L'analyse statistique du Trésor de la Langue Française

Étienne Brunet

L’analyse statistique du Trésor de la Langue Française

Etienne Brunet

La plus importante banque de données linguistiques du monde est française. Disponible. Encore inexploitée. Et presque inexplorée. Cette immense forêt qui s’étend sur deux siècles, 350 auteurs, 1000 titres, 70 millions de mots[1], attend son Livingstone ou son Stanley. Aux amateurs de grands espaces et de grands nombres s’ouvre l’aventure concrète et solitaire bien loin des théoriciens qui se bousculent en piétinant la même surface pelée comme les minimes devant la cage de but.

Il n’est d’ailleurs pas nécessaire de s’enfoncer d’un coup au cœur de la forêt, au risque de s’y perdre. Le mystère peut être loti. Ainsi a-t-on suivi jusqu’à leur source quelques cours d’eau, – quelques thèmes – qui jaillissent ou s’étiolent, se ramifient ou convergent dans le paysage littéraire des deux derniers siècles[2]. On a aussi découpé quelques parcelles pour constituer un échantillonnage. On a enfin détaché quelques territoires d’un seul tenant, pour réaliser des monographies d’auteurs. C’est cette dernière voie que nous avons choisie comme exercice d’entraînement, avec Giraudoux, puis Chateaubriand, le premier (650 000 mots) faisant la courte échelle au second (1 300 000 mots).

Ce n’est pas le lieu de développer ici, dans leur détail, les conclusions de notre thèse[3] à laquelle nous renvoyons le lecteur. Les données quantitatives y confirment bien l’impression du lecteur, occasionnel ou familier, de Giraudoux. L’attention jamais lassée que Giraudoux porte aux saisons et aux heures, aux mille nuances du jour et

---

[1] 100 millions si l'on tient compte du corpus XVIII[e] siècle en voie d'achèvement.

[2] Cf. le thème du rêve par M. Faïk.

[3] *Le vocabulaire de Jean Giraudoux. Structure et évolution,* 1178 pages dactylographiées. Nice, 1976.

de l'année, aux jeux de lumière, aux reflets d'eau, au langage des pierres, des fleurs, des animaux, donne aux termes qui représentent ces thèmes une fréquence exceptionnelle. Et l'ordinateur s'accorde ici avec l'intuition. C'est le cas aussi lorsqu'on examine les catégories propres à Giraudoux : le sens des correspondances où tout est signe, écho, miroir et double, le goût de l'équilibre et de la précision, des poids, des mesures et des nombres, l'exigence du vrai, du neuf, du pur, et de cette ambition constante d'être le premier ou le seul. La liste ci-dessous rend compte avec éclat de cet écart caractéristique (a : fréquence dans la prose du XX^{e} du T.L.F. ; b : fréquence observée chez Giraudoux ; c : fréquence théorique chez Giraudoux ; d : écart réduit).

	a	***b***	***c***	***d***
premier	35 161	1 018	646	14,78
vrai	20 497	577	375	10,73
seul	39 533	1 225	726	18,69
faux	5 615	286	103	18,18
exact	2 025	73	37	5,93
juste	7 025	289	129	14,21
neuf	5 028	151	92	6,13
double	2 980	93	54	5,22
reflet	1 635	94	30	11,78
écho.	1 505	70	27	8,16
miroir	1 343	80	25	11,25
pur	7 260	173	133	3,47

L'étude statistique précise donc les traits connus ou pressentis d'un auteur, mais elle en révèle aussi les aspects insoupçonnés, notamment les ombres ou les creux du portrait (les mots non employés ou sous-représentés : par exemple ANALYSE, ÉCONOMIQUE, ORGANISATION, RÉVOLTE, OBJECTIF, STRUCTURE, PSYCHOLOGIQUE, PHILOSOPHIQUE, FONDAMENTAL, IMPLIQUER, ORGANIQUE, SUBJECTIF ...), tous également absents des textes

giralduciens[4], comme aussi les particularités stylistiques qui caractérisent l'emploi des mots-outils et qui restent assez peu sensibles à la conscience du lecteur, parce que leur abondance et leur aspect utilitaire découragent l'attention. Ainsi découvre-t-on que Giraudoux hait les indéfinis (QUELQUE, CERTAIN, PLUSIEURS, TEL, MÊME, PERSONNE, QUELCONQUE, AUTRE), qu'il évite le neutre (ÇA, CE, CECI, CELA, IL, TOUT, RIEN, EN, Y), qu'il préfère le pluriel (LES, DES, AUX), et le masculin (sauf lorsqu'il s'agit de personnes). Un principe semble gouverner ses choix : l'horreur pour les épaisseurs adipeuses dans le tissu conjonctif de la phrase : les outils que Giraudoux écarte sont souvent les plus lourds, les moins variables, quelquefois aussi les plus usés et les moins précis, comme les particules adversatives (CEPENDANT, QUOIQUE, POURTANT, MAIS, TOUTEFOIS, NÉANMOINS), consécutives (PAR CONSÉQUENT, DONC), causales (PARCE QUE, PUISQUE) ou finales (AFIN QUE). L'intuition est plus désarmée encore lorsqu'il s'agit d'apprécier la structure interne des mots, leur longueur, leur contexture graphémique, ou les associations de graphèmes, domaine *infra*- lexical et dans une large mesure *infra*-conscient, où le hasard pourtant ne règne pas puisque de puissants courants le traversent qui appartiennent au genre et au temps et que l'analyse factorielle met en lumière[5]. Même à l'échelle supra-lexicale, lorsqu'il s'agit d'apprécier le rythme du discours, la ponctuation, la longueur des phrases, la succession des répliques, des scènes, des chapitres et les lois organiques de la composition, l'analyse statistique supplée à l'infirmité essoufflée de la lecture. À plus forte raison le manque de recul et le manque de puissance empêchent le lecteur de considérer, d'un coup et dans le même instant, toute œuvre d'un auteur, et d'y suivre le cheminement chronologique ou les failles qui marquent le passage d'un genre à l'autre. Comment prouver mieux qu'avec des chiffres l'évolution de Giraudoux, qui le conduit du concret à l'abstrait, de la nature à l'homme, de la richesse à la sobriété, et du substantif au verbe ?

On n'attendait pas que les chiffres révélassent si clairement la marque du temps sur un front si lisse et les contraintes du genre sur une plume si libre et si déliée.

---

[4] La liste est longue des mots que Giraudoux écarte comme appartenant aux jargons philosophique, psychologique, économique, politique, religieux. Il s'agit le plus souvent d'abstractions molles qu'il appelait des « éponges lexicales ».

[5] Cf. notre thèse, p. 180-253.

Voit-on pareillement les rides de l'âge sur le visage tourmenté de Chateaubriand ? Le problème est ici plus délicat dans la mesure où la coquetterie de plume pousse cet auteur non point comme Giraudoux à rester le même, mais à désorienter ses détracteurs comme ses admirateurs en devançant son temps – c'est-à-dire en le suivant mais par devant. Le corpus de Nancy[6] comprend 24 sous-ensembles dans l'œuvre de Chateaubriand de *l'Essai* aux *Mémoires,* soit cinquante ans de production littéraire, où l'évolution semble aller à rebours de ce que l'on attend. Il semble en effet que la richesse lexicale soit le propre de la jeunesse alors que dans le cas de Chateaubriand les œuvres de la première période font un appel plus discret et plus classique aux ressources du lexique, et que la digue de retenue est rompue dans les *Mémoires d'Outre-Tombe.* Si l'on établit un classement d'après l'indice W[7], les 13 sous-ensembles des *Mémoires* occupent les 13 premières places. Chateaubriand a donc renouvelé son vocabulaire en abordant sa dernière grande œuvre, soit que le recul de la rédaction posthume l'ait affranchi de la réserve classique, soit qu'un goût nouveau pour la rareté et la variété lexicales se soit manifesté chez lui à partir des années 20.

Le tableau ci-après suggère cependant des indications plus nuancées : les premiers écrits *(L'Essai* et le *Génie)* précèdent sous le rapport de la richesse lexicale ceux qui les suivent dans la chronologie *(Martyrs* et *Natchez)* comme si le goût de Chateaubriand s'était progressivement fermé aux audaces du vocabulaire, jusqu'à la rupture des *Mémoires* où Chateaubriand fait volte-face. Mais la conversion n'est pas définitive : Chateaubriand se fatigue et revient peu à peu à la sobriété. Or on sait depuis les travaux de M. Levaillant que la rédaction des *Mémoires,* comme celle de la *Recherche du Temps Perdu,* n'a pas été linéaire et que les derniers livres ont été rédigés tout de suite après les premiers.

[6] Le découpage d'un texte long n'y obéit qu'à des considérations pratiques, le but étant de constituer des unités de taille voisine.

[7] L'indice W, qui permet la mesure de la richesse lexicale, est calculé à l'aide des deux paramètres N et V, selon la formule :

$$W = N^{V^{-0,172}}$$

			Rapport virgule./point	Rang	N	V	W	Rang
Essai	I		1,94	11	49 081	4 645	12,522	17
	II		2,29	6	56 924	5 113	12,408	15
Génie	I		2,74	1	57 854	5 097	12,454	16
	II		2,33	5	55 536	4 566	12,889	21
	III		2,43	3	58 562	4 995	12,627	18
	IV		2,65	2	64 667	5 460	12,360	14
Martyrs	I		2,39	4	57 468	4 836	12,711	19
	II		1,91	13	48 437	4 217	13,033	22
	III		2,08	9	48 735	4 147	13,052	23
Natchez	I		2,12	7	64 706	5 075	12,779	20
	II		1,83	15	81 634	5 013	13,634	24
Mémoires d'Outre-Tombe	Partie I	1	1,95	10	40 886	5 004	11,621	6
		2	1,94	12	51 890	5 972	11,380	2
		3	1,83	16	45 799	5 792	11,186	1
		4	2,10	8	31 192	4 528	11,378	3
	Partie II		1,83	17	74 784	6 259	12,077	10
	Partie III	1	1,73	18	73 335	6 370	11,891	9
		2	1,58	19	44 620	5 052	11,732	7
		3	1,41	23	81 965	6 199	12,326	13
		4	1,45	22	67 620	5 824	12,211	11
		5	1,36	24	46 270	4 677	12,218	12
	Partie IV	1	1,56	20	59 729	6 016	11,616	5
		2	1,56	21	70 544	6 769	11,529	4
		3	1,85	14	65 215	6 125	11,846	8

L'indice W confirme ces enseignements de la critique externe puisqu'il regroupe à des rangs voisins les livres du début et de la fin des *Mémoires*. Les problèmes de datation peuvent donc être abordés par l'approche quantitative. Surtout lorsqu'on obtient la convergence des mesures et des indices. Ainsi l'étude de la ponctuation très personnelle de Chateaubriand, – qui fait de la virgule un usage rythmique, affectif et respiratoire autant que logique, l'intercalant assez souvent entre le sujet et le verbe[8] – met en évidence l'abandon progressif de la virgule au bénéfice du point et le fossé qui sépare les *Mémoires* de la production antérieure et à l'intérieur des *Mémoires* les époques de la rédaction. La place nous manque ici pour développer les conclusions que nous inspire cette monographie. Le lecteur peut imaginer la thématique de Chateaubriand au vu d'un simple extrait des listes négatives et positives de son vocabulaire caractéristique (il ne s'agit ici que des substantifs ; la liste des verbes et celle des adjectifs offrent la même facilité d'interprétation).

[8] Cf. la thèse de Jean MOUROT, *Rythme et sonorité dans les Mémoires d'Outre-Tombe.*

Tableau 1 : Le vocabulaire significatif de Chateaubriand (extrait)
a : par rapport à la prose de la première moitié du XIX^e^ siècle.
b : par rapport à l'ensemble du corpus XIX^e^ et XX^e^ siècles.

VOCABULAIRE NÉGATIF				VOCABULAIRE POSITIF			
MOT	TOTAL	ÉCARTS RÉDUITS a.	b.	MOT	TOTAL	ÉCARTS RÉDUITS a.	b.
PASSION	173	-11,6	-06,3	CHRISTIANISME	487	+42,7	+55,2
INSTANT	232	-11,5	-10,7	SAUVAGE	605	+39,5	+44
SITUATION	32	-11,0	-11,0	RELIGION	960	+38,7	+50,2
MAISON	398	-10,6	-12,4	PATRIE	632	+36,4	+49,2
SOIR	324	-10,2	-15,6	GÉNIE	870	+35,9	+49,0
FEMME	1218	-09,8	-08,0	FORET	495	+33,2	+31,6
HEURE	655	-09,7	-13,0	TOMBEAU	473	+30,6	+43,5
MARI	111	-09,6	-08,6	TERRE	1 403	+30,0	+23,3
DOMESTIQUE	38	-09,6	-07,3	RÉVOLUTION	649	+27,2	+37,3
AFFECTION	33	-09,5	-06,9	ROI	1 707	+26,6	+54,4
IMPRESSION	51	-09,4	-10,4	FILS	1 041	+26,4	+30,1
ARGENT	153	-09,2	-09,8	FRÈRE	846	+25,4	+29,0
DOCTEUR	36	-09,2	-09,5	PEUPLE	1 305	+25,3	+40,5
SENS	231	-09,1	-15,2	MER	710	+25,0	+19,6
PAUVRE	442	-08,8	-05,9	BORD	546	+24,9	+18,3
BESOIN	281	-08,7	-10,6	SOLDAT	586	+24,7	+28,0
EFFET	328	-08,7	-09,9	BOIS	614	+23,9	+15,1
EXPRESSION	101	-08,7	-06,2	EMPIRE	523	+22,6	+32,4
LIVRE	285	-08,7	-11,1	MŒURS	475	+22,2	+40,2
HABITUDE	89	-08,6	-09,6	MÉMOIRE	507	+21,9	+24,5
VISITE	72	-08,6	-08,3	CHAMP	440	+21,9	+15,8
MOIS	270	-08,5	-06,9	CIEL	1 018	+21,5	+19,7
ATTENTION	59	-06,3	-09,2	VERTU	694	+21,1	+30,6
DIABLE	47	-08,3	-06,6	SANG	527	+20,5	+13,1
IDÉE	677	-08,3	-08,5	SOLITUDE	371	+20,4	+22,1

Notons dans le tableau 1 combien le vocabulaire de l'affectivité répugne à l'auteur de *René* : PASSION, AFFECTION, IMPRESSION, EXPRESSION, SENSATION, SENS, PLAISIR, CARESSE, ÉMOTION, JOUISSANCE, AMOUR, SENTIMENT, BONHEUR appartiennent à la zone d'ombre que fuit Chateaubriand, par un souci assez classique de la dignité. C'est le classique aussi que révèle l'étude des mots outils : par une réserve classique et aristocratique, Chateaubriand se tient à distance de l'objet, qu'il évite de montrer du doigt (déficit des démonstratifs), et de l'interlocuteur, qu'il veut

ignorer (déficit de la seconde personne), il s'abstient de trop parler de soi (déficit de la première personne, d'autant plus surprenant qu'il s'agit de mémoires), de contredire (déficit des négations), d'insister (déficit des intensifs), de s'indigner ou d'applaudir (interjections). Trop grand seigneur pour compter (déficit des numéraux) ou pour s'appesantir à discutailler (déficit des causales, des consécutives, des concessives, et même des temporelles), Chateaubriand désigne les tierces personnes par le nom propre[9] plutôt que par le pronom de rappel (déficit des pronoms de la troisième personne) : la forte proportion des possessifs par rapport aux personnels est le signe d'un style soutenu, mais le déficit des conjonctions, des relatives, des interrogations, montre que la structure de la phrase de Chateaubriand est peu complexe non qu'elle soit brève et sèche (on y compte 21,3 mots par phrase alors que la prose de la première moitié du XIX^{e} n'en a que 17,8 et l'ensemble du corpus XIX^{e} et XX^{e} 14,3) : c'est une phrase *horizontale* comme les fermes sans étage, qui étalent autour du centre les communs, les débarras et les remises. La phrase de Chateaubriand a deux pentes douces, souvent inégales. Protase et apodose se prolongent par des extensions, des circonstances où la préposition joue un rôle essentiel (excédent généralisé des prépositions). Le style de Chateaubriand – spécialement dans les *Mémoires* – est celui d'un homme qui prend du recul, à qui la réflexion et l'éloignement permettent un ton serein et digne, éloigné de la passion et de l'éloquence comme de la vulgarité et du pédantisme.

Aux monographies, déjà constituées, de Giraudoux et de Chateaubriand s'ajouteront bientôt deux projets en cours, concernant Rousseau et Proust[10], notre ambition étant de mesurer l'originalité de chacun des grands écrivains et d'apprécier aussi les invariants de la

[9] La proportion des noms propres est extrêmement forte chez Chateaubriand, qu'ils désignent les personnes ou les lieux. L'étude de cette catégorie lexicale – qu'on néglige le plus souvent – permet de définir l'histoire et la géographie d'un auteur. Dans le cas de Chateaubriand, on ne sera pas surpris en constatant que le nom le plus souvent cité est celui de Napoléon (734 occurrences) ou mieux celui de Bonaparte (977), quand ce n'est pas Buonaparte (16). Mais on peut s'étonner que le poète ossianique, le « prince des songes », se plaise assez peu dans les brumes du Nord, et que sur les coordonnées historique et géographique, il se situe systématiquement sous le ciel méditerranéen, le plus souvent en Grèce.

[10] D'autres auteurs sont sur le chantier, comme Maupassant étudié par M. Dugast. Le lecteur consultera avec profit l'inventaire que M. Musso réalise au *Trésor de la Langue Française*.

création littéraire qui s'imposent pareillement aux auteurs, notamment l'effet de l'âge et les contraintes du genre.

L'étude qui nous retient présentement dépasse le cadre d'une monographie puisqu'il s'agit de la totalité du corpus XIX^e^ et XX^e^ siècles du T.L.F. Disons tout de suite que cette tâche est écrasante et non achevée. Car si la loi des grands nombres est un avantage scientifique, le poids des grands nombres est un handicap matériel. Les résultats sont cependant acquis, sauf en ce qui concerne la frange lexicale des mots rares, soit environ 50 000 vocables qui ont au maximum 25 occurrences et qui se répartissent en 635 620 occurrences sur un total de 70 millions. Notre base représente donc 99 % du corpus (exactement 69 681 614 occurrences pour 21 029 vocables)[11].

Les données ont été directement fournies par le T.L.F.[12] et ne correspondent pas au dictionnaire des fréquences (qui publie des fréquences relatives par genre, par siècle et demi-siècle) ni au dictionnaire des variations de fréquences (qui est limité aux 7 000 vocables les plus fréquents et ne restitue que les 15 sous-fréquences des tranches chronologiques, sans distinction du genre). Il nous fallait *à la fois* les subdivisions par genre et par période et pour *tous* les mots, même de fréquence basse (car lorsque la fréquence baisse, la classe correspondante augmente et la statistique conserve ses droits en abandonnant les fréquences pour les effectifs). Nous avons donc reçu de Nancy cinq bandes magnétiques contenant pour chaque forme le détail des 105 sous-fréquences distinguées (15 tranches X 7 genres). Nous avons dû procéder à la lemmatisation, au codage grammatical, à la distinction des homographes et au tri alphabétique sur l'orthographe du lemme en alphabet pauvre, toutes ces opérations étant fort longues et fort coûteuses.

Une fois constituée cette base exploitable, on a évalué la taille de chacun des 105 sous-ensembles, ce qui était nécessaire aux calculs de

---

[11] Le centième restant ne sera pas négligé et doit faire l'objet prochainement d'une recherche spécifique, soutenue par le C.E.R. IBM de La Gaude. Dans cette masse flottante qui relève des excentricités du discours sans appartenir nécessairement au lexique, du moins au lexique normalisé des dictionnaires, devraient ressortir les faits éphémères de la mode, du néologisme, des idiotismes. Cette étude de la frange mouvante du lexique peut se révéler riche d'enseignements sur la création verbale, sur la naissance et la mort des mots. On apprend beaucoup sur le soleil quand on étudie ses protubérances.

[12] Grâce à l'obligeance des services informatiques du T.L.F. dirigés par M. Maucourt.

pondération ultérieurs. En fait, huit d'entre eux furent trouvés vides, ce qui permit une réduction du traitement à 97 sous-fréquences.

TABLEAU 2 : Évolution de quelques substantifs et adjectifs de 1789 à 1964
La « tendance » est mesurée par l'indice de Spearman ou coefficient des rangs significatif au seuil de 5 % quand p > /0.52/).

PERIODE			1789 1815	1816 1832	1833 1841	1842 1849	1850 1859	1860 1869	1870 1879	1880 1892	1893 1907	1908 1918	1919 1926	1927 1932	1933 1937	1938 1945	1946 1964
MOT	TENDANCE	TOTAL	E C A R T S R E D U I T S														
ADJECTIFS EN PROGRESSION																	
PIRE	+0,97	3126	-14,4	-07,7	-07,9	-06,5	-05,4	-05,3	-03,2	-04,5	+05,0	+04,4	+04,4	+05,4	+19,6	+08,9	+06,9
TECHNIQUE	+0,93	1663	-10,2	-10,6	-10,7	-08,7	-05,6	-09,4	-08,7	-06,4	-06,7	-04,7	-03,9	+07,5	+16,4	+12,2	+47,9
ESTHETIQUE	+0,91	1379	-11,0	-08,7	-09,7	-07,2	-02,8	-03,6	-05,5	+03,1	+00,4	-00,1	+11,4	+03,9	+02,7	+18,6	+08,6
INCONSCIENT	+0,90	1863	-12,9	-12,0	-11,9	-10,6	-10,8	-07,5	-05,5	+06,4	+04,9	-00,2	+00,4	+02,8	+00,8	+20,1	+33,8
MULTIPLE	+0,89	1009	-06,9	-06,8	-07,0	-05,3	-04,7	-02,5	-04,1	-01,8	+06,1	-02,7	+04,7	+03,6	-01,2	+05,0	+22,5
IDIOT &c.	+0,88	1656	-11,6	-09,2	-09,8	-07,1	-02,2	+01,3	+02,4	+03,8	+00,7	+02,3	+00,9	+05,5	+11,5	+08,2	+05,1
ADJECTIFS EN REGRESSION																	
IMMORTEL	-0,91	1094	+10,5	+03,0	+08,9	+03,0	+00,6	-00,2	+02,5	-02,6	+01,2	-01,3	-03,4	-06,3	-06,9	-05,0	-05,4
SAGE	-0,90	5902	+12,0	+13,5	+06,3	-01,4	+01,6	-02,1	-01,8	+02,5	-02,6	-02,4	-05,2	-06,5	-02,4	-06,2	-07,2
INFORTUNE	-0,90	1315	+40,6	+10,9	+02,6	+02,6	-01,4	-01,4	-04,3	-07,9	-04,3	-05,0	-07,4	-06,5	-08,5	-08,5	-07,1
ROMAIN	-0,87	3185	+17,3	+33,7	-00,7	+13,2	-02,4	+09,2	-07,1	-03,7	-06,3	-03,0	-07,8	-08,8	-06,2	-09,0	-12,2
BARBARE	-0,80	2939	+14,2	+19,0	-02,1	+01,7	-04,1	+15,9	-04,2	+00,2	-00,0	+00,5	-07,4	-09,5	-07,7	-07,5	-09,0
VENERABLE	-0,78	1093	+05,4	+03,0	+00,4	+07,0	-01,4	-00,2	+01,9	-01,4	+01,6	+00,0	-00,1	-05,2	-01,5	-03,4	-06,2
SOLITAIRE	-0,75	3871	+05,6	-00,6	+10,0	+13,2	-00,1	+01,9	-05,8	-01,0	-00,9	-03,8	-01,5	-08,4	-05,1	-02,5	-01,9
AUGUSTE	-0,74	1170	+06,2	+03,4	+00,1	+01,2	+02,2	+01,6	+08,0	-01,0	+01,0	+05,2	-03,3	-04,5	-06,7	-05,3	-07,8
INGRAT	-0,74	1530	+07,8	+00,4	+05,8	+04,3	+01,1	+00,0	-01,4	-04,7	+01,6	+02,7	-02,2	-04,1	-03,8	-05,7	-02,7
CHASTE &c.	-0,72	1286	-00,8	+01,0	+09,8	+03,1	+08,4	+01,4	+01,0	+03,4	-00,4	-00,2	-03,4	-04,1	-06,4	-05,5	-06,7
SUBSTANTIFS EN PROGRESSION																	
SOUCI	+0,98	3742	-15,0	-10,4	-05,6	-07,4	-07,2	-03,7	-02,5	+00,0	+03,7	+03,3	+05,5	+05,5	+06,7	+17,8	+10,0
TENTATION	+0,98	1965	-09,6	-08,0	-03,4	-03,4	-02,9	-02,1	-00,8	-00,5	+00,7	-00,6	+02,3	+04,0	+03,5	+13,0	+08,5
RYTHME	+0,97	2070	-12,0	-10,3	-11,5	-08,1	-07,9	-06,8	-05,4	-05,9	+05,3	+04,4	+09,1	+07,4	+08,1	+15,7	+10,5
APPEL	+0,97	4007	-14,8	-12,7	-11,7	-10,0	-08,0	-04,2	-00,9	+00,0	+05,1	+03,2	+08,0	+13,5	+05,6	+12,5	+15,4
CIGARETTE	+0,97	1577	-11,9	-11,1	-10,6	-07,7	-07,1	-07,1	-00,9	+02,3	-02,0	+04,6	+05,0	+06,5	+16,3	+12,9	+12,2
PROBLEME	+0,97	5936	-19,9	-14,7	-13,4	-13,5	-10,9	-08,6	-13,1	-08,3	-00,3	-05,5	+04,6	+25,3	+17,5	+18,7	+42,1
CONTACT	+0,97	3043	-10,6	-10,3	-09,5	-08,4	-06,4	-02,0	-06,3	-04,1	+03,3	-01,3	+09,5	+09,3	+03,4	+10,3	+22,7
USINE	+0,96	1356	-09,9	-08,7	-09,4	-06,3	-07,0	-05,4	-02,4	-00,8	-00,4	-01,2	+01,5	+05,1	+28,4	+05,0	+12,9
VACANCES &c.	+0,96	1596	-11,3	-09,3	-06,3	-02,9	-04,9	-02,8	-00,6	-02,5	-00,1	+08,5	+03,9	+04,8	+08,6	+05,7	+10,9
SUBSTANTIFS EN REGRESSION																	
PLEURS	-0,95	2699	+14,7	+10,1	+25,8	+00,6	+08,7	-03,1	+00,6	-04,2	-02,4	-06,4	-06,5	-09,8	-09,1	-10,8	-10,8
FLAMBEAU	-0,95	1511	+11,0	+08,8	+09,3	+05,1	+08,8	+02,0	+01,5	-03,1	-01,9	-04,2	-07,7	-07,8	-08,6	-06,7	-07,7
FORTUNE	-0,95	9414	+08,1	+16,6	+19,1	+31,1	+05,8	+05,1	-02,9	+03,7	-08,7	-09,2	-13,5	-10,7	-12,3	-15,6	-16,5
VOILE	-0,95	5290	+08,4	+03,1	+14,6	+05,0	+07,2	+04,8	+00,2	-01,2	-02,7	-02,9	-04,7	-08,2	-07,5	-08,3	-09,3
EPOUX	-0,94	3396	+44,7	+05,8	+05,6	-04,5	+04,5	+02,8	-02,6	-04,2	-06,3	-06,3	-08,3	-07,4	-08,9	-08,0	-11,8
JOUG	-0,94	1004	+17,8	+08,8	+04,7	+01,8	+00,8	-02,9	-01,7	-03,3	-00,1	-03,2	-03,8	-03,0	-05,2	-06,1	-06,8
ANTIQUITE	-0,93	1548	+17,5	+08,1	+10,2	+06,0	+02,4	-00,5	-05,2	-01,6	-04,5	-05,9	-04,4	-06,3	-06,0	-05,9	-06,3
TOMBEAU &c.	-0,93	4209	+24,1	+04,3	+16,6	+20,4	+04,5	+02,0	-01,3	-07,3	-05,9	-02,9	-06,9	-13,3	-11,6	-10,8	-11,7

Dans un premier temps, nous nous sommes intéressé aux 15 tranches chronologiques, tous genres réunis, en reconstituant le dictionnaire des variations de fréquences, mais sur une base trois fois plus large. Un fichier spécial recueillit les vocables dont la fréquence est supérieure à 1000 (il y en a 4 378), après que les fréquences absolues eurent été transformées en écarts réduits[13].

Or sur ces écarts réduits un classement peut être établi que l'on compare au classement chronologique. Le coefficient de corrélation

[13] On aurait pu choisir aussi légitimement la transformation en fréquences relatives – mais les variations aléatoires n'apparaissent guère alors, non plus que le sens des écarts.

obtenu (ou indice de Spearman) permet de mesurer l'évolution du mot : si l'indice est inférieur à -0,52, la régression du mot est significative, s'il est supérieur à + 0,52, la progression échappe au hasard. Deux listes ont ainsi été établies, l'une suivant l'ordre alphabétique, l'autre donnant, pour chaque catégorie grammaticale, la suite classée des mots qui progressent ou régressent le plus. Les courts extraits que nous présentons dans le tableau 2 mettent en relief les profits et pertes de la langue. Parmi les gains, beaucoup s'expliquent par les faits de civilisation (la CIGARETTE, L'USINE, la TECHNIQUE, l'INCONSCIENT, les VACANCES), certains par une tendance de la langue à l'abstraction (PROBLÈME, CONTACT, ESTHÉTIQUE). Au rayon des vieilleries lexicales, on découvre un lot d'adjectifs nobles et sévères (INFORTUNÉ, AUGUSTE, VÉNÉRABLE, SOLITAIRE, etc.) et un stock de substantifs (poétiques) qui sentent la poussière (il faudrait dire la poudre) du tombeau (FLEURS, FLAMBEAU, VOILE, JOUG, TOMBEAU, ANTIQUITÉ). Le deuil ne sied plus à notre époque et le vocabulaire poétique de la mort est mort. L'étude de ces variations de fréquence suggère mille autres observations qui peuvent contribuer à préciser l'histoire des mentalités à travers l'histoire des mots. Dans le cadre étroit de cet article, nous ne pouvons que souligner l'extrême sensibilité du langage comme témoin et produit de l'histoire. Même dans la zone des fréquences moyennes et élevées rien n'apparaît stable. Les mots sont animés de mouvements, de courants qui les entraînent à la surface ou les précipitent au fond. Plus d'un mot sur trois franchit le seuil de 5 %, exactement 1 597 sur 4 378, dont 757 vont à la grandeur (plutôt des verbes) et 840 à la décadence – plutôt les catégories nominales.

Encore le coefficient de Spearman n'est-il sensible qu'aux courants réguliers et continus qui traversent dans le même sens la masse lexicale.

Mais il ne réagit pas aux troubles passagers ou aux mouvements contraires. Pour observer ces derniers phénomènes, on doit examiner le détail des écarts réduits. Le procédé le plus efficace pour préserver une certaine lisibilité est d'opérer un tri dans chaque tranche par valeurs décroissantes de cet écart. On constitue ainsi le vocabulaire significatif – positif et négatif – de la tranche considérée. L'extrait présenté dans le tableau 3 n'a rien de très flatteur pour la littérature de notre temps : on y chercherait vainement une once de poésie, de fantaisie ou de sentiment. C'est le vocabulaire de l'analyse politique, économique, philosophique. En s'en tenant aux seuls adjectifs, on voit que les préoccupations de notre époque concernent l'action (VOLONTAIRE, INVOLONTAIRE, RESPONSABLE, LIBRE, CONSCIENT, MENTAL, RÉEL, POLITIQUE, SOCIAL, etc.) et que ses prétentions ne sont guère dissimulées (FONDAMENTAL, VITAL, CENTRAL,

ESSENTIEL, NORMAL, ÉLÉMENTAIRE). Notre littérature apparaît visiblement engagée.

TABLEAU 3 : Vocabulaire significatif (positif) de la dernière tranche (1946-1964) du corpus XIX^e^ et XX^e^ s. du T.L.F.

MOT	ÉCARTS			MOT	ÉCARTS		
	RÉEL	THÉOR	RÉDUIT		RÉEL	THÉOR	RÉDUIT
CONDITION	1 755	852	+32,2	PLAN	1 417	642	+31,8
POSSIBILITÉ	630	198	+31,9	ACTIVITÉ	1 019	405	+31,7
ANALYSE	721	249	+31,1	LIBERTÉ	2 167	1 148	+31,3
ÉQUILIBRE	665	226	+30,4	SENS	3 779	2 378	+29,9
PRODUCTION	656	229	+29,3	LIMITE	852	330	+29,9
GROUPE	1 146	519	+28,6	COMMANDEMENT	619	207	+29,7
RAPPORT	1 932	1 047	+28,5	CONCERNER	546	175	+29,0
RÉEL	1 270	607	+28,0	MESURE	1 873	1 000	+28,7
DÉVELOPPEMENT	818	330	+27,9	RÉVOLTE	589	204	+28,0
MILITAIRE	1 108	512	+27,4	DOMAINE	776	312	+27;3
IMAGE	1 872	1 040	+26,8	ENTREPRISE	644	248	+26,2
UNIVERS	1 102	543	+25,0	LIBRE	1 759	1 003	‘ +24,8
CONTENU	545	209	+24,2	SOCIAL	1 201	624	+24,0
ORGANISATION	710	305	+24,1	CHOIX	739	325	+23,8
ATTITUDE	867	406	+23,8	FONCTION	948	462	+23,5
RÉGION	728	319	+23,8	AUTORITÉ	1 018	514	+23,1
OBJET	2 276	1 437	+23,0	REPRÉSENTATION	620	262	+22,9
MARCHÉ	798	373	+22,9	ACTE	1 681	990	+22,9
EXPÉRIENCE	1 310	721	+22,8	CONTACT	571	235	+22,8
MONDE	6 470	4 938	+22,7	PROJET	1 030	537	+22,1
RECHERCHE	827	398	+22,4	POLITIQUE	1 989	1 242	+22,0
AGIR	2 137	1 367	+21,7	DESTIN	610	266	+22,0
INTERPRÉTATION	296	94	+21,5	OPÉRATION	844	422	+21,3
SITUATION	1 244	702	+21,3	APPARAÎTRE	1 334	767	+21,3
ÉMOTION	1 125	620	+21,1	ESSENTIEL	729	350	+21,1

Une inquiétude subsiste cependant : les faits observés sont-ils spécifiques à l'époque ou au genre, ne tiennent-ils pas plutôt aux variations du dosage des genres à travers les époques ? En particulier le caractère prosaïque et utilitaire de la dernière tranche n'est-il pas imputable à l'abondance exceptionnelle des productions techniques ? Reste donc à réduire ce facteur de trouble. Isoler le genre conduit à la restitution des 97 sous-ensembles que nous avions simplifiés pour une première approche. Pour chaque vocable on aura cette fois une série de 97 écarts réduits qu'on soumettra au tri. Et l'on obtient le vocabulaire significatif de chaque subdivision du temps et du genre, par exemple

celui de la poésie entre 1820 et 1830 ou de la prose (corpus technique exclu) entre 1946 et 1964. On s'aperçoit alors que les caractères que nous discernions précédemment pour l'ensemble de cette dernière tranche ne doivent rien à l'influence des textes techniques et qu'on les retrouve aussi nets dans la prose littéraire. On peut aussi isoler un genre, par exemple le corpus poétique, et y suivre l'évolution du lexique tout au long de deux siècles. Évolution restreinte d'ailleurs : le genre poétique, malgré les manifestes et les révolutions de palais, est le plus rebelle au changement. Du romantisme au surréalisme les poètes se ravitaillent aux mêmes stocks lexicaux, même s'ils multiplient les combinaisons et tordent le cou à l'éloquence, à la syntaxe et à la logique. On peut aussi opérer des regroupements sélectifs comme nous avons fait pour la prose littéraire de la première moitié du XIXe siècle en réunissant les quatre premières tranches tout en éliminant les textes techniques et les textes strictement poétiques – ce qui justifiait la comparaison avec Chateaubriand. Ainsi peut se résoudre le problème de la norme, qu'on constituera souplement par un déplacement approprié, latéral ou vertical, du genre ou de l'époque.

Quand l'attention est focalisée sur un mot – et le lexicographe est le plus souvent dans cette situation – l'analyse détaillée de sa distribution peut être représentée dans une courbe – ou dans une superposition de courbes – chacun des 97 points étant désigné par son genre et prenant place dans la suite linéaire du temps (axe horizontal), tandis que la valeur de l'écart réduit fixe la position dans le plan vertical. La zone médiane où l'écart réduit n'est pas significatif (entre -2 et + 2) est plus ou moins encombrée suivant que la distribution du mot est régulière ou non. Ainsi la courbe du mot HOMME et celle du mot CHOSE (deux des trois substantifs les plus employés de la langue) frappent par leur évidence visuelle. On y voit d'emblée que la part de l'homme s'amenuise quand celle des choses grandit. Est-ce la fin de l'humanisme et l'aliénation de l'homme au sein des objets ? Ces conclusions seraient prématurées au vu de ces deux seules courbes, dont l'une d'ailleurs est sujette à caution dans la mesure où le mot CHOSE n'est souvent qu'un fantôme sémantique, un neutre abstrait qui voisine avec TOUT, RIEN, CE et CELA[14]. Elles sont pourtant corroborées par d'autres courbes qui signalent le déclin de L'ÂME, du

[14] Le neutre grammatical fait des progrès rapides dans la langue, marqués par une corrélation chronologique très forte. Le coefficient s'élève à 0,62 pour ÇA, et 0,50 pour CE, 0,19 pour CELA, 0,45 pour TOUT et 0,49 pour RIEN, alors que le seuil significatif est franchi avec une valeur de 0,20 (pour 97 paires d'observations).

CŒUR, de la VERTU, de la BEAUTÉ, du BONHEUR... Nous disposons ainsi d'un millier de courbes dont le défilé rapide laisse à l'image rétinienne, comme dans la lanterne magique, le mouvement de la langue. Mais il est souvent aussi intéressant de faire un arrêt sur l'image et d'étudier les oppositions internes qui agitent un même mot et dont les lignes de force ne sont pas toujours parallèles. Il arrive qu'un mot change de registre, qu'il acquière ou qu'il perde des connotations poétiques, ou techniques. On voit ainsi que le genre poétique résiste mieux que les autres genres à la désaffection qui frappe L'HOMME comme à la promotion dont bénéficient les CHOSES.

A ceux qui se méfient de l'impression visuelle, notre programme de courbes[15] propose deux aides à la synthèse : un coefficient de Spearman qui résume d'un mot – d'un chiffre – l'évolution d'ensemble, et la série de 15 écarts réduits qui suivent la chronologie sans distinction de genre (dernière ligne du graphique). Mais il existe une aide à la synthèse, plus sophistiquée et plus puissante, que la science mathématique (aidée par l'ordinateur) met à notre disposition. Il s'agit de l'analyse factorielle et notamment de l'analyse de correspondance de Jean-Paul Benzécri. Les données proposées à cette méthode sont celles qui figurent dans le tableau 2, c'est-à-dire pour chaque mot une suite chronologique de 15 écarts réduits. Cinq analyses ont été exécutées concernant les mots dont la fréquence est supérieure à 1 000, soit 776 verbes, puis 1 129 adjectifs ou participes, 2 003 substantifs, 131 adjectifs/substantifs et enfin 167 mots de relation. Les résultats de chaque analyse sont reproduits sur un graphique où apparaissent à la fois les variables (les époques) et les individus (les mots). En s'en tenant ici à l'analyse des verbes, on voit aisément se dessiner l'évolution de la langue. Les verbes dont l'usage se répand n'ont plus le frémissement affectif, l'épaisseur concrète, la puissance évocatrice qu'ils avaient au milieu du XIXe. Les verbes qui créent un nuage autour du noyau de 1955 (bas du graphique), soit DÉFINIR, CONCERNER, CONSTITUER, SIGNIFIER, IMPLIQUER, COMPORTER, IMPORTER, VISER, PROVOQUER, MANIFESTER, ENVISAGER, CONSIDÉRER, CONSTATER, AFFIRMER, MARQUER, CORRESPONDRE, DÉSIGNER, RÉVÉLER, etc. sont dépourvus de chair. Monnaie fiduciaire passablement dévaluée, ils font regretter le temps où les échanges linguistiques se faisaient par le troc, ou par l'or, avec des mots qui représentaient quelque chose. Il suffit de faire la comparaison

[15] Nos programmes originaux – écrits en PLI, en Fortran et en Cobol – sont à la disposition des chercheurs. S'adresser à CUMFID, UER Lettres et Sciences Humaines, 98, boulevard Herriot 06200 Nice.

avec la constellation des verbes qui entourent la tranche de 1835 (haut du graphique) et où l'on relève : FRÉMIR, EFFRAYER, BÉNIR, DÉVOUER, VENGER, RESPIRER, PROMETTRE, REPENTIR, BLESSER, FRAPPER, RAVIR, BLÂMER, ACCABLER, SOULAGER, PÉRIR, MÉNAGER, TOURMENTER, DÉCHIRER, etc.). Il est vrai que l'époque 1800 – la Révolution et l'Empire et les Idéologues – partage plus d'un point commun avec notre temps : et notamment le goût des verbes abstraits qui définissent plus qu'ils n'évoquent. Y aurait-il des cycles dans l'histoire de la pensée et du langage, comme ces phases Simian qu'on a cru découvrir dans les variations économiques ?

Qu'il s'agisse des verbes ou des autres catégories grammaticales, les graphiques obtenus suggèrent en effet l'image d'une boucle qui tend à se refermer et où les extrêmes (les tranches 1 et 15) se rapprochent après qu'un large croissant a été décrit dont la corne horizontale suit le mouvement linéaire du XIXe siècle, et la corne verticale celui du XXe. Car – et c'est là le résultat le plus remarquable – une extraordinaire convergence permet de superposer les cinq graphiques, alors même que les analyses sont indépendantes, puisqu'elles s'appliquent à des unités lexicales différentes, à des catégories étrangères les unes aux autres. L'influence du temps (car les 15 variables sont ici des divisions chronologiques) apparaît donc primordiale dans tous les coins et recoins du lexique, dont aucun élément n'apparaît stable, et pas davantage les mots de relation que les classes nominales. Mais le temps n'est pas seul en cause. Quand on donne pour variables à l'analyse factorielle non plus les 15 tranches chronologiques, mais les 97 sous-ensembles qu'on obtient en croisant le genre et le temps, on voit que le premier l'emporte sur le second. Trois analyses opérées successivement sur les substantifs, les verbes et les mots de relation soulignent la prépondérance du genre, qui constitue la ligne principale de partage du graphique, et qui oppose la partie droite à la partie gauche, les textes techniques aux textes littéraires. Les deux autres pentes de la pyramide (le bas et le haut du graphique) appartiennent à la chronologie. Dans chaque genre, en prose, dans le dialogue, dans le soliloque, dans les textes techniques, partout une chaîne linéaire parcourt, dans l'ordre et de bas en haut, les tranches chronologiques. Seul le corpus en vers se protège contre le temps et les variations y sont beaucoup plus faibles. Encore l'analyse ne s'arrête-t-elle pas là ; elle fournit d'autres facteurs tout aussi lisibles qui donnent de la réalité une appréciation plus délicate.

Mais la place nous est trop chichement mesurée pour déployer ces cartes. Et le temps autant que l'espace nous manque. Car on s'essouffle vite à suivre le rythme de l'ordinateur qui accumule les résultats comme

les stocks de pièces dans une chaîne d'usine. Peut-on appeler résultats d'ailleurs ce qui n'est qu'une base de départ, un terminus (a quo plutôt que ad quem) où s'arrête lâchement l'ordinateur pour passer le témoin au chercheur ? Un super ordinateur viendra-t-il un jour nous délivrer de ceux qu'on a ?