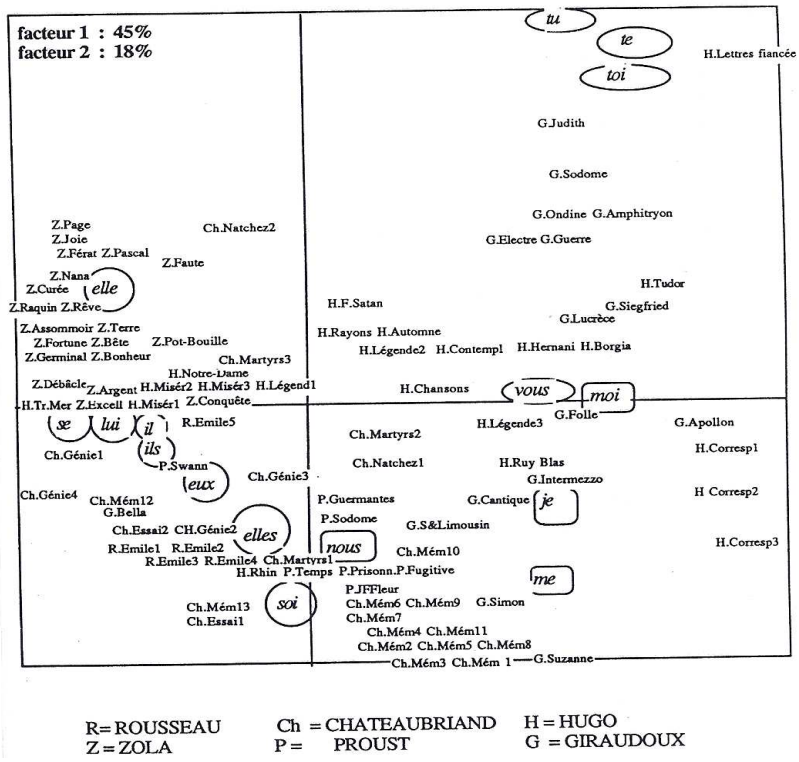

Figure 3. Analyse factorielle des pronoms personnels (96 textes, 6 écrivains)

Les pronoms personnels ont la propriété d'être la catégorie la plus sensible aux variations de la situation de discours et les écarts y sont toujours très considérables. C'est ce que montre le graphique qui groupe en haut la seconde personne, en bas à droite la première, la troisième se répartissant toute la moitié gauche[29]. Or les textes de Zola, qui appartiennent au même genre romanesque, se concentrent bien autour du pronom ELLE[30] ; les livres de l'*Émile* forment aussi un bloc compact (sauf le livre 5 qui s'oriente vers l'idylle et le roman et se rapproche de ELLE) et

29. Les formes LE, LA, LES et LEUR n'ont pas été oubliées, mais écartées pour cause d'ambiguïté.

30. Le féminin est fort prisé par Zola qui cite plus souvent la FEMME que l'HOMME, alors que chez Hugo la FEMME obtient à peine le tiers des occurrences de l'HOMME.

de même l'unité organique de la *Recherche du temps perdu* est solidement établie dans le graphique.

Mais quelle diversité chez Giraudoux qui place des représentants dans trois quadrants, et aussi chez Chateaubriand qui s'installe de préférence en bas du graphique à cheval entre la première et la troisième personne, sans s'interdire pourtant de franchir l'axe horizontal. Quant à Hugo, il occupe tout l'espace du graphique, annexant la seconde personne quand il s'occupe de théâtre ou de poésie, la première quand il s'agit d'un recueil de correspondance, et la troisième enfin dans le cas des romans. La distance est ainsi plus faible entre les *Misérables* et *Germinal* qu'entre les *Misérables* et les *Contemplations*. Tout cela était certes prévisible, s'agissant des pronoms personnels. Mais d'autres expériences ont été faites avec d'autres critères, qui aboutissent aux mêmes conclusions : certaines pièces de Corneille diffèrent plus entre elles qu'elles ne diffèrent de celles de Racine.

4 – Aux monographies d'écrivains l'histoire littéraire aimerait ajouter des monographies de mots ou de thèmes. Au lieu de s'intéresser à tous les mots d'un auteur ou d'un texte, on souhaite alors embrasser tous les textes en ne s'intéressant qu'à un seul mot, ou à un groupe de mots préalablement constitué en champ sémantique ou en classe syntaxique. Par exemple, quelles sont les représentations de l'esprit ou de l'âme au cours des deux derniers siècles et dans les différents genres littéraires[31] ? La figure 4 rend compte des 47 687 occurrences de l'ESPRIT relevées dans le corpus. Un déclin s'y dessine au cours du XIX^e^ siècle, suivi d'un léger redressement au XX^e^ siècle. La courbe de l'ÂME est, elle, régulièrement descendante, comme aussi celle du CŒUR, alors que le CORPS et la MATIÈRE sont en progrès. Dira-t-on dès lors que le matérialisme a vaincu l'esprit ? Ce serait aller trop vite en besogne et négliger le regain du spiritualisme au croisement des deux siècles. Ce renouveau s'accomplit en changeant le vocabulaire et le mot CONSCIENCE reçoit l'héritage de l'ÂME et de l'ESPRIT. La vogue de certains courants de pensée, dont le freudisme, ajoutera à la CONSCIENCE les substantifs CONCEPT et PSYCHOLOGIE et aussi beaucoup d'adjectifs, les facultés cessant d'être des substances (et des substantifs) pour devenir des modalités (et des adjectifs)[32].

---

31. NDÉ : voir tome III, chapitre 14, « Peut-on mesurer l'esprit ? Un essai statistique d'après les données du *Trésor de la langue française* » (1984b).

32. Il est vrai que certains adjectifs, comme INCONSCIENT, sont très vite substantivés et s'affranchissent tout de suite de toute servitude. On peut même estimer que l'inconscient a plus d'autonomie que le libre arbitre de la philosophie classique.

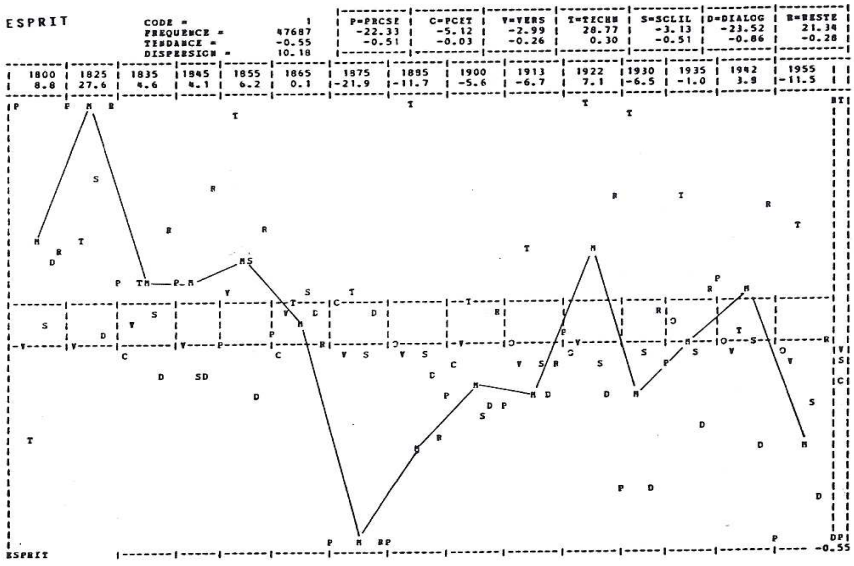

Figure 4. Courbe du mot ESPRIT

Un simple coup d'œil donné à la figure 5 permet d'apprécier le changement de perspective et de vocabulaire. L'analyse factorielle dessine un fer à cheval qui suit la chronologie, de la droite à la gauche. Sur cet arc, une quinzaine de notions liées à l'ESPRIT (par synonymie ou antonymie) se répartissent suivant les affinités qui les rattachent aux tranches les plus lointaines ou les plus proches. Le GÉNIE, le TALENT, l'ÂME et le CŒUR appartiennent au début du XIX^{e} siècle (et en réalité au siècle classique), la CHAIR s'étale à l'époque naturaliste, et le XX^{e} siècle donne la faveur à la CONSCIENCE, à la PSYCHOLOGIE et au CONCEPT. Certains choix sont plus mous : celui du CORPS et de la MATIÈRE pour l'époque contemporaine, et celui du JUGEMENT et de l'ESPRIT pour le siècle dernier. Et au centre du graphique l'INTELLIGENCE et la PENSÉE ne savent que penser et demeurent indécises.

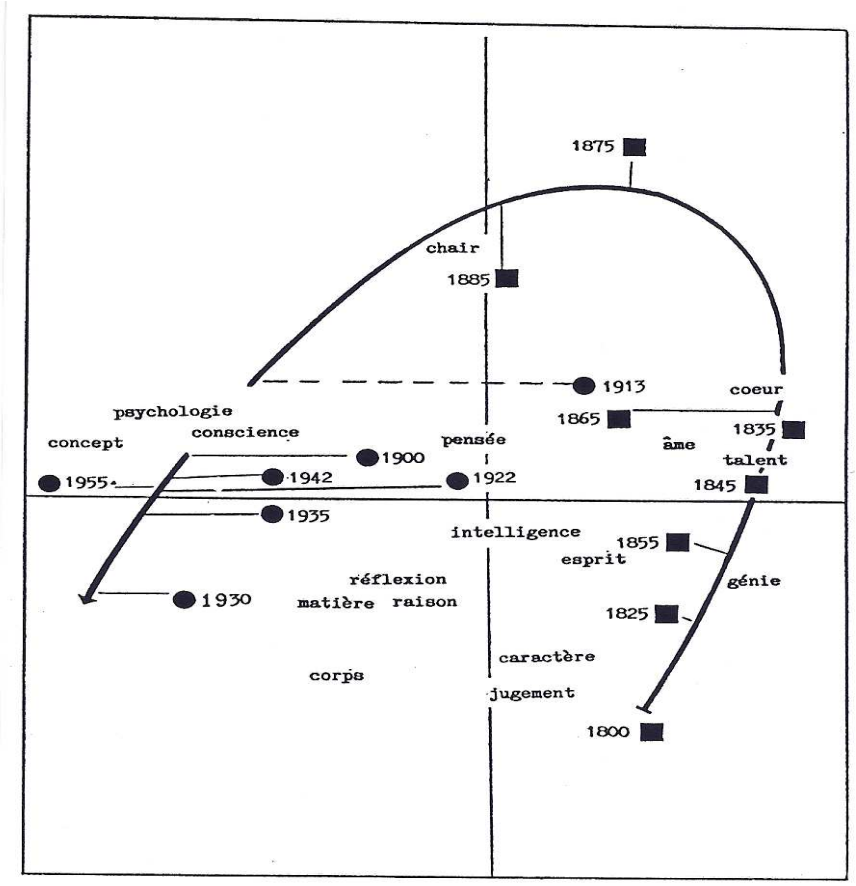

Figure 5. Analyse factorielle des représentations mentales

5 – L'histoire littéraire peut-elle s'élever plus haut encore, sous la poussée de l'ordinateur ? Peut-on échapper au rase-motte de la lecture humaine et demander à la machine une perspective panoramique sur l'ensemble des textes et sur l'ensemble des mots ? Persuadé que deux infirmités peuvent s'épauler, celle de la machine qui ne comprend rien et celle de l'homme qui ne retient presque rien, et qu'un aveugle robuste pouvait porter un paralytique clairvoyant, nous avons tenté l'exploration. Et plus heureux qu'Icare, nous avons pu rendre compte de l'aventure avant qu'elle ne s'achève (Brunet, 1981a). On ne saurait ici que survoler quelques expériences, choisies parmi les plus triviales ou au contraire les plus troublantes.

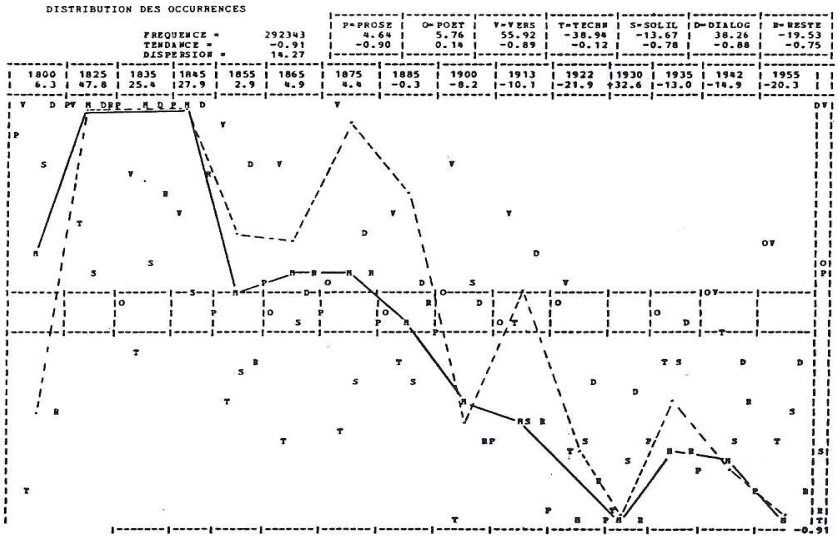

Figure 6a. L'évolution des suffixes. Le suffixe –EUR

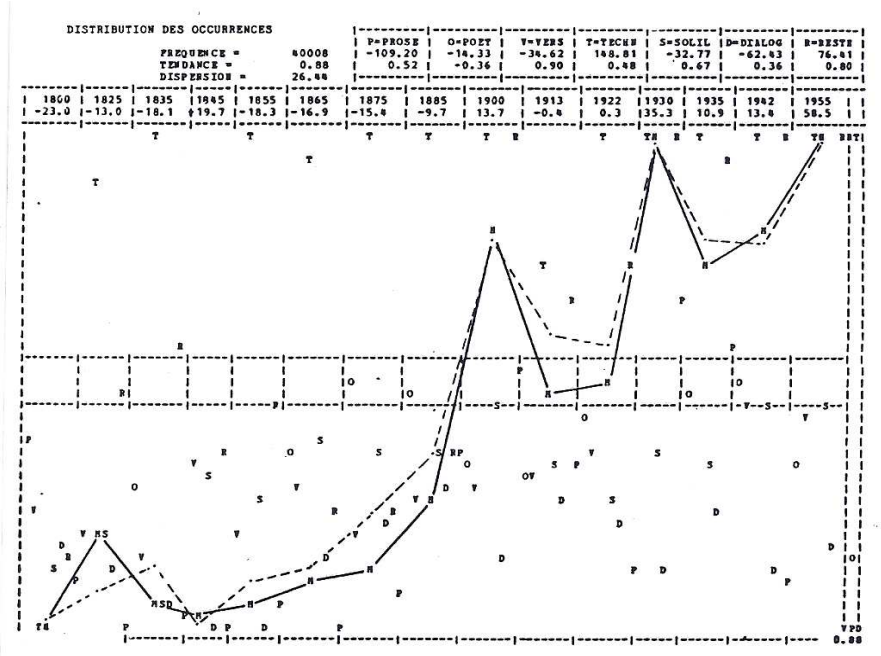

Figure 6b. L'évolution des suffixes. Le suffixe –ISME

Les résultats relatifs aux suffixes – au moins aux deux variétés que nous présentons dans la figure 6 parmi une cinquantaine étudiée – sont à ranger parmi les faits attendus. Nul ne s'étonnera du progrès des mots en –ISME (figure 6b). Le déclin des mots en –EUR (figure 6a) est moins facile à interpréter car ce suffixe a plusieurs fonctions ; il peut désigner un agent ou un instrument (MOTEUR, MENTEUR) ou plus souvent une qualité, un sentiment (COULEUR, PEUR, HONNEUR). Comme c'est cette seconde valeur qui est majoritaire (quand on considère les occurrences), le déclin du suffixe accompagne la baisse des valeurs. Il faut aussi souligner que l'emploi littéraire de ces deux suffixes est radicalement différent : les mots en –ISME sont légion dans les essais (écart réduit de + 148 dans les textes dits « techniques »), alors que le suffixe –EUR – au moins pour la seconde fonction – se rencontre plus volontiers dans la prose littéraire et plus encore dans la poésie (+ 56 dans les vers et - 39 dans la prose technique).

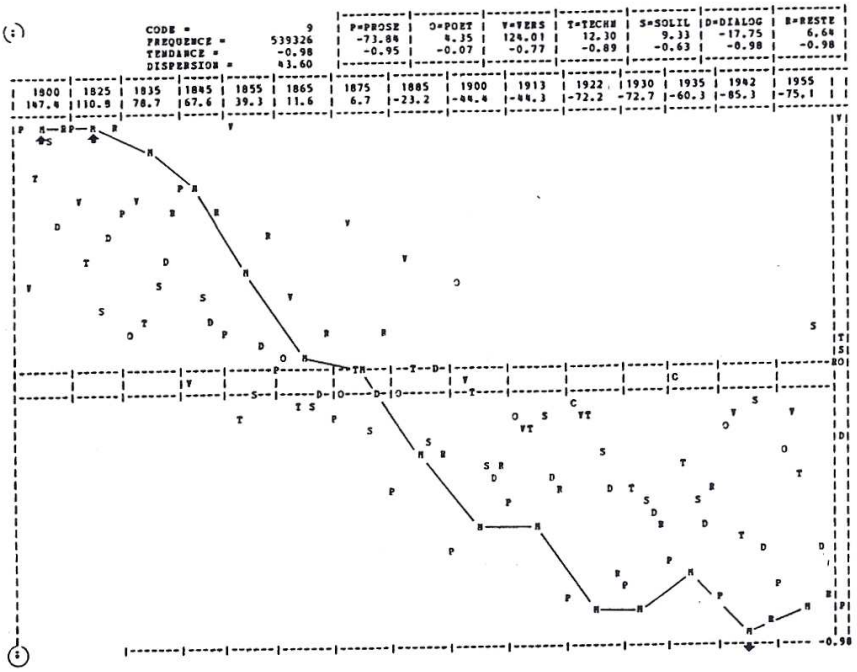

Figure 7. La courbe du point-virgule

La figure 7 représente un signe de ponctuation qui eut une grande force dans les siècles passés et qui est en voie d'extinction. Il s'agit du point-virgule. Ce signe indécis semble hérisser la plume des

contemporains, qui ne l'emploient presque plus[33]. Cet abandon est-il inconscient ? Une enquête a montré que non.

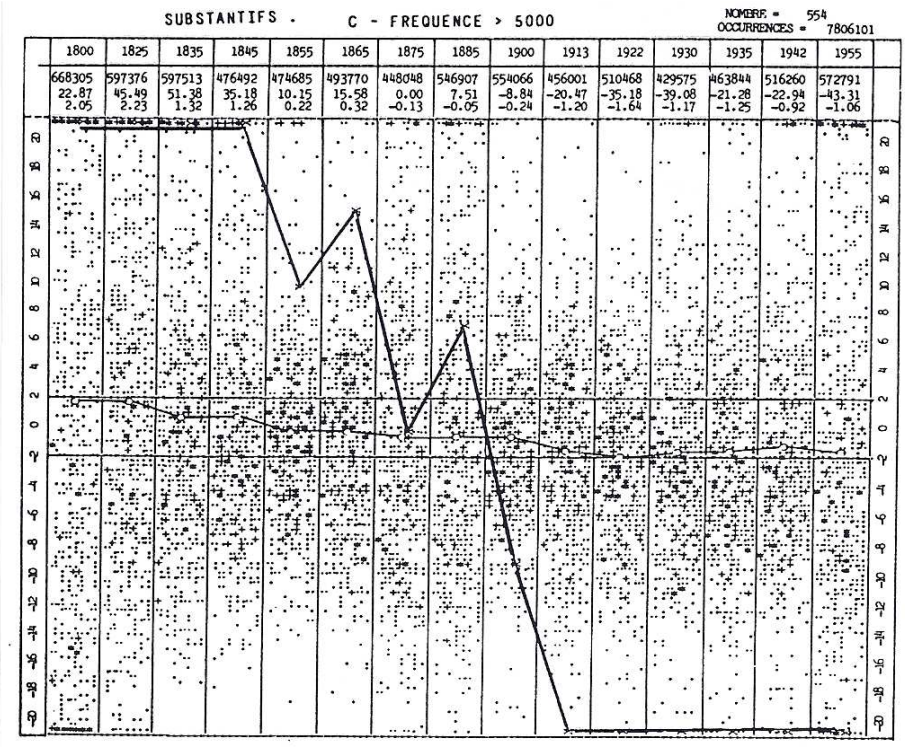

Figure 8. Les substantifs fréquents

Une autre enquête serait fort utile, à l'endroit des substantifs. On va répétant que le « style substantif » caractérise le langage actuel et qu'en particulier la presse d'information est friande de substantifs, non seulement dans les titres, mais dans la structure de la phrase. Un présentateur dira volontiers : « La question du pouvoir d'achat des chômeurs en fin de droits a fait l'objet d'un débat de principe à l'Assemblée. » Or les écrivains semblent résister à cette tendance. La figure 8, qui est établie sur les 554 substantifs les plus fréquents et qui comporte près de 8 millions d'occurrences invite à conclure à un déclin, au moins depuis la Révolution. Mais si l'on considère les substantifs

33. Ce signe, devenu simple symbole, est au contraire fort prisé dans les langages artificiels, précisément parce qu'il est disponible et qu'il n'est pas requis comme le point décimal (ou la virgule en France) par la notation des chiffres.

moins fréquents, la tendance se renverse. Quand, comme ici, les nombres ne répondent pas à l'attente, la confiance vient à manquer, au bénéfice du doute.

Tableau 4. L'époque de l'affaire Dreyfus

MOTS GRAMMATICAUX		
voici	15620	+24.81
la	1870137	+17.73
ha	34674	+16.47
c'	306914	+15.50
il	1150105	+15.45
pas	508086	+14.19
que	1570612	+13.92
de	3940365	+13.39
ne	1033896	+12.79
se	827153	+12.72
très	68535	+12.29
ainsi	59728	+12.28
pourquoi	31403	+12.25
elles	58958	+12.02
l'	1370002	+11.70
puisque	19145	+11.63
contre	45380	+11.22
ou	162381	+10.93
aujourd'hui	50999	+10.54
comment	30076	+10.51
VERBES		
affirmer	4360	+16.25
déterminer	3103	+14.06
subsister	2235	+10.86
supposer	7103	+10.71
condamner	2480	+10.18
PARTICIPES PASSES		
clos	2834	+18.01
condamné	3809	+12.39
accusé	1734	+11.86
défini	1458	+11.39
organisé	1839	+11.19
fabriqué	653	+10.93
déterminé	2923	+10.47
violé	521	+10.45
ADJECTIFS		
social	8072	+24.68
individuel	2974	+23.76
lorrain	726	+20.11
faux	9857	+17.29
collectif	1339	+16.60
professionnel	829	+13.65
musical	1352	+13.05
mathématique	1750	+12.99
radical	1284	+12.73
subjectif	753	+12.41
douloureux	4001	+12.29
conséquent	3532	+11.95
unique	6350	+11.81
identique	1104	+11.74
parlementaire	839	+11.73
scientifique	2745	+11.67
vierge	5272	+11.55
ignoble	1249	+11.54
catholique	3781	+11.18
complice	1570	+10.97
militaire	6624	+10.73
spécial	3086	+10.70
innocent	3723	+10.58
judiciaire	666	+10.23
idéal	4261	+10.03
SUBSTANTIFS		
colonel	3652	+36.52
justice	9620	+35.68
solidarité	814	+34.56
abeille	1513	+29.01
déterminisme	562	+28.54
document	1691	+27.75
moine	2392	+27.74
ruche	751	+27.57
état-major	1775	+27.17
dossier	1655	+26.96
action	17397	+24.84
office	2154	+24.23
socialiste	1409	+24.23
évolution	1969	+23.10
juif	3981	+22.79
acte	12798	+22.53
figaro	715	+22.11
prolétariat	567	+22.01
traître	1651	+21.49
messe	3417	+21.05
organisme	2131	+20.64
volonté	13869	+20.63
enquête	1291	+20.41
cloître	1172	+19.90
hérédité	683	+19.56
conscience	15179	+19.12
propriété	5907	+18.74
loi	21190	+18.32
conception	3040	+18.32
intelligence	10332	+18.30
individu	6795	+17.49
article	8445	+17.10
juge	5547	+16.89
socialisme	697	+16.86
procès	2700	+16.71
vérité	20075	+16.70
variation	1206	+16.55
synthèse	1064	+16.46
novice	519	+16.28
crime	7411	+15.97
science	13555	+15.93
accusation	1150	+15.70
énergie	4715	+15.43
canaille	1140	+15.16
iniquité	534	+14.86
âme	41800	+14.32
série	3221	+14.31
abbaye	937	+14.26
mensonge	3838	+13.91
miel	1245	+13.57
pause	1115	+13.45
allégresse	917	+13.32
fonction	5976	+13.17
mercure	521	+13.16
psaume	641	+13.15
réalité	9579	+13.15
esquisse	649	+12.98
pauvre	29903	+12.90
étang	1172	+12.65
capitaliste	638	+12.45
député	2717	+12.33
condamnation	956	+12.19
société	16161	+12.18
joie	17134	+12.01
oeuvre	19033	+11.92
suggestion	501	+11.77
intuition	1315	+11.70
crise	3741	+11.61
phénomène	6258	+11.39
cellule	2380	+11.28
oraison	847	+11.20

La situation inverse peut se produire, avec le même effet. Certains résultats en effet peuvent paraître trop beaux pour être vrais. Le tableau 4 en porte témoignage. Il représente le vocabulaire caractéristique de l'une des quinze tranches du corpus, celle qui commence en 1893 et s'achève en 1907. Comment ne pas reconnaître l'affaire Dreyfus dans la liste des adjectifs en faveur à l'époque (FAUX, IGNOBLE, COMPLICE, INNOCENT, MILITAIRE,

JUDICIAIRE, PARLEMENTAIRE, CATHOLIQUE) ou celle des substantifs (COLONEL, ÉTAT-MAJOR, JUSTICE, DOCUMENT, DOSSIER, ENQUÊTE, PROCÈS, JUGE, ACCUSATION, CRIME, MENSONGE, TRAÎTRE, CONDAMNATION, CELLULE). Les verbes ne sont pas neutres (AFFIRMER, DÉTERMINER, SUPPOSER, CONDAMNER), ni les participes (CLOS, CONDAMNÉ, ACCUSÉ, VIOLÉ). Et même les mots grammaticaux semblent prendre parti (POURQUOI, COMMENT, PUISQUE, CONTRE). Faut-il invoquer un climat obsessionnel, la présence tentaculaire de l'Affaire dans les cerveaux et dans les textes ? Probablement non. Les mots qu'on vient de citer et que l'Affaire a répandus n'ont en réalité qu'un territoire assez circonscrit, limité à quelques textes de Clemenceau, de Barrès, de Maurras et de Jaurès. Mais ces textes représentent le dixième de la tranche et la production intensive et spécialisée qui est la leur a fini par donner à l'ensemble la coloration qui leur est propre. On doit se méfier, en statistique comme en politique, des minorités agissantes.

De la même façon, les effets qui tiennent au temps et ceux qui tiennent au genre peuvent se mêler et créer la confusion. Il suffit que le dosage des genres diffère dans les tranches pour que la chronologie en soit perturbée. C'est pourquoi il a fallu neutraliser l'une après l'autre les deux variables et étudier le genre à une époque donnée, ou le temps à genre constant[34]. On a ainsi été amené à constituer, pour chacun des 70 000 mots du corpus, un tableau de 105 sous-fréquences (7 genres x 15 tranches). Les 554 substantifs regroupés dans la figure 8 ont été soumis à une analyse factorielle que nous ne reproduisons pas ici, parce que les détails y sont trop nombreux et trop menus. Outre les 554 mots recensés, le graphique contient les 105 variables obtenues en croisant le genre et le temps. On trouve là la confirmation de la primauté du genre et surtout de la distinction radicale qui oppose les textes techniques aux textes littéraires. Mais à l'intérieur d'un genre une aimantation se produit qui est celle du temps. L'ampleur du mouvement, faible en poésie, s'accentue dans la prose et particulièrement dans la variété technique, plus dépendante des mouvements de l'histoire et des changements de civilisation. Voici dans l'agencement d'une phrase, ceux qu'on trouve au haut du graphique, c'est-à-dire à une extrémité de la chaîne chronologique : « En cette ÉPOQUE, où le TITRE vaut LOI, où la RELIGION cède à l'EMPIRE de la PASSION, et où la FAVEUR de l'OPINION met en DANGER le PEUPLE, le PAYS, la PATRIE, et la NATION même, cet OUVRAGE est moins le FRUIT des CIRCONSTANCES que celui du JUGEMENT, puisqu'il respecte les USAGES de la

34. C'est la bonne vieille méthode de Claude Bernard : pour étudier un facteur, il faut l'isoler et neutraliser les autres.

LANGUE, les INTÉRÊTS de la VERTU et le GÉNIE du CHRISTIANISME. » Et voici les mots qu'on lit à l'autre bout de la série : « Le PROBLÈME est de prendre CONSCIENCE de la RÉALITÉ en JEU et d'adopter un TYPE d'ATTITUDE qui réponde à la QUESTION. » Nous laissons au lecteur le soin de démêler les deux bouts.