Revue, Informatique et Statistique dans les Sciences humaines
XXI, 1-4, 1985, pp. 35-52

Les Rougon-Macquart. Aspects quantitatifs

Etienne BRUNET

Nous ne sommes point très sûr d'avoir l'aval de Zola quand nous soumettons les *Rougon-Macquart* aux calculs et à la statistique. Zola n'a pas beaucoup d'estime pour la statistique dont il parle incidemment dans trois romans. Ce ne sont guère que des notations rapides, mais d'autant plus significatives qu'elles sont gratuites. C'est par exemple dans *Son Excellence Eugène Rougon* (1) le défilé comique des "six messieurs, grands pieds, grosses mains, larges figures massives", que le préfet présente "comme des membres distingués de la société de statistique". C'est encore cette réplique savoureuse de M. Vabre qui passe son temps à répertorier les tableaux, à faire des fiches, des classements et des décomptes et à qui l'on demande s'il s'intéresse aux arts : "Mais non, je n'ai pas besoin de voir les tableaux. Il s'agit d'un travail de statistique" (2). Qu'on remplace les tableaux par les textes et voilà condamnée la statistique lexicale. C'est d'ailleurs un tel procès qui est fait à Mouret explicitement (3) : "Vous n'imagineriez jamais à quoi Mouret passe son temps dans la pièce où il s'enferme ? (...) Eh bien ! il compte les *s* qui se trouvent dans la Bible. Il a craint de s'être trompé, et il a recommencé trois fois son calcul (...) Ma foi ! vous aviez raison, il est fêlé du haut en bas, ce farceur-là !".

Nous n'avons pas compté les *s* dans les *Rougon-Macquart* (4), et la statistique n'autorise pas à parler des textes sans les lire. Bien au contraire il y a deux lectures qui se complètent et se corrigent : celle de l'homme et celle de la machine, la première plus sélective mais intermittente, la seconde toujours attentive mais jamais sensible. On attend que la seconde, dans les meilleurs cas, confirme les intuitions, les convictions ou les constats de la première. Mais il arrive qu'il y ait désaccord et qui faut-il croire alors ? Ainsi nous étions persuadé en

abordant l'oeuvre de Zola après celle de Proust, de Giraudoux et de Chateaubriand, que la matière lexicale de Zola est fort abondante et nous avions en tête telle ou telle page descriptive où Zola cultive à tel point la variété et l'exhaustivité lexicales qu'il semble faire concurrence au dictionnaire comme Balzac faisait concurrence à l'état civil. Et anticipant sur la présente enquête nous avons déclaré imprudemment : "la soumission au réel peut engendrer la multiplication documentaire du lexique, comme cela se produit chez Zola." (5)

Or les chiffres sur ce point nous apportent un désaveu formel. On sait que la "richesse" ou variété lexicale se définit comme un rapport entre le nombre *N* d'occurrences observées dans un texte (c'est la longueur du texte) et le nombre *V* de vocables (ou mots différents) qu'on relève dans le même texte et qui mesure l'étendue de son lexique. On sait aussi que ce rapport n'est pas simple et qu'il ne se réduit pas à un seul quotient. Mais avant d'entrer dans le détail des calculs, une simple comparaison avec des corpus déjà établis donne une indication très claire et assez surprenante : le vocabulaire de Zola est plus pauvre que celui de Proust. Un des sous-ensembles de la *Recherche du temps perdu*, la *Prisonnière*, peut être comparé pour la longueur aux plus longs des romans de Zola : l'*Assommoir* (165 017 mots), *Germinal* (172 410), la *Terre* (171 492) ou la *Débâcle* (195 340). Or la *Prisonnière*, dont la longueur est un peu moindre, a un vocabulaire nettement plus étendu (8 633 vocables contre 7 383, 7 775, 7 955 et 7 825 respectivement). Et cela est d'autant plus probant que les longs romans de Zola figurent parmi les plus riches, tandis que la *Prisonnière* est le plus "pauvre" des textes de la *Recherche*. Bien entendu l'écart avec Giraudoux se creuse encore : en réunissant les trois romans : *Suzanne*, *Siegfried et le Limousin* et *Juliette* on obtient 10 108 vocables pour 164 683 occurrences. Là encore pour une valeur moins élevée de *N* on a une valeur plus forte de *V*. Et cela est vrai des mêmes romans pris individuellement : avec respectivement 63 811 et 63 912 occurrences, *Suzanne* et *Siegfried* comptent 6 136 et 6 590 vocables, soit beaucoup plus que le *Rêve* qui n'en contient que 5 326 et qui est pourtant plus long (70 193 occurrences). Si l'on prend le témoignage de Chateaubriand, la conclusion ne change pas. Ainsi les données relatives à la dernière partie des *Mémoires d'Outre-Tombe* (*N* = 63 361, *V* = 5 600) montrent une variété lexicale plus grande que celle du *Rêve*, qui figure pourtant en bonne place parmi les textes de Zola.

La où les données ne conduisent pas à des conclusions aussi sûres et aussi directes que celles qui précèdent, notamment quand N et V sont à la fois plus grands (ou plus petits) dans le texte A que dans le texte B, force est de recourir à des formules plus compliquées qui permettent la comparaison. Or qu'on s'appuie sur la loi binomiale ou sur l'indice w on est invité pareillement à conclure à la sobriété lexicale de Zola. Ainsi quand notre indice w

$$w = N^{V^{-a}}$$

vaut 10 pour Zola pour $a = 0,189$, dans les mêmes conditions (c'est-à-dire avec la même valeur de a) il s'élève à 9,007 pour Proust, 8,66 pour Giraudoux, 8,89 pour Chateaubriand, 8,85 pour la période 1860-1870 et respectivement 8,87 , 8,97 , 8,69 pour les tranches suivantes. Rappelons que l'indice évolue en sens inverse de la richesse lexicale et que les valeurs plus faibles de w indiquent une richesse plus grande. Le lexique de Zola paraît donc moins étendu que celui des écrivains de son époque, alors même que le genre romanesque est plus favorable que d'autres – et notamment le théâtre – à l'extension du vocabulaire. Et cela est vrai non seulement des *Rougon-Macquart* dans leur ensemble, mais aussi des romans individuels qui constituent le cycle. En choisissant *Nana*, qui sous ce rapport occupe une position moyenne chez Zola, tous les textes de Chateaubriand qu'on rapproche de *Nana* apparaissent plus riches, même les *Martyrs*.

La cause est donc entendue. Reste à expliquer notre illusion initiale que beaucoup de lecteurs de Zola ont sans doute partagée. On ne cherchera pas l'explication du côté du sujet, car Zola a varié ses thèmes expressément, veillant à ce que chaque roman découpe une face originale de la réalité sociale et abordant tour à tour le commerce, le rail, la banque, la terre, la mine, la guerre, etc. Il y a moins un recouvrement qu'une juxtaposition des thèmes et cela devrait favoriser le renouvellement lexical. La situation inverse s'observe précisément chez Proust qui ne s'écarte jamais très loin de son thème central et de ses obsessions. Pourquoi donc alors le vocabulaire de Proust est-il plus varié ? Cela ne tient pas non plus aux différences de registre : le langage populaire que Zola s'emploie parfois à reproduire tendrait à étendre le lexique et c'est ce qu'on observe dans les romans qui utilisent l'argot. L'explication nous paraît être dans un trait fondamental du naturalisme, dans la volonté de s'intéresser aux choses plutôt qu'aux mots,

de recenser la multiplicité des êtres et des réalités plutôt que d'explorer les ressources du langage, bref de découvrir des choses. La fabrication des mots a des ressources inépuisables quand on s'intéresse aux concepts et aux rapports. Composition et suffixation étendent à l'infini le champ lexical jusqu'à la frontière incertaine où le lexique touche à la syntaxe. Mais Zola fuit l'abstraction : toutes les variétés de suffixes qui représentent un procès, une qualité, un concept sont largement déficitaires dans les *Rougon-Macquart*, tandis que les suffixes qui désignent un agent ou un instrument sont en excédent. Si l'on établit la comparaison avec l'époque de Zola (soit les tranches du corpus du Trésor de la langue française qui vont de 1860 à 1907), on obtient le tableau 1.

Tableau 1

	Concept (substantifs)				Agent, instrument (substantifs)		
	Zola	époque	éc.réd.		Zola	époque	éc.réd.
tion	11 810	94 662	- 44,55	(ure)	11 363	60 049	5,60
(ment)	11 653	65 968	- 2,54	ée	8 139	43 992	2,49
té	11 376	75 183	- 20,78	ier	7 348	34 516	15,67
ie	11 232	66 771	- 8,22	aire	2 955	15 422	3,60
isme	373	3 941	- 14,01	oir	1 957	5 660	32,34
ence	5 086	31 715	- 9,30	ette	2 754	12 606	11,10
ance	4 388	30 105	- 15,65	euse	250	1 223	2,18
at	2 310	16 679	- 14,09				

Il y a certes des suffixes impurs qu'on hésite à ranger dans un camp ou dans l'autre (par exemple, *ment, ure, âge*), mais la tendance générale est claire, qui d'ailleurs se confirme parmi les adjectifs : les suffixes habituels à l'abstraction (ible, - 11; able, - 15; ique, - 30; el, - 28; al, - 12; if, - 8; aire, - 16; iste, - 17) étant rares sous la plume de Zola. Or il faut observer que, des deux espèces de suffixes, la variété abstraite est de beaucoup la plus sollicitée et la plus productive dans le discours. Et c'est elle qui crée le plus fort contingent de mots rares et contribue le plus à l'étendue du lexique. Comme Zola s'en éloigne, son lexique s'en trouve diminué d'autant, ce qui est paradoxal dans une oeuvre où foisonnent les objets, où l'on trouve toutes les variétés de fleurs, d'arbres, de fromages, de poissons, de viandes, d'outils, d'étoffes, toute la technologie du train et de la mine, toutes les techniques commerciales,

financières, agraires, tous les objets du culte et de l'art. En réalité son écriture fait penser à une encyclopédie non à un dictionnaire de langue.

Y a-t-il sous ce rapport quelque évolution de Zola ? Comme notre corpus recouvre une vingtaine d'années de rédaction et une vingtaine de textes régulièrement échelonnés (6), la chronologie offre à notre étude d'intéressantes opportunités. Le résultat apparaît dans le tableau 2 qui envisage chacun des textes du corpus et compare pour chacun le vocabulaire observé au vocabulaire théorique, calculé d'après l'étendue du texte, selon la loi binomiale (7). Sans s'attarder au phénomène bien connu de la spécialisation lexicale qui explique que tous les écarts réduits sont négatifs, on observera que la suite des

Tableau 2

La richesse lexicale chez Zola
(Loi binomiale)

Oeuvre	Rang	N	V absent	Vréel	Vthéo	Ecart	réduit	Rang
Raquin	1	68 937	13 062	4 606	6 275	- 1 669	- 25,63	19
Férat	2	102 348	11 922	5 397	7 415	- 2 018	- 29,84	22
Fortune	3	121 943	11 387	6 940	7 950	- 1 010	- 14,77	6
Curée	4	109 069	11 730	6 843	7 607	- 764	- 11,25	2
Ventre	5	114 532	11 580	6 977	7 757	- 780	- 11,44	3
Conquête	6	118 527	11 475	5 952	7 863	- 1 910	- 27,97	21
Faute	7	121 708	11 392	6 520	7 945	- 1 425	- 20,82	12
Excellence	8	131 251	11 156	7 212	8 181	- 969	- 14,10	5
Assommoir	9	165 017	10 422	7 383	8 915	- 1 532	- 22,10	14
Page	10	106 661	11 798	5 809	7 539	- 1 730	- 25,51	18
Nana	11	145 605	10 827	7 136	8 510	- 1 374	- 19,91	11
Pot-Bouille	12	139 977	10 953	6 590	8 384	- 1 794	- 26,04	20
Bonheur	13	153 491	10 657	7 371	8 680	- 1 309	- 18,92	10
Joie	14	123 628	11 344	6 404	7 993	- 1 589	- 23,21	15
Germinal	15	172 410	10 278	7 775	9 059	- 1 284	- 18,50	9
Oeuvre	16	138 301	10 991	7 864	8 346	- 482	- 7,00	1
Terre	17	171 492	10 296	7 955	9 041	- 1 086	- 15,65	8
Rêve	18	70 193	13 012	5 326	6 325	- 999	- 15,31	7
Bête	19	133 481	11 103	6 519	8 234	- 1 715	- 24,94	17
Argent	20	150 833	10 713	7 696	8 624	- 928	- 13,42	4
Débâcle	21	195 340	9 864	7 825	9 473	- 1 648	- 23,70	16
Pascal	22	120 011	11 436	6 469	7 901	- 1 432	- 20,95	13

écarts n'a rien de régulier et que la chronologie ne joue pas un rôle prépondérant. Avec une valeur de 0,12 le coefficient de Spearman ne permet pas de rejeter l'hypothèse nulle. C'est donc du côté du genre ou du thème qu'il faut chercher la loi de distinction. Les romans qu'on retrouve en tête (*Oeuvre* 1, *Curée* 2, *Argent* 3, *Ventre* 4, *Rêve* 6, *Terre* 7) ont une intention plus descriptive, alors que le récit et le dialogue se font plus vifs dans les derniers de la série (*Férat* 22, *Conquête* 21, *Raquin* 20, *Page d'Amour* 19, *Bête Humaine* 18 ou *Pot-Bouille* 17). Ce classement ne change guère lors qu'on utilise notre indice *w* avec $a = 0,185$: sur 22 textes 19 gardent strictement le même rang, trois textes voisins échangeant leur place.

La distinction obtenue entre ces textes à dominante descriptive et ceux où prévaut le récit a un retentissement dans la répartition des catégories grammaticales aussi bien que dans l'étendue du lexique. Les deux faits sont liés comme on s'en rendra compte en considérant la figure 1. Il s'agit d'une analyse factorielle, dite "de correspondance", qui distribue les textes sur le plan graphique selon qu'ils s'accordent avec telle ou telle catégorie grammaticale. La partie haute de la figure est occupée par les textes que la richesse lexicale avait placés en tête de liste. C'est là qu'on trouve à profusion les adjectifs, les substantifs et les prépositions. Au contraire les textes les plus pauvres se groupent dans la partie basse, là où dominent les verbes, les adverbes et les mots grammaticaux. On voit que la richesse lexicale est liée à la prédominance des catégories nominales et descriptives et que l'avènement du verbe – par quoi s'expriment l'action et le dialogue – marque un rétrécissement du lexique. Ce sont là des clivages déjà observés dans d'autres corpus et qui tiennent aux spécificités des genres littéraires. Car la richesse lexicale n'est pas un attribut indépendant comme le poids spécifique d'un élément chimique. Elle n'est qu'un aspect, qu'une manifestation parmi d'autres, d'une réalité sous-jacente qui structure le discours. Elle est liée à l'organisation syntaxique et à la répartition des catégories grammaticales comme elle est liée à la polarisation sémantique et aux tendances particulières qui orientent l'écrivain vers les choses ou vers les idées.

L'étendue du vocabulaire peut être aussi étudiée dans son mouvement chronologique, dans un saisie dynamique qui considère l'apport lexical de chaque texte. L'accroissement lexical du texte est une notion relative qui prend appui sur le vocabulaire déjà rencontré au moment où le texte *x* s'ajoute à ceux qui

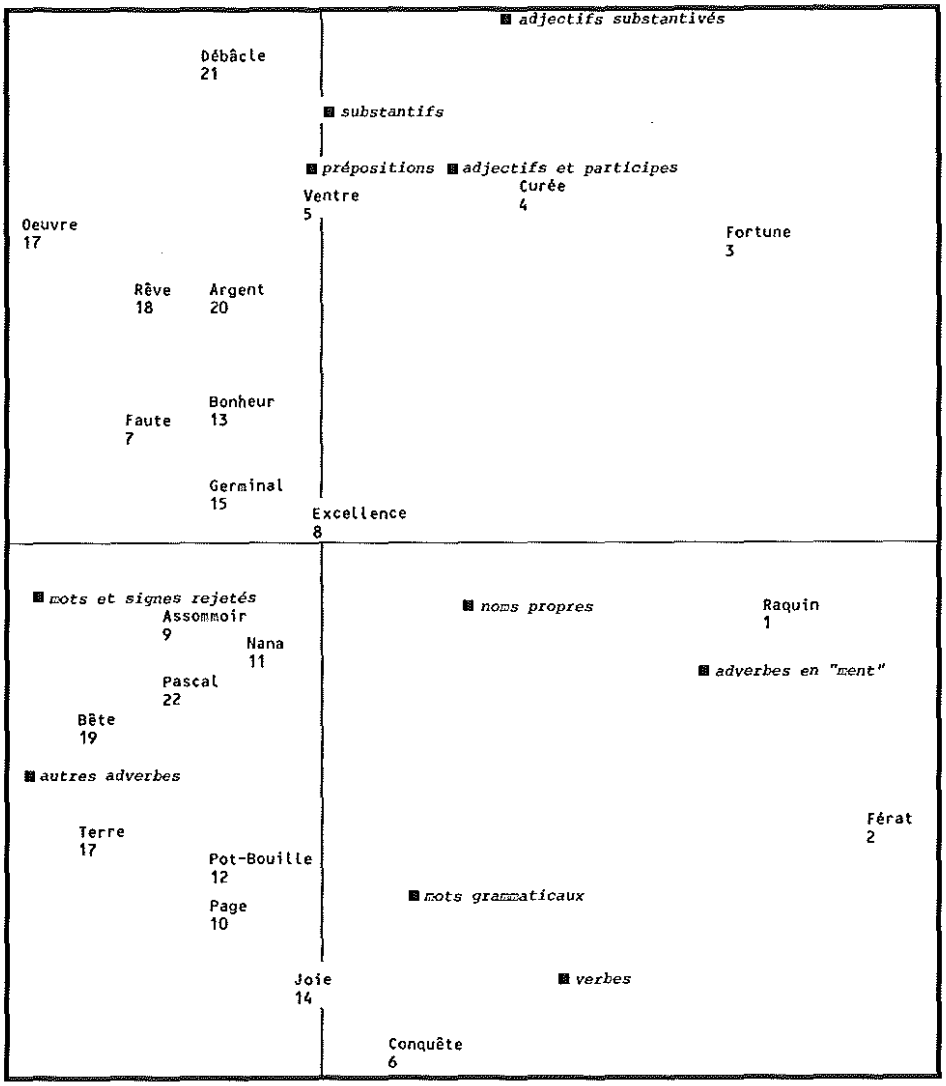

Figure 1

Catégories grammaticales
(Facteurs 1 et 2)

précèdent. Dans le cas du premier texte de la série l'accroissement, partant de zéro, se confond avec le vocabulaire, soit 4606 pour *Thérèse Raquin*. A partir du second texte, *Mademoiselle Férat*, le rapport entre le vocabulaire 5397 et l'accroissement 1984 s'élève à 2,72. Il ne peut que croître puisque les stocks s'épuisent et que le renouvellement du vocabulaire s'amenuise. Ainsi le 22e texte de la série, le *Docteur Pascal*, ne compte-t-il que 210 mots nouveaux sur les 6469 qu'il contient, soit un rapport de 30,80. Voici la suite de ce rapport :

Raquin	Férat	Fortune	Curée	Ventre	Conquête	Faute
1,00	2,72	3,05	4,57	5,53	11,38	8,27
Excellence	**Assommoir**	**Page**	**Nana**	**Pot-Bouille**	**Bonheur**	**Joie**
10,59	7,10	21,59	17,15	20,34	17,80	19,83
Germinal	**Oeuvre**	**Terre**	**Rêve**	**Bête**	**Argent**	**Débâcle**	**Pascal**
13,50	17,71	16,47	21,56	28,22	20,36	21,32	30,80

La série est croissante mais la régularité n'est pas parfaite : ainsi certains textes tranchent sur leur entourage dans un sens ou dans l'autre : dans le sens du renouvellement l'*Assommoir, Germinal*, la *Terre*, l'*Argent* et la *Débâcle* se mettent en relief, inversement l'accroissement lexical se ralentit dans la *Conquête de Plassans, Une Page d'Amour, Pot-Bouille,* la *Bête humaine* et le *Docteur Pascal*. Un programme d'ajustement de courbe permet de mesurer les écarts en choisissant le modèle adéquat (fonction puissance : y = a xb) on obtient une approximation excellente (r = 0,96), avec pour paramètres : a = 1,02 et b = 1,18. Voici les écarts obtenus :

Raquin	Férat	Fortune	Curée	Ventre	Conquête	Faute
+ 0,18	- 0,32	+ 0,57	+ 0,29	+ 0,58	- 4,01	+ 0,35
Excellence	**Assommoir**	**Page**	**Nana**	**Pot-Bouille**	**Bonheur**	**Joie**
- 0,70	+ 4,05	- 9,16	- 3,45	- 5,37	- 1,55	- 2,31
Germinal	**Oeuvre**	**Terre**	**Rêve**	**Bête**	**Argent**	**Débâcle**	**Pascal**
+ 5,29	+ 2,37	+ 4,89	+ 1,09	- 4,28	+ 4,86	+ 5,19	- 2,99

Dans les romans bourgeois le vocabulaire se répète, alors que les mots foisonnent quand Zola poursuit son enquête dans des milieux populaires.

Cependant notre procédé de calcul qui pondère l'accroissement par le seul élément V, ignore le facteur N ou étendue des textes, qu'il convient sans doute d'intégrer au calcul. Pour ce faire l'indice w peut être utile, en prenant soin toutefois d'inverser la formule :

$$w = N^{V^a}$$

et d'opérer sur les données cumulées, c'est-à-dire sur les valeurs que prennent N et V quand on parvient au texte x et qu'on a retenu tous ceux qui précèdent. Dans le tableau 3 qui donne le détail des calculs, (pour $a = 0{,}25$ et $w = 3{,}53$), on s'attachera à la colonne des écarts (pondérés par N). Le fait marquant est le renouvellement lexical qui accompagne l'*Assommoir* (rang 2), mais aussi le *Ventre de Paris* (3), l'*Oeuvre* (4), l'*Argent* (5), *Germinal* (6) et la *Terre* (7). On retrouve bien les textes que le calcul précédent avait relevés, mais parmi eux se glisse un texte que sa faible étendue avait défavorisé et qui occupe ici le premier rang : le *Rêve*. Dans le sens négatif les romans bourgeois prennent position comme précédemment mais la dernière place est occupée par les deux premiers textes *Raquin* et *Férat*, que le traitement précédent avait épargnés. On s'en tiendra là, en s'abstenant de recourir à la méthode binomiale qui, dans le cas de l'accroissement, souffre d'un biais dangereux.

On corroborera les résultats en inversant le cours de l'histoire, c'est-à-dire en lisant les *Rougon-Macquart* à rebours, en commençant par le dernier de la série, *Le Docteur Pascal*. Cette procédure, apparemment saugrenue, fait apparaître les thèmes qui ont cessé d'être exploités, les filons abandonnés, alors que la procédure normale est sensible au jaillissement des thèmes et aux nouveaux chantiers. On comparera les deux valeurs superposées dans la colonne des écarts du tableau 3 et les deux classements obtenus. Ainsi en remontant la chronologie on rencontre la *Terre* qui gagne trois places avant *Germinal* qui en perd 3. Et parmi les romans bourgeois, les derniers de la série (devenus les premiers) sont mieux classés, *Au bonheur des dames* passant de la 15e place à la 10e, et *Nana* de la 17e à la 12e, tandis que la *Fortune des Rougon* recule de 3 rangs et le *Ventre de Paris* de 4. Mais au total les deux perspectives se recouvrent largement, ce qui signifie que les romans constitutifs des *Rougon-Macquart* sont des unités largement autonomes et originales dont l'individualité s'affirme dans le sens chronologique comme dans l'autre, à l'image des montagnes volcaniques

Tableau 3

Accroissement du vocabulaire
(Indice W, pour A = 0,25)

	N Occurrences	V Vocables	A Accroissement	NC Occurrences cumulées	VC Vocables cumulés	T Occurrences théoriques cumulées	(NC - T)/N Ecart pondéré	Rang normal	Rang inverse	Ecart moyen	Rang moyen
Raquin	68 937	4 606	4 606	68 937	4 606	32 511	- 0,528	22	22	- 0,584	22
			45	2 874 755	19 337	2 874 255	- 0,640				
Férat	102 348	5 397	1984	171 285	6 590	86 043	- 0,477	21	21	- 0,496	21
			91	2 805 818	19 292	2 849 968	- 0,515				
Fortune	121 943	6 940	2 273	293 228	8 863	206 148	- 0,015	11	14	- 0,081	12
			195	2 703 470	19 201	2 800 365	- 0,147				
Curée	109 069	6 843	1 499	402 297	10 362	335 650	0,187	8	8	0,107	8
			217	2 581 527	19 006	2 696 391	0,027				
Ventre	114 532	6 977	1 261	516 829	11 623	486 223	0,315	3	7	0,262	6
			279	2 472 458	18 789	2 584 330	0,210				
Conquête	118 527	5 952	523	635 356	12 146	562 026	- 0,360	20	20	- 0,404	20
			136	2 357 926	18 510	2 445 737	- 0,448				
Faute	121 708	6 520	788	757 064	12 934	693 150	0,077	9	6	0,157	7
			325	2 239 399	18 374	2 380 361	0,237				
Excellence	131 251	7 212	681	888 315	13 615	824 588	0,001	10	11	- 0,017	10
			287	2 117 691	18 049	2 229 750	- 0,035				
Assommoir	165 017	7 383	1 040	1 053 332	14 655	1 061 983	0,439	2	2	0,486	2
			617	1 986 440	17 762	2 103 140	0,533				
Page	106 661	5 809	269	1 159 993	14 924	1 131 312	- 0,350	19	19	- 0,388	19
Nana	145 605	7 136	416	1 305 598	15 340	1 245 472	- 0,216	17	12	- 0,133	14
			376	1 714 762	16 986	1 789 153	- 0,051				
Pot-Bouille	139 977	6 590	324	1 445 575	15 664	1 340 502	- 0,321	18	17	- 0,312	18
			277	1 569 157	16 610	1 651 028	- 0,312				
Bonheur	153 491	7 371	414	1 599 066	16 078	1 470 138	- 0,155	15	10	- 0,0186	13
			463	1 429 180	16 333	1 554 776	- 0,017				
Joie	123 628	6 404	323	1 722 694	16 401	1 577 983	- 0,128	14	16	- 0,156	15
			332	1 275 689	15 870	1 403 828	- 0,184				
Germinal	172 410	7 775	576	1 895 104	16 977	1 785 743	0,205	6	9	0,098	9
			611	1 152 061	15 538	1 302 891	- 0,009				
Oeuvre	138 301	7 864	444	2 033 405	17 421	1 960 199	0,261	4	3	0,382	3
			850	979 651	14 927	1 132 105	0,503				
Terre	171 492	7 955	483	2 204 897	17 904	2 165 049	0,195	7	4	0,295	4
			1 188	841 350	14 077	924 272	0,395				
Rêve	70 193	5 326	247	2 275 090	18 151	2 276 178	0,583	1	1	0,584	1
			667	669 858	12 889	685 078	0,586				
Bête	133 481	6 519	231	2 408 571	18 382	2 384 168	- 0,191	16	15	- 0,181	16
			771	599 665	12 222	573 760	- 0,171				
Argent	150 833	7 696	378	2 559 404	18 760	2 569 640	0,230	5	5	0,269	5
			1 830	466 184	11 451	463 105	0,309				
Débâcle	195 340	7 825	367	2 754 744	19 127	2 760 538	- 0,023	12	13	- 0,040	11
			3 152	315 351	9 621	265 663	- 0,058				
Pascal	120 011	6 469	210	2 874 755	19 337	2 874 755	- 0,048	13	18	- 0,184	17
			6 469	120 011	6 469	81 641	- 0,320				
	2 874 755	19 337	19 337								

Ligne 1 : ordre chronologique
Ligne 2 : ordre inverse

qui ont le même aspect qu'on vienne de la droite ou de la gauche (8). On admirera en particulier combien l'*Assommoir* domine le paysage environnant de quelque côté qu'on l'aborde, et combien les deux premiers romans, extérieurs au cycle des *Rougon*, paraissent rabotés et sans relief (*Thérèse Raquin* 22e rang dans les deux séries et *Mademoiselle Férat* 21e).

Si on superpose les deux visées et qu'on fasse la moyenne des deux écarts, on mesure une sorte d'originalité lexicale relative à tous les textes considérés, qu'ils soient situés en deçà ou au delà sur l'axe chronologique. C'est le rang moyen qui figure dans la dernière colonne du tableau 3. D'autres approches sont possibles qui cernent cette notion d'originalité : l'étude des hapax ou celle de la fréquence 1 (9). Nous n'aborderons, rapidement, que ce dernier point de vue, qui fait appel à la loi binomiale à partir du tableau de distribution des fréquences. Sachant que dans le corpus de Zola on rencontre 3750 mots de fréquence 1, 1854 de fréquence 2, 1299 pour f = 3, 1001 pour f = 4, etc., on peut calculer dans chaque texte le nombre de mots de fréquence 1. Voici le classement obtenu :

Raquin	Férat	Fortune	Curée	Ventre	Conquête	Faute
22	21	8	4	3	18	13
Excellence	**Assommoir**	**Page**	**Nana**	**Pot-Bouille**	**Bonheur**	**Joie**
6	7	20	9	17	10	16
Germinal	**Oeuvre**	**Terre**	**Rêve**	**Bête**	**Argent**	**Débâcle**	**Pascal**
12	1	2	11	19	5	15	14

Reste à savoir ce que l'on mesure quand on s'attache à la fréquence 1. S'agit-il de la variété ou de l'originalité lexicale ? Observons que les mots de fréquence 1 (il faudrait préciser de sous-fréquence 1) ne sont pas nécessairement des hapax ou mots exclusifs. Bien au contraire les vrais hapax sont minoritaires parmi eux, puisque sur l'ensemble des 22 textes on compte un effectif (cumulé) de 22785 mots ayant la fréquence 1 dans l'un des textes, alors que le nombre des hapax s'élève à 3750. La logique voudrait donc que les hapax contribuent à l'originalité lexicale et soient en rapport direct avec l'accroissement du lexique et qu'au contraire la fréquence 1 soit le reflet de la richesse lexicale. C'est en effet ce que disent les chiffres : la corrélation de la fréquence 1 est plus forte avec la richesse lexicale (r = 0,91) qu'avec l'accroissement du lexique (r = 0,82). Quant à la liaison directe entre richesse et accroissement, le coefficient de Spearman montre qu'elle n'est pas très

élevée : r = 0,77. Si l'on compare de plus près les classements de la richesse et de l'accroissement, on voit apparaître en effet quelques distorsions notables. Certains textes, surtout au début des Rougon-Macquart, cultivent plutôt la richesse que l'originalité : c'est le cas de la *Fortune des Rougon* (rang 6 et 12 respectivement), de la *Curée* (2 et 8), du *Ventre de Paris* (3 et 6) et de *Son Excellence Eugène Rougon* (5 et 10). Mais quand Zola choisit un sujet hors des sentiers battus, c'est le contraire qui se produit, l'exemple le plus éclatant étant celui de l'*Assommoir*, dont la richesse est moyenne (rang 14) et l'originalité éclatante (2e). Il en est ainsi de la *Faute de l'abbé Mouret* (rangs 12 et 7), de la *Terre* (8 et 4), du *Rêve* (7 et 1) et de la *Débâcle* (16 et 11).

Pour mettre en relief la distinction qui sépare richesses et originalité il est utile de prendre appui sur le grand corpus du Trésor de la langue française, d'y repérer les mots rares et ceux qui le sont moins, et de relever les uns et les autres dans l'oeuvre de Zola. Pour permettre une étude plus fine, on a distingué dans le T.L.F. vingt groupes de mots (10) selon que la fréquence observée est de 1, 2, 3 à 4, 5 à 8, 9 à 16, etc. Le dernier groupe représente les mots très fréquents qui ont au moins 262144 occurrences. En doublant d'un groupe à l'autre le nombre de classes de fréquences envisagées, on a atténué l'arbitraire qui s'attache inévitablement à la notion de fréquences basses, moyennes ou hautes. Or qu'observons-nous chez Zola ? Le groupe 1 qui concerne les hapax du T.L.F. n'a que 167 représentants dans le corpus de Zola. La distribution en est très irrégulière :

Raquin	Férat	Fortune	Curée	Ventre	Conquête	Faute
0	1	1	4	7	2	15
Excellence	**Assommoir**	**Page**	**Nana**	**Pot-Bouille**	**Bonheur**	**Joie**
2	38	2	5	4	7	4
Germinal	**Oeuvre**	**Terre**	**Rêve**	**Bête**	**Argent**	**Débâcle**	**Pascal**
13	5	25	16	3	4	5	4

On voit que l'*Assommoir* s'approprie près du quart du lot (38 sur 167) et qu'il en est ainsi du groupe suivant (11) – fréquence 2 dans le T.L.F. – où la part de l'*Assommoir* est de 28 sur 148. Rien ne montre mieux l'exterritorialité de l'*Assommoir*, non seulement dans le cycle des *Rougon-Macquart*, mais dans la littérature française. Zola dans cette oeuvre cultive un domaine

nouveau du lexique et il en était lui-même si conscient qu'il proposé à ses lecteurs un glossaire explicatif. Le cas se reproduit à un moindre degré avec la *Terre* et *Germinal*. Dans la répartition des hapax ces deux textes viennent derrière *Germinal* avec respectivement 25 et 13 unités (et aussi 16 et 18 dans le groupe 2, 29 et 27 dans le groupe 3). Les raisons sont les mêmes, Zola explorant dans ces deux romans un univers (et donc un lexique) assez étranger au monde bourgeois qui prédomine dans les Lettres françaises.

Pour accéder à une vue synthétique des relations que les textes de Zola établissent avec le vocabulaire littéraire de ces deux derniers siècles (le corpus du T.L.F. étant accepté comme toile de fond), on a réalisé une analyse factorielle qui prend pour données l'effectif, dans chacun des textes de Zola, des vingt groupes de fréquences distinguées dans le T.L.F. Le résultat graphique est visible dans la figure 2. Les vingt groupes de fréquences formant une série, il est naturel qu'il soient liés les uns aux autres et qu'ils forment une chaîne du premier au dernier. Et c'est effectivement ce qu'on observe dans le graphique, où la progression est régulière qui part du haut à droite, s'oriente vers la gauche en descendant, pour revenir à droite dans la partie basse. Dans le quadrant supérieur droit se groupent les très basses fréquences et c'est là que l'*Assommoir*, la *Terre* et le *Rêve* prennent position. Le quadrant supérieur gauche, occupé par des fréquences basses, accueille le *Ventre, Germinal, la Curée, la Débâcle, l'Oeuvre* et *la Faute*. Puis en franchissant l'axe médian on aborde les fréquences moyennes où se cantonnent la *Fortune, Raquin, Son Excellence* et *Férat*. Enfin le trajet s'achève dans le quadrant inférieur droit où l'on retrouve, parmi les hautes fréquences, les textes qui ne spéculent nullement sur l'originalité et dont le cadre n'a rien qui puisse surprendre un lecteur ordinaire : *Une Page d'amour*, *La Joie de vivre*, *Pot-Bouille*, *La Conquête de Plassans* et *Le Docteur Pascal*. En réalité cette répartition reproduit assez fidèlement le classement que l'étude de l'accroissement lexical avait constitué. Et dans le même mouvement il s'écarte sensiblement de celui de la richesse lexicale.

La figure 3 donne une autre présentation des mêmes phénomènes, les vingt groupes de fréquence ayant été réduits à quatre où sont regroupés les fréquences très basses (de 1 à 128), basses (de 129 à 1024), moyennes (de 1025 à 16384) et hautes (plus de 16384). Dans le graphique des basses et très basses fréquences on remarquera les sommets atteints par

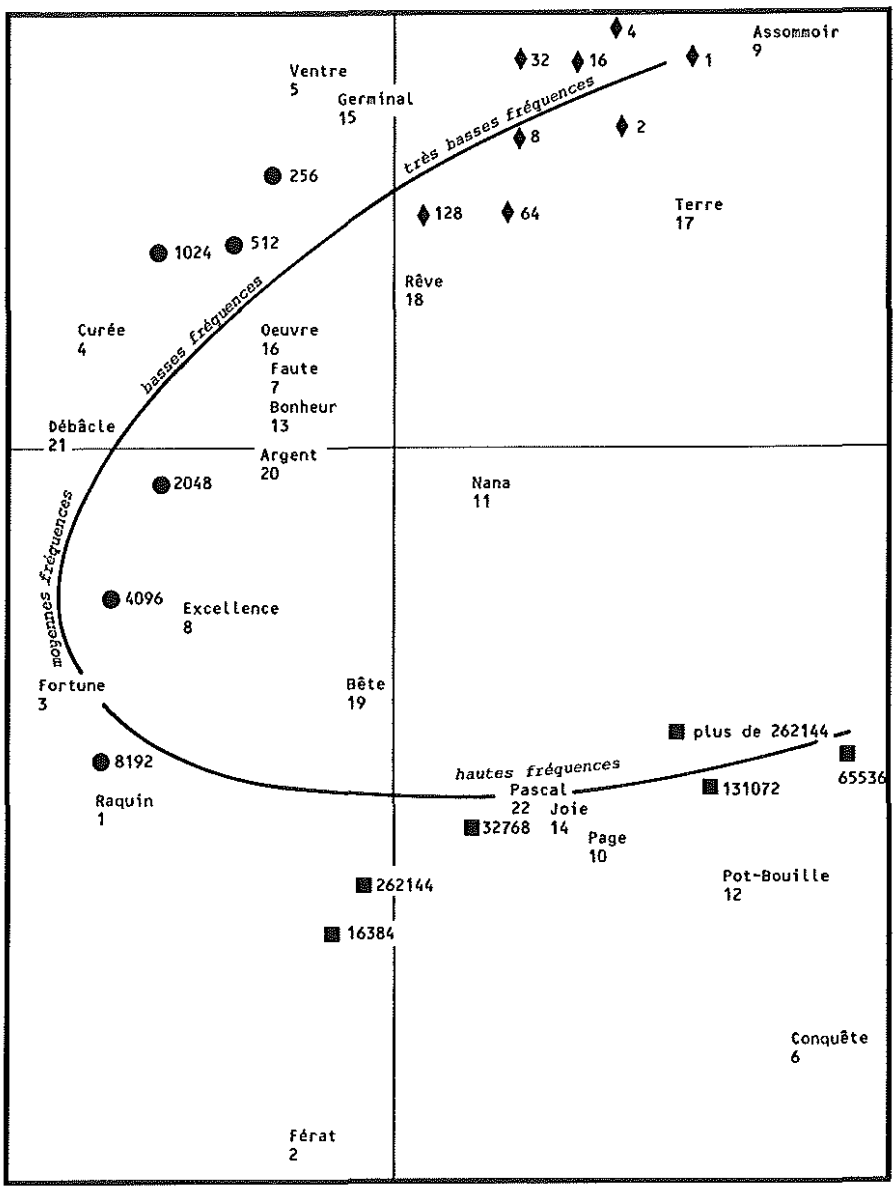

Figure 2

Analyse factorielle des groupes de fréquences

Les groupes de fréquence sont représentés par leur limite supérieure dans le grand corpus du T.L.F.

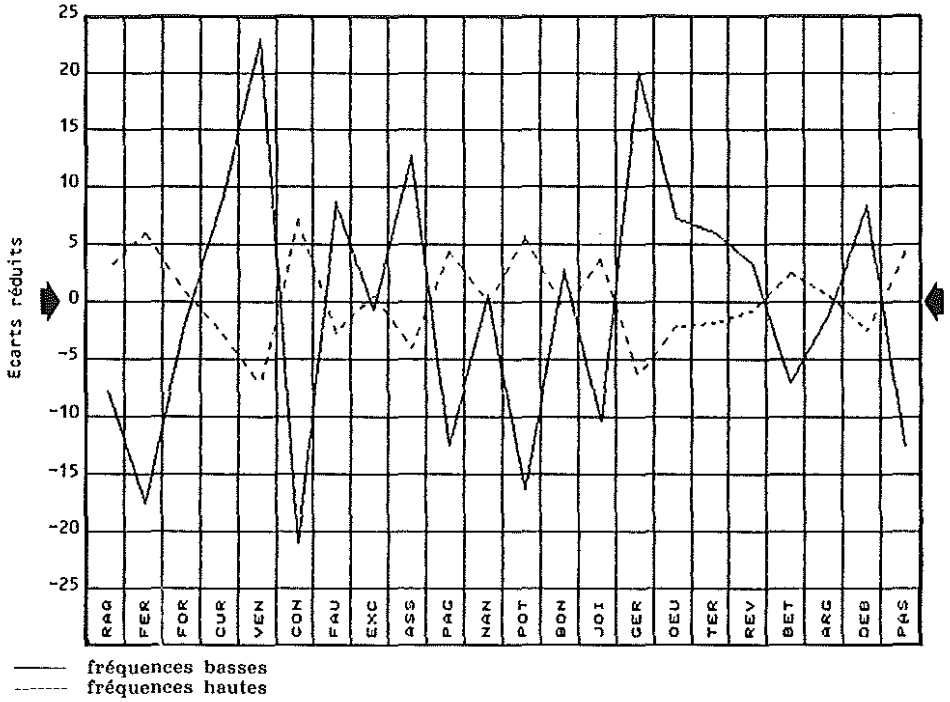

Figure 3

Courbe comparative des hautes et des basses fréquences

l'*Assommoir*, *Germinal*, le *Ventre de Paris* et la *Terre* et de façon corollaire les pics de la *Conquête*, de *Pot-Bouille*, de la *Joie* et du *Docteur Pascal* dans la courbe hautes fréquences.

Reste à expliquer en quoi richesse et originalité sont deux notions distinctes et pourquoi les deux mesures donnent des résultats divergents dans le cas de Zola. L'originalité du lexique est une notion relative, qui prend en compte l'identité des vocables d'un texte et leur présence (ou leur absence) dans les textes comparés. La richesse lexicale au contraire est un rapport interne à un texte et sans référence à l'extérieur. Ce rapport, établi entre les paramètres *N* et *V*, fait abstraction de l'identité des vocables et n'intéresse que la structure du vocabulaire. Encore n'est-ce qu'une mesure globale qui peut

recouvrir des situations fort différentes. Dans la définition de la richesse lexicale les mots fréquents comptent autant que les basses fréquences puisque le rapport est établi entre l'étendue du texte *N* (où les mots fréquents jouent un grand rôle), et l'étendue du texte *V* (où les mots rares sont déterminants). La richesse est élevée quand *V* est grand et *N* petit (relativement); c'est le cas observé chez Giraudoux. Elle est faible quand V est petit et N grand : c'est le cas de Corneille. Mais il peut y avoir des situations mixtes : par exemple beaucoup de mots fréquents et beaucoup de mots rares (c'est l'exemple de Proust) ou bien peu des uns et des autres et Zola se situe précisément dans ce dernier cas.

Observons en effet la répartition des vingt groupes de fréquence non plus dans chacun des textes de Zola mais dans l'ensemble des *Rougon-Macquart*. Les écarts observés sont repris au tableau 4.

Tableau 4

Groupe	Plage de fréquence dans le T.L.F.	Effectif observé	Effectif théorique	Ecart réduit
1	1	167	867	- 24,28
2	2	193	545	- 15,40
3	3-4	396	850	- 15,92
4	5-8	741	1 408	- 18,16
5	9-16	1 425	2 453	- 21,19
6	17-32	3 666	4 592	- 13,97
7	33-64	7 279	8 536	- 13,90
8	65-128	14 254	16 184	- 15,50
9	129-256	27 869	29 083	- 7,27
10	257-512	48 496	46 278	10,53
11	513-1024	72 400	68 307	15,99
12	1025-2048	109 057	100 682	26,95
13	2049-4096	144 926	142 411	6,80
14	4097-8082	180 504	175 190	12,96
15	8093-16384	174 214	182 424	- 19,63
16	16385-32768	199 520	181 691	42,71
17	32769-65536	201 268	199 815	3,32
18	65537-131072	159 201	158 005	3,07
19	131073-262144	175 556	179 803	- 10,23
20	plus de 263 144	1 353 730	1 390 555	- 31,89

Déficit dans les fréquences basses et hautes, excédents dans les fréquences moyennes, le choix de Zola est celui de la poésie. Et ce rapprochement qu'impose la structure lexicale est suggéré aussi par l'étude du contenu. Comme Zola les poètes fuient l'abstraction et montrent le corps et le monde physique. Comme Zola ils privilégient la catégorie nominale et particulièrement l'adjectif. Zola, dans beaucoup des traits accessibles à l'analyse quantitative, apparaît donc comme l'héritier de Hugo et de Balzac, et le naturalisme comme prolongement stylistique du romantisme.

Notes

(1) *Son Excellence Eugène Rougon*, la Pléiade, tome 2, p. 253.

(2) *Pot-Bouille*, la Pléiade, tome 3, p. 85.

(3) *La Conquête de Plassans*, la Pléiade, tome 1, p. 1127.

(4) Nous devons avouer que nous les avons comptés dans le grand corpus du Trésor de la langue française et nous en avons trouvé 25 137 577 mais, moins "fêlé" que Mouret, nous n'avons pas recommencé le calcul trois fois. Voir notre *Vocabulaire français de 1789 jusqu'à nos jours*, tome 1, p. 216, éditions Slatkine.

(5) *Proust et Giraudoux*, Revue d'Histoire littéraire de la France, n° 5-6, 1983, p. 829.

(6) Les *Rougon-Macquart* comportent 20 textes dont le premier paraît en 1871 et le dernier en 1893. Notre étude ajoute deux romans antérieurs : *Thérèse Raquin* (1867) et *Madeleine Férat* (1868). Le tout représente 19.342 vocables et 2.874.862 occurrences (ou bien respectivement 19.337 et 2.874.755 dans un état antérieur du corpus, non entièrement expurgé).

(7) Une fois de plus nous renvoyons sur ce point aux ouvrages de Charles Muller.

(8) La perspective inverse donne le moyen de supprimer le biais mathématique qui est à craindre lorsque la valeur des paramètres a et w est mal ajustée et favorise les premiers ou les derniers de la série. L'ajustement des paramètres a été fait de telle sorte que l'équilibre est réalisé entre la perspective chronologique et la perspective inverse. Ainsi la somme des rangs des 11 premiers titres est égale à 142 dans les deux sens et celle des 11 derniers est aussi identique (111 dans les deux séries). De plus le coefficient de Spearman a la même valeur (r = 0,30) dans les deux cas et l'ajustement linéaire a les mêmes paramètres (a = - 30 et b = 15). Enfin la corrélation directe entre les deux classements est extrêmement forte : r = 0,92.

(9) On a relevé 3750 hapax chez Zola, entendons par là 3750 vocables employés une seule fois dans les *Rougon-Macquart*. Mais parmi les

hapax du grand corpus du Trésor de la langue française 167 seulement (sur 21193) se trouvent chez Zola, soit beaucoup moins que l'espérance mathématique 867 et cela confirme l'étroitesse relative de son lexique.

(10) Voir notre *Vocabulaire français de 1789 à nos jours*, tome 1, p. 105.

(11) Ce privilège de l'*Assommoir* se maintient dans les groupes suivants :
fréquences 3 et 4 : 58 sur 251
fréquences 5 à 8 : 76 sur 428
fréquences 9 à 16 : 130 sur 713
fréquences 17 à 32 : 224 sur 1250
En réalité tous ces groupes représentent les très basses fréquences.