On a compté trois millions de mots chez Zola. Et alors ?

Étienne Brunet

On a compté trois millions de mots chez Zola. Et alors ?[1]

Etienne Brunet

On a compté trois millions de mots chez Zola, ou plus exactement dans les *Rougon-Macquart*[2]. Evidemment certains penseront que pour rendre compte d'un texte, il vaut sans doute mieux lire les mots que les compter et Zola, pourtant acquis aux thèses positivistes, est probablement de cet avis, puisqu'il ridiculise la statistique dans l'innocente manie de Mouret :

> *vous n'imagineriez pas à quoi Mouret passe le 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 ! »* (*La Conquête de Plassans*, la Pléiade, tome 1, p. 1127).

Or Mouret n'est qu'un amateur, comme M. Vabre qui de la même façon répertorie les tableaux, dans *Pot-Bouille,* et les soumet, « sans les voir », aux décomptes, aux tris et aux classements. Qu'eût pensé Zola des professionnels modernes qui traitent les textes avec des machines, qui n'ont même plus besoin de lire pour compter et qui poussent la prétention blasphématoire jusqu'à s'attaquer à la Bible, comme Mouret, et aux grands textes comme ceux de Shakespeare, de Proust et de Zola lui-même ?

En réalité nous ne sommes pas certain que Zola n'eût pas été fasciné, s'il avait pu le connaître, par la nouvelle idole. Lui qu'on a vu si sensible aux mutations techniques, aux changements de civilisations, à l'avènement de la grande industrie, à l'envol du grand commerce, aux

---

1. Article publié dans *Computers in literary and linguistic Computing* , Champion-Slatkine, 1985, p. 63-91. Seule la mise en page a été revue et corrigée.

2. Soit un ensemble de vingt titres, de la *Fortune des Rougon* au *Docteur Pascal*, à quoi l'on a ajouté deux romans antérieurs, extérieurs au cycle : *Thérèse Raquin* et *Madeleine Férat.*

combats de la haute finance, au déploiement des grands travaux, n'eût pas été indifférent au mythe nouveau du grand calcul. Si cent ans exactement après *Germinal*, à l'orée du XXI^e^ siècle, il était donné à Zola d'ajouter un 21^e^ volume à la série des *Rougon-Macquart*, on peut penser qu'il le consacrerait au Minotaure moderne, l'ordinateur, et qu'un nouveau monstre serait introduit dans la galerie zolienne, à côté du Voreux de Germinal.

I . L'Index des *Rougon-Macquart*

Quoi qu'il en soit, que reste-t-il de Zola après que le Voreux électronique a englouti toute son œuvre ? Apparemment rien d'autre qu'un tas de mots réduits en poussière, cette sorte de terril lexical qu'on appelle un index. Notre index des *Rougon-Macquart*, même limité à un million de références, est presque aussi volumineux que l'édition de la Pléiade où le texte des *Rougon-Macquart* s'étend sur 5 volumes et 7154 pages. Même en écartant une centaine de mots grammaticaux, même en tassant au maximum les informations jusqu'à la limite de la lisibilité, il reste plus de 4000 pages qui pour rester légères ont emprunté le voile de la microfiche. Nous n'en montrerons, de façon fugitive, que la première page (tableau 1), où les options de notre *Index de Proust*[3] ont été reprises : lemmatisation, présentation synoptique et tests statistiques.

Pour chaque vocable et chaque forme, on restitue les références habituelles (tome, titre, page, zone de la page) sous la forme d'un histogramme renversé qui reproduit la distribution du mot. L'index est riche de données statistiques, soit brutes (fréquence dans l'ensemble et dans chacun des 22 textes), soit élaborées (écarts réduits confrontant d'une part chaque partie à l'ensemble, d'autre part le corpus de Zola au corpus général du *Trésor de la langue française*, et au corpus plus spécifique de la fin du XIX^e^ siècle). Deux informations ont été ajoutées, qui n'existaient pas dans l'index de Proust : un test de répartition (un écart-type) et un coefficient de corrélation chronologique (Bravais-Pearson). Ainsi pour chaque mot peut-on mesurer l'originalité de Zola, sa régularité et son évolution.

3. *Le Vocabulaire de Proust*, 1983, éditions Slatkine, Genève, vol. 1, 261 p., vol. 2 et 3, 1644 p.

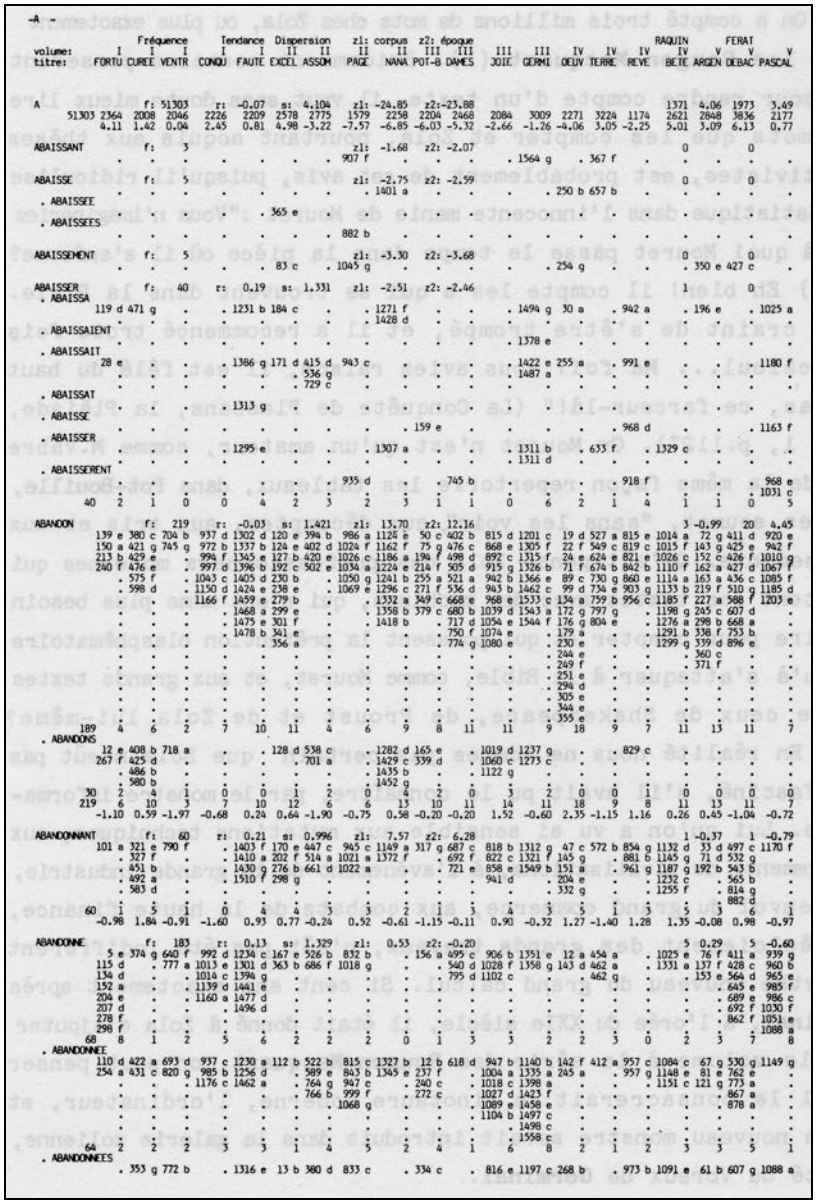


Tableau 1. Extrait de l'index de Zola

Mais nous nous garderons de les explorer un à un. Les limites étroites de cet exposé nous condamnent à un parcours précipité et à un discours de prétérition.

Aussi bien nous ne dirons rien des méthodes qu'on a utilisées ici et qui conviennent aux grands corpus. Comme notre étude se situe dans les grands nombres, la loi normale parait devoir fournir une approximation acceptable et, faute d'autre modèle, on a eu recours au schéma d'urne – même s'il est contesté par certains – car l'étendue du corpus, la constance des normes de dépouillement, l'homogénéité d'un ensemble fermé, constitué autour d'un même auteur, d'un même genre et d'une même époque, offrent des garanties suffisantes pour la validité des méthodes statistiques traditionnelles, que confirment d'ailleurs d'autres méthodes, dont certaines ne sont pas paramétriques. En particulier dans un ensemble de 22 objets ordonnés par la chronologie, la corrélation des rangs est une arme efficace, comme aussi l'analyse factorielle dont l'utilisation est ici systématique.

On ne s'appesantira pas sur les difficultés – surmontées non sans mal – de la lemmatisation. Les options prises sont celles de l'Institut de la langue française qui, malgré leur imperfection, sont les seules à permettre, dans le domaine français, la compatibilité et la comparaison. On ne s'attardera pas non plus sur le dépouillement, sinon pour évoquer les problèmes techniques que pose le passage d'une édition à l'autre et qui ont reçu une solution originale. Si les 13 premiers textes des *Rougon-Macquart* ont été enregistrés dans l'édition de la Pléiade (3 premiers tomes), les 7 derniers romans, qui avaient été dépouillés à partir d'une édition antérieure, ont été confrontés au texte de la Pléiade et l'on a procédé à l'ajustement de 3000 pages par un procédé semi-automatique.

II . La structure lexicale

1. On évitera de s'engager trop avant dans les questions de richesse et d'originalité lexicales qui ont monopolisé trop souvent les efforts de la linguistique quantitative. Si nous les abordons cependant ici, c'est parce qu'elles nous ont apporté une surprise. Qui n'a en mémoire ces pages descriptives, où Zola accumule les variétés de fromage ou d'étoffe, les espèces de poissons, de fleurs ou de légumes, au point que la richesse semble poussée jusqu'à l'exhaustivité et que Zola parait vouloir concurrencer, non l'état civil, mais le dictionnaire ? Beaucoup parieront que le vocabulaire de Zola est riche et nous avions fait imprudemment ce pari en déclarant dans notre étude de Proust : « La soumission au réel

peut engendrer la multiplication documentaire du lexique, comme cela se produit chez Zola. » Or l'ordinateur, qu'on accuse parfois de n'apporter que la confirmation quantitative de l'évidence, nous oppose ici un démenti formel. Les chiffres montrent de façon claire que le lexique de Zola est plus limité que celui de Proust, de Giraudoux, de Chateaubriand, pour ne citer que les écrivains dont nous avons exploré les données.

La démonstration en est simple et se réduit à la seule comparaison des paramètres N (étendue du texte) et V (étendue du vocabulaire). Ainsi la *Prisonnière* qui est le plus pauvre des textes de la *Recherche du temps perdu* l'emporte pourtant en richesse sur les romans de Zola l'*Assommoir*, *Germinal*, la *Terre*, la *Débâcle* dont l'étendue est plus grande (N = 165.017, 172.410, 171.492 et 195.340 respectivement contre 164.683 pour la *Prisonnière*) et le vocabulaire plus restreint (V = 7.383, 7.775, 7.955 et 7.825 respectivement contre 8.633).

L'écart s'accroît encore lorsque la comparaison est faite avec Giraudoux ou Chateaubriand, qu'on envisage le corpus entier de Zola ou les romans individuels, qu'on utilise la loi binomiale ou des formules plus spécifiques (l'indice *w* par exemple), ou tout simplement, dans les cas favorables, les seules indications de N et V.

Comment expliquer cette sobriété lexicale qui ne correspond à aucune volonté expresse de Zola et qui s'accorde mal avec l'impression du lecteur ? Cela ne tient pas au choix des thèmes, qui sont fort variés dans les *Rougon-Macquart* et qui auraient dû conduire à la multiplication des mots. On n'invoquera pas davantage le choix du registre car les excursions que Zola tente du côté du langage populaire ne pouvaient qu'enrichir son vocabulaire. L'explication gît dans la nature du vocabulaire zolien qui est essentiellement concret. Lorsqu'il s'agit de désigner les réalités matérielles et sociales, les inventaires, même techniques, ne sont pas infinis et les dictionnaires encyclopédiques ont des limites. Mais lorsque l'écrivain s'engage dans la voie de l'invention verbale, en exploitant les ressources illimitées de la suffixation et de la composition, les bornes du lexique reculent indéfiniment en empiétant sur la syntaxe. Quand par exemple Proust risque les mots *mélancolieusement* ou *inefficacement*, la création lexicale se substitue aux procédés ordinaires de la syntaxe.

OEUVRE	rang	N	V réel	réduit	rang	accrois normal inverse	coeff normal inverse	coeff. moyen	rang
Raquin	1	68937	4606	-25.63	19	4606 45	-0,528 -0,640	-0,584	22
Férat	2	102348	5397	-29.84	22	1984 91	-0,477 -0,515	-0,496	21
Fortune	3	121943	6940	-14.77	6	2273 195	-0,015 -0,147	-0,081	12
Curée	4	109069	6843	-11.25	2	1499 217	0,187 0,027	-0,107	8
Ventre	5	114532	6977	-11.44	3	1261 279	0,315 0,210	0,262	6
Conquête	6	118527	5952	-27.97	21	523 136	-0,360 -0,448	-0,404	20
Faute	7	121708	6520	-20.82	12	788 325	0,077 0,237	0,157	7
Excellence	8	131251	7212	-14.10	5	681 287	0,001 -0,035	-0,017	10
Assommoir	9	165017	7383	-22.10	14	1040 617	0,439 0,533	0,486	2
Page	10	106661	5809	-25.51	18	269 159	-0,350 -0,427	-0,388	19
Nana	11	145605	7136	-19.91	11	416 376	-0,216 -0,051	-0,133	14
Pot-Bouille	12	139977	6590	-26.04	20	324 277	-0,321 -0,312	-0,312	18
Bonheur	13	153491	7371	-18.92	10	414 463	-0,154 -0.017	-0,086	13
Joie	14	123628	6404	-23.21	15	323 332	-0,128 -0,184	-0,156	15
Germinal	15	172410	7775	-18.50	9	576 611	0,205 -0.009	0,098	9
Oeuvre	16	138301	7864	-7.00	1	444 850	0,261 0,503	0,382	3
Terre	17	171492	7955	-15.65	8	483 1188	0,195 0.395	0,295	4
Rêve	18	70193	5326	-15.31	7	247 667	0,583 0,586	0,584	1
Bête	19	133481	6519	-24.94	17	231 771	-0,191 -0,171	-0,181	16
Argent	20	150833	7696	-13.42	4	378 1830	0,230 0,309	0,269	5
Débâcle	21	195340	7825	-23.70	16	367 3152	-0,023 -0.058	-0,040	11
Pascal	22	120011	6469	-20.95	13	210 6469	-0,048 -0,320	-0,184	17

Tableau 2. La richesse lexicale chez Zola (loi binomiale)

Or Zola s'abstient de recourir à cette invention lexicale qui fait du neuf avec du vieux et qui n'est qu'une combinatoire de radicaux et d'infixes. Il fuit surtout le langage abstrait, qui abuse de la suffixation. Une étude des différentes espèces de suffixes nous montre combien Zola, accueillant aux variétés qui représentent un agent ou un instrument, répugne par contre à utiliser celles qui désignent un concept, une qualité ou un procès. En prenant appui sur le corpus du *Trésor de la langue française* et plus particulièrement sur les tranches du corpus qui vont de 1860 à 1907, on obtient la répartition suivante :

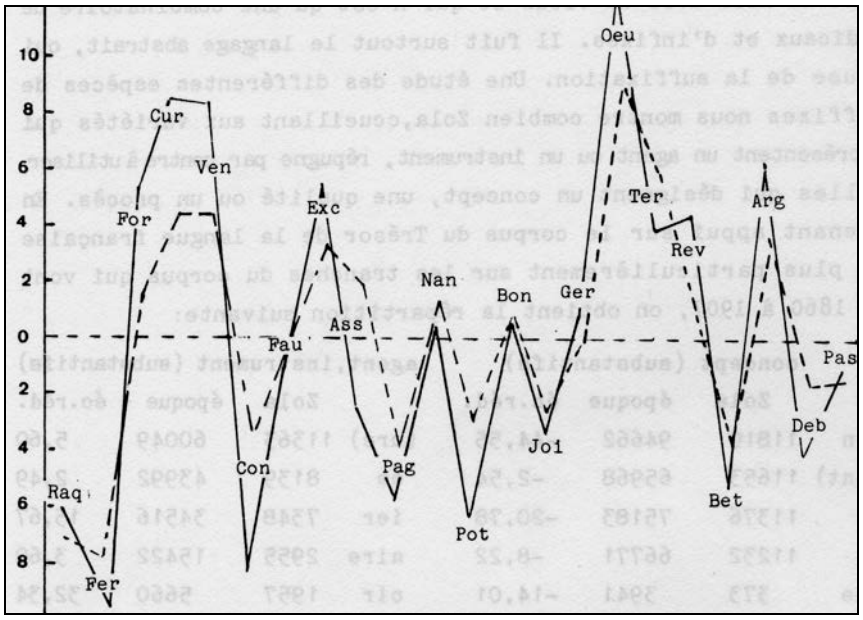
	concept (substantifs)				agent, instrument (substantifs)		
	Zola	époque	éc.réd.		**Zola**	époque	éc.réd.
tion	11810	94662	-44,55	(ure)	11363	60049	5,60
(ment)	11653	65968	-2,54	ée	8139	43992	2,49
té	11376	75183	-20,78	ier	7348	34516	15,67
ie	11232	66771	-8,22	aire	2955	15422	3,60
isme	373	3941	-14,01	oir	1957	5660	32,34
ence	5086	31715	-9,30	ette	2754	12606	11,10
ante	4388	30105	-15,65	euse	250	1223	2,18
at	2310	16679	-14,09				

Certains de ces suffixes ne sont pas d'une pureté absolue et l'on hésite à les ranger dans un camp ou dans l'autre – c'est le cas de *ment*, *ure*, *âge* – mais le choix de Zola est clair, qui s'exerce pareillement parmi les adjectifs : ceux où s'exprime quelque processus abstrait sont déficitaires dans les *Rougon-Macquart : ible* -11, *able* -15, *ique* -30, *el* -12, *if* -8, *aire* -16, *iste* -17. Or la catégorie abstraite est de loin la plus productive et la plus sollicitée dans le discours. En s'en abstenant, Zola se prive du même coup des facilités qui contribuent à la « richesse » du lexique. Mais peut-être faudrait-il distinguer dans le trésor lexical l'encaisse-or et la monnaie fiduciaire. Zola ne prise que la première, qui est fondée sur des garanties concrètes, et se méfie de la seconde, qui est issue de la planche à billets. Zola ne contribue en rien à l'inflation lexicale qu'on constate dans le discours depuis 1789. Peu soucieux d'inventer des mots, de se payer de mots, il se préoccupe davantage de découvrir les réalités et de les peindre. L'impression d'abondance que donne la lecture de son oeuvre n'est donc pas tout à fait illusoire : elle ne tient pas au langage ou aux curiosités lexicales mais à la richesse d'un univers, d'une arche de Noé où toute la création se trouve rassemblée, ce qui était bien l'ambition de Zola : « Je voudrais coucher l'humanité sur une page blanche, tous les êtres, toutes les choses ; une oeuvre qui serait l'arche immense. »[4]

2. Quant à l'évolution de Zola sous ce rapport, on peut la saisir à différents niveaux, soit qu'on s'intéresse à la variété lexicale des différents textes des *Rougon-Macquart*, soit qu'on envisage plutôt l'originalité du vocabulaire de chacun des textes comparés, c'est-à-dire l'accroissement

4. *Oeuvres complètes*, la Pléiade, tome IX, p. 351.

du vocabulaire[5]. Le tableau 2 réunit les deux études qui trouvent leur expression graphique dans les graphiques 1 et 2.

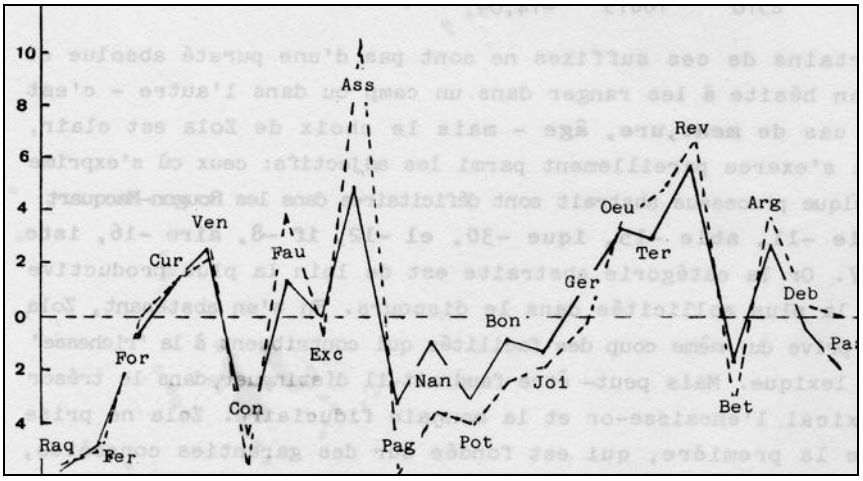


Graphique 1. Richesse (en traits pleins) et fréquence 1 (en pointillés)

Deux enseignements principaux peuvent âtre retenus : d'une part les deux courbes s'accordent pour séparer les textes de Zola qui privilégient la description (*Curée*, *Ventre de Paris*, *Assommoir*, *Oeuvre*, *Terre*, *Rêve*, *Argent*) de ceux où dominent la narration et l'action (*Thérèse Raquin*, *Madeleine Férat*, *Conquête*, *Page d'amour*, *Pot-Bouille*, *Joie de vivre*, *Bête humaine*, *Docteur Pascal*). Mais d'autre part des divergences importantes apparaissent à certains endroits. Certains textes, surtout au début du cycle, prétendent plutôt à la richesse qu'à l'originalité (c'est le cas de la *Fortune*, de la *Curée*, du *Ventre de Paris* et de *Son Excellence Eugène Rougon* qui occupent les rangs respectifs 6, 2, 3 et 5 pour la

5. L'accroissement lexical peut à son tour être saisi selon deux perspectives : l'une, normale, suit le cours du temps, et l'autre, inverse, remonte la chronologie. Nous avons suivi les deux voies successivement mais ne donnons ici qu'un résultat unique qui fait la moyenne des deux valeurs obtenues et mesure une originalité globale, dirigée vers l'amont comme vers l'aval. On a utilisé l'indice *w* pour cette étude, la loi binomiale étant appliquée à celle de la richesse et de la fréquence 1, tandis que l'étude des hapax relève de la loi normale. On a accordé l'échelle des différents graphiques en les centrant autour d'une moyenne 0 ou en leur appliquant un facteur 10.

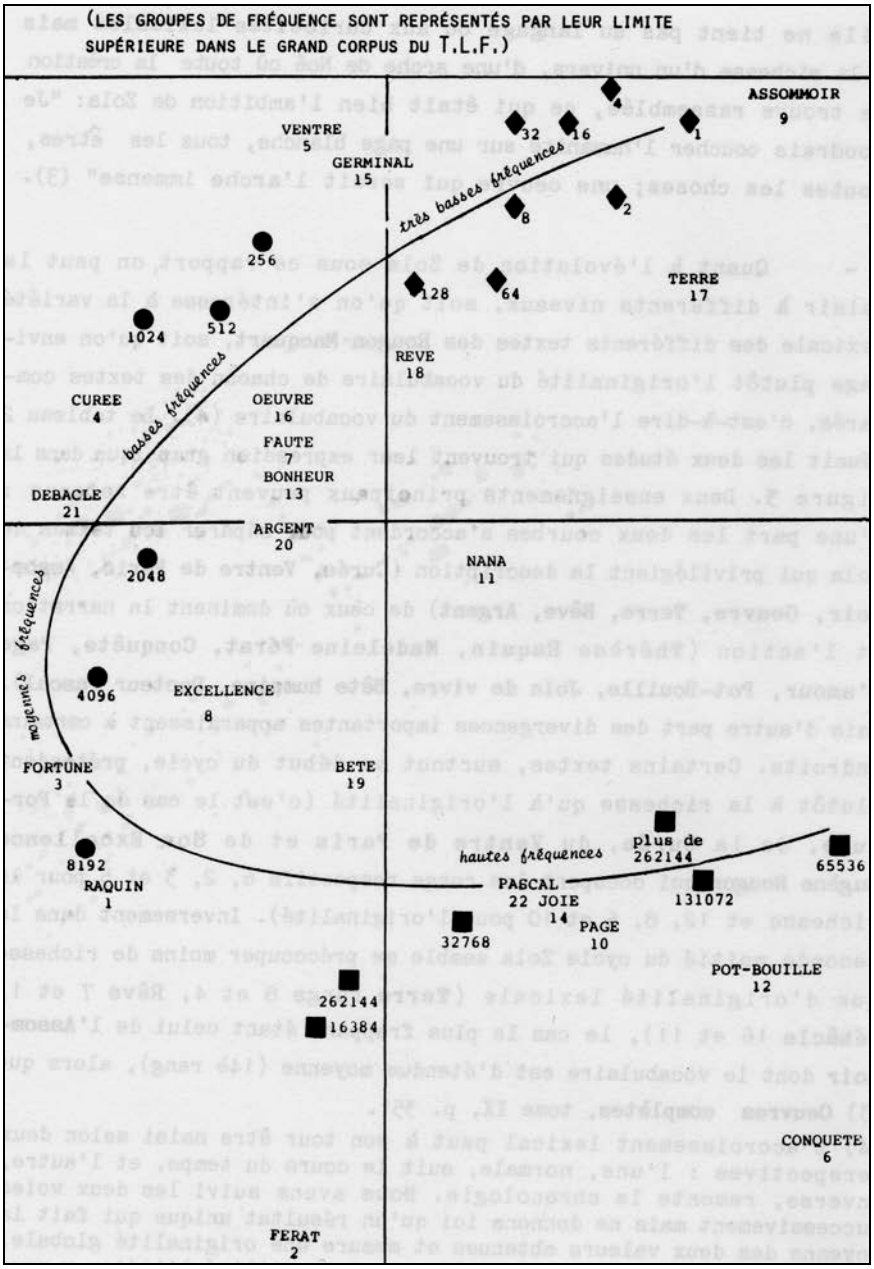
richesse et 12, 8, 6 et 10 pour l'originalité). Inversement dans la seconde moitié du cycle, Zola semble se préoccuper moins de richesse que d'originalité lexicale (*Terre* rangs 8 et 4, *Rêve* 7 et 1, *Débâcle* 16 et 11), le cas le plus frappant étant celui de l'*Assommoir* dont le vocabulaire est d'étendue moyenne (14^{e} rang), alors que le renouvellement lexical y est considérable (rang 2). Et l'exterritorialité de *l'Assommoir* est encore mieux marquée quand on mène l'enquête du côté des hapax (graphique 2). C'est que l'*Assommoir* explore un milieu et un langage peu familiers au roman bourgeois et il en est de même – à un degré moindre – de la *Terre* et de *Germinal.*



Graphique 2. Accroissement (en traits pleins) et hapax (en pointillés)

3. On aura une illustration très claire de cette spécificité des romans « populaires » de Zola si l'on examine l'analyse factorielle réalisée, dans le graphique 3, à partir de la distribution dans les *Rougon-Macquart* de 20 groupes de fréquence.

Ces groupes sont extraits du grand corpus du *T.L.F.* de telle sorte que le nombre de classes de fréquence considérées double à chaque pas : 1, 2, 3 à 4, 5 à 8, 9 à 16, etc. Les très basses fréquences se cantonnent dans le quadrant supérieur droit du graphique, là où prennent place les grandes fresques populaires *Asommoir*, *Terre*, *Germinal* et *Ventre* de *Paris.* Puis les fréquences moyennes apparaissent progressivement lorsque la chaîne se porte à gauche au voisinage de l'axe des *x*. Et la forme en croissant, caractéristique des données sérielles, s'achève à droite, dans le quadrant inférieur, où se concentrent les hautes fréquences, dans l'entourage des romans bourgeois traditionnels.



Graphique 3. Analyse factorielle des groupes de fréquences

Ces mêmes groupes de fréquence (réduits à 12) nous permettent en outre de préciser à quoi tient la sobriété lexicale de Zola. En comparant les effectifs observés chez Zola à ceux du corpus, on obtient les écarts suivants :

groupe	plage de fréquences	effectif dans TLF	effectif chez Zola	effectif théorique	écart réduit
1	1-2	34523	360	1412	-28,59
2	3-8	55216	1137	2258	-24,10
3	9-32	172241	5091	7046	-23,78
4	33-128	604225	21533	24717	-20,68
5	129-512	1842220	76365	75361	3,73
6	513-2048	4130986	181457	168990	30,97
7	2049-8092	7763827	325430	317602	14,18
8	8093-32768	8900871	373734	364116	16,27
9	32769-65536	4884506	201268	199815	3,32
10	65537-131072	3862455	159201	158005	3,07
11	131073-262144	4395316	175556	179803	-10,23
12	+ de 263144	33992266	1353730	1390555	-31,89

Les déficits sont dans les fréquences basses mais aussi dans les hautes fréquences, Zola utilisant de préférence les moyennes fréquences, ce qui est aussi le choix des poètes.

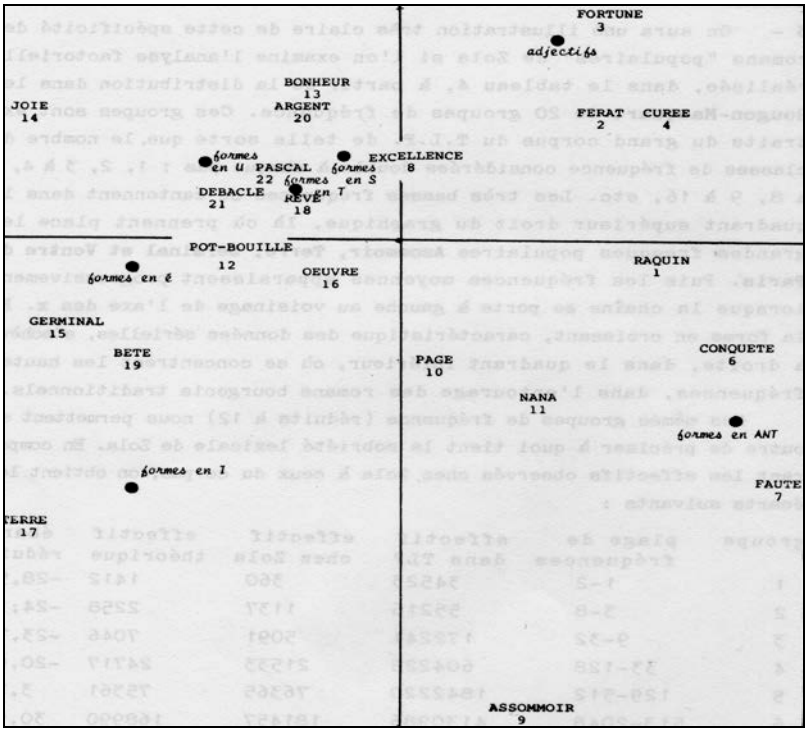
III. La syntaxe

1. Les choix de Zola sont parfois ceux des poètes en matière de syntaxe, ou tout au moins dans le dosage des catégories grammaticales. En particulier, Zola partage avec les poètes le goût de l'adjectif et du participe. Etablissons d'abord les relevés :

		Zola	Epoque	écart réduit
Substantifs	N	556581	3142616	-15,40
	V	8684	18110	
Adjectifs	N	304991	1666440	8,63
+ participes	V	6889	13260	
Verbes	N	433094	2226682	54,53
	V	2914	4676	
Adverbes	N	18190	92988	12,03
en *ment*	V	527	1141	
Mots	N	1562006	8797590	-22,36
grammaticaux	V	328	523	

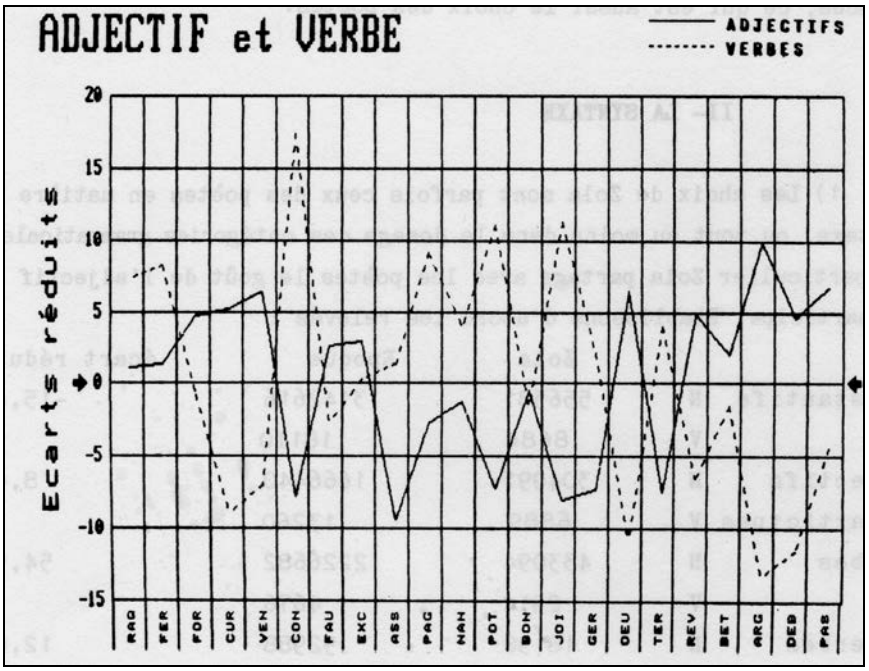
Pour mieux cerner les traits de l'écriture de Zola, nous avons choisi pour référence non pas le corpus entier mais les tranches chronologiques qui coïncident avec la rédaction des *Rougon-Macquart* et où précisément l'adjectif jouit d'une grande faveur. Or on constate une surenchère caractéristique de la part de Zola (écart réduit de +8,63), en un temps où

la concurrence des Goncourt et de l'écriture artiste est pourtant forte. Tout un mouvement parcourt le XIX^e^ siècle qui grossit la faveur de l'adjectif, l'héritage passant des romantiques aux réalistes et aux naturalistes. Il trouve son apogée chez Zola avant de connaître une chute brutale à l'orée du XX^e^ siècle. Mais il y a sans doute quelque abus à parler d'une catégorie composite où se mêlent les participes et les adjectifs. Aussi avons-nous tenté une décantation des uns et des autres en opérant d'abord une sélection sur la désinence et en complétant à la main les opérations de tri. On découvre alors que Zola est moins friand d'adjectif que de participes et particulièrement de participes en *ant* : voici la série des écarts réduits, la prose littéraire du temps servant de pôle de référence : adjectifs -10,57, participes en *é* 6,93, participes en *i* 9,94, participes en *s* 9,79, participes en *ant* 43,34. La prédilection pour la forme en *ant* décline cependant au cours de la rédaction des *Rougon-Macquart*, au bénéfice du participe passé. L'analyse factorielle du graphique 4 décrit clairement cette évolution, les premiers romans se situant à droite avec les adjectifs et les participes présents, et les derniers à gauche avec les formes en *é*, en *i*, en s et en *t*.



Graphique 4. Analyse factorielle des adjectifs et des participes chez Zola

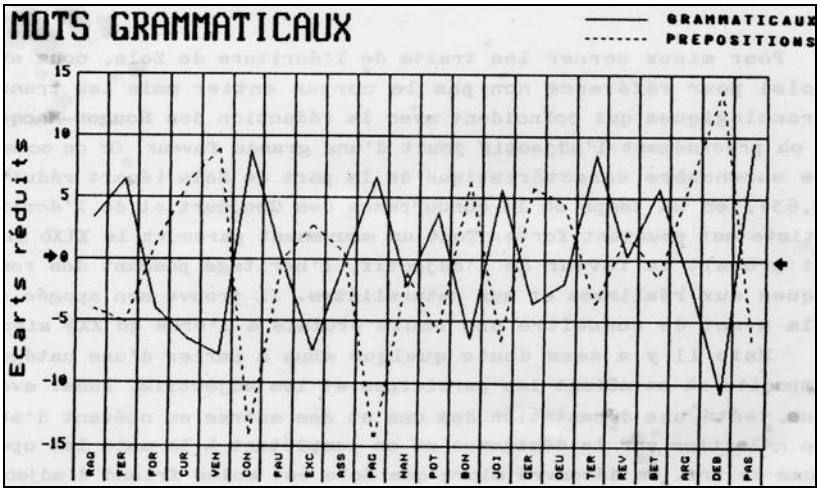
L'afflux des participes est à mettre en relation avec l'abondance des verbes (433.094 occurrences relevées pour 401.800 attendues selon la « norme » de l'époque ou 398.311 selon celle du corpus entier). Car beaucoup des participes sont employés avec l'auxiliaire et constituent des formes verbales au même titre que les formes simples. Mais les deux catégories ne se confondent pas et le graphique 5 les montre complémentaires puisqu'un excédent de l'une accompagne un déficit de l'autre.



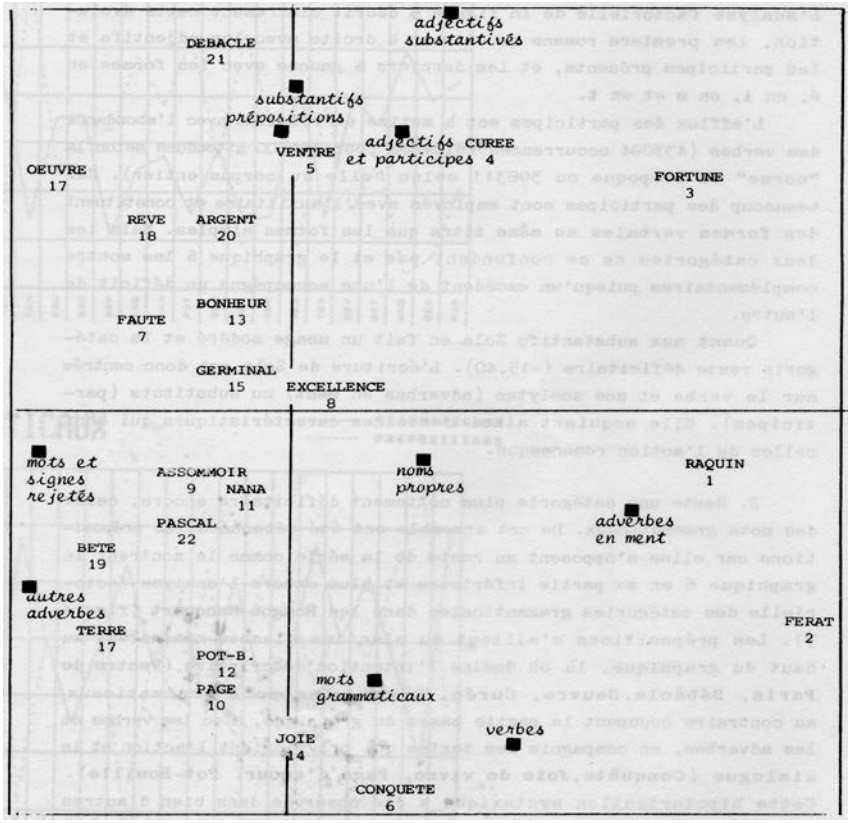
Graphique 5. Adjectifs et verbes

Quant aux substantifs, Zola en fait un usage modéré et la catégorie reste déficitaire (-15,40). L'écriture de Zola est donc centrée sur le verbe et ses acolytes (adverbes en *ment*) ou substituts (participes). Elle acquiert ainsi certaines caractéristiques qui sont celles de l'action romanesque.

2. Reste une catégorie plus nettement déficitaire encore, celle des mots grammaticaux. De cet ensemble ont été détachées les prépositions car elles s'opposent au reste de la série comme le montrent le graphique 6 et plus encore l'analyse factorielle des catégories grammaticales dans les *Rougon-Macquart* (graphique 7).



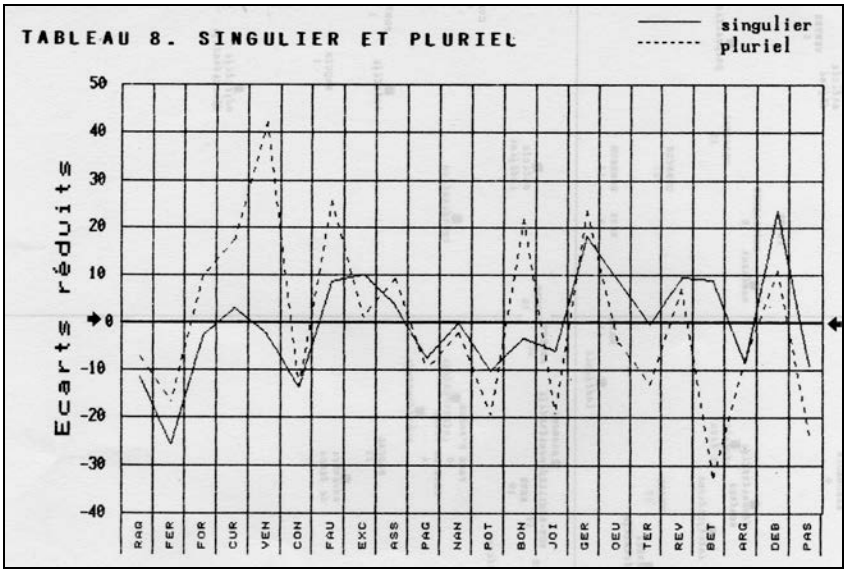
Graphique 6. Mots grammaticaux



Graphique 7. Analyse factorielle des catégories grammaticales (facteurs 1 et 2)

Les prépositions s'allient au clan des classes nominales, au haut du graphique, là où domine l'intention descriptive (*Ventre de Paris, Débâcle, Oeuvre, Curée, Rêve*). Les mots grammaticaux au contraire occupent la partie basse du graphique, avec les verbes et les adverbes, en compagnie des textes qui privilégient l'action et le dialogue (*Conquête, Joie de vivre, Page d'amour, Pot-Bouille*). Cette bipolarisation syntaxique a été observée dans bien d'autres corpus : elle correspond sans doute à quelque caractéristique fondamentale des genres littéraires.

Un exposé aussi rapide ne saurait entrer dans le détail des distributions des mots grammaticaux. Une vingtaine d'espèces ont été distinguées dont chacune donne lieu à des comparaisons externes et internes et autant de courbes ont été établies dont aucune ne peut trouver place ici. Un seul extrait suffira qui mesure la part du singulier et du pluriel dans les *Rougon-Macquart* (graphique 8).



Graphique 8. Singulier et pluriel

On voit que le pluriel s'accorde avec les mouvements de foule comme ceux de *Germinal* ou du *Bonheur des Dames*, ou avec l'entassement luxuriant du *Ventre de Paris* ou de l'épisode du Paradou dans la *Faute de l'abbé Mouret*, mais il y a aussi une évolution d'ensemble qui profite au singulier au détriment du pluriel. Le pluriel l'emporte (proportionnellement) dans les 7 premiers textes alors que le singulier domine dans les 7 derniers. Cette observation est d'autant plus

troublante qu'on la retrouve dans les monographies de Giraudoux et de Proust aussi bien que dans le corpus du XX^{e} siècle. S'agit-il de quelque constante liée à l'âge et au vieillissement ? Ou s'agit-il d'un mouvement général de la langue ? Cette tendance collective est bien réelle quand on l'observe sur deux siècles[6] mais elle n'est pas décelable dans les limites courtes qui délimitent la carrière d'un écrivain. Aussi l'hypothèse individuelle, et peut-être physiologique, ne peut être rejetée. Le tableau d'ensemble des mots grammaticaux (tableau 9) suggère deux indications : d'une part un déficit généralisé de l'ensemble des espèces distinguées, à l'exception des adverbes de lieu et de temps, des interjections et des catégories étroitement liées à la classe nominale : numéraux, articles et prépositions. Zola ne retient guère que les « embrayeurs » et les outils qui actualisent les lieux, les êtres et les choses.

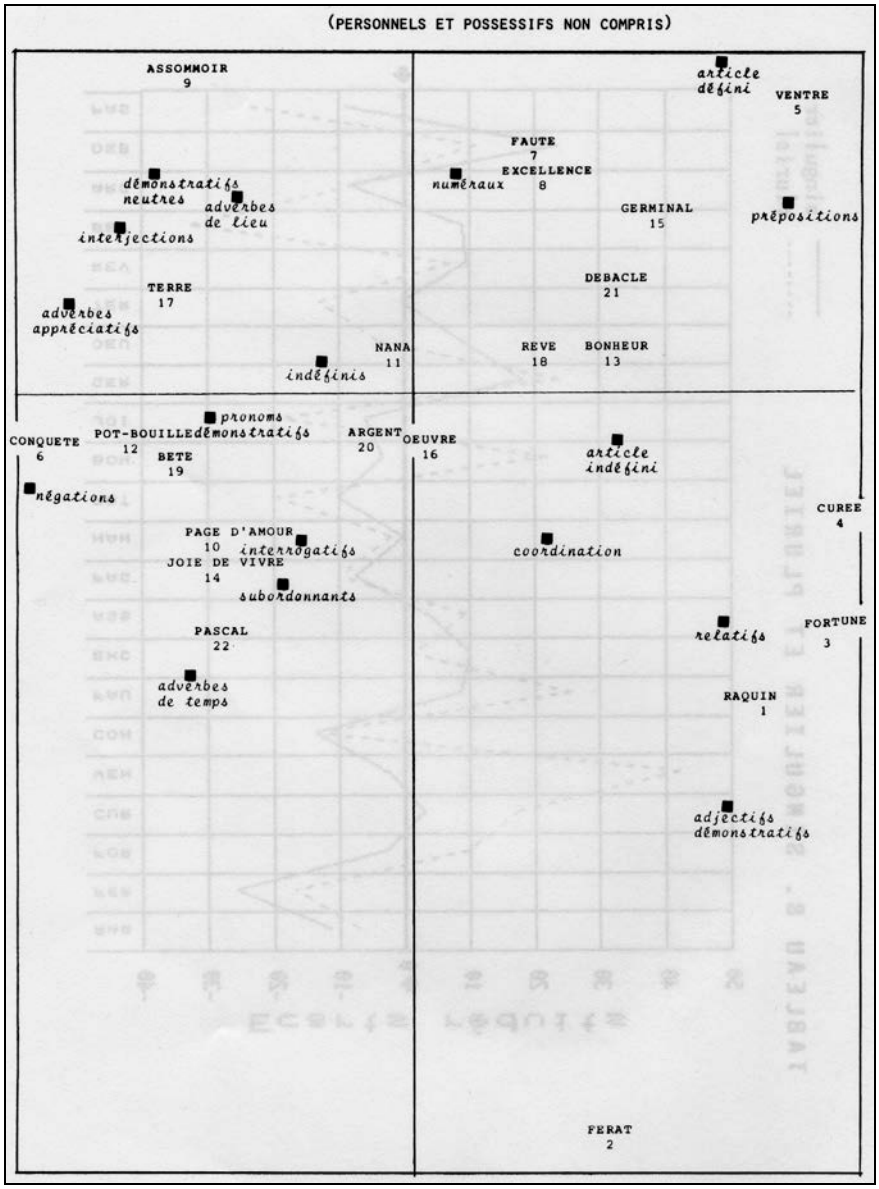
Catégorie interne	Effectif Zola	Effectif époque	Ecarts réduits	Tendance
article défini	244231	1314064	+16,13	+0,23
article indéfini	81329	407671	+38,63	-0,46
démonstratif	52275	342676	-38,03	+0,51
numéraux	20076	93985	+26,44	+0,50
interrogatif	10405	70449	-22,61	+0,84
relatif	29025	180083	-21,27	-0,03
indéfini	43209	265020	-23,30	+0,78
préposition	370491	2034638	+6,10	+0,14
coordination	67289	509980	-90,07	+0,51
subordination	83344	485801	-16,11	+0,46
adverbe d'appréciat.	55795	317937	-7,27	+0,41
adverbe de négation	74599	425369	-8,60	+0,30
adverbe de temps	22331	114761	+12,46	+0,39
adverbe de lieu	23984	128806	+5,37	+0,68
interjection	5865	27379	+14,53	+0,34
lère personne	41561	505401	-181,56	-0,16
2ème personne	27344	197036	-48,10	-0,21
3ème personne	249273	991768	+183,59	-0,61

Tableau 3. La distribution des mots grammaticaux

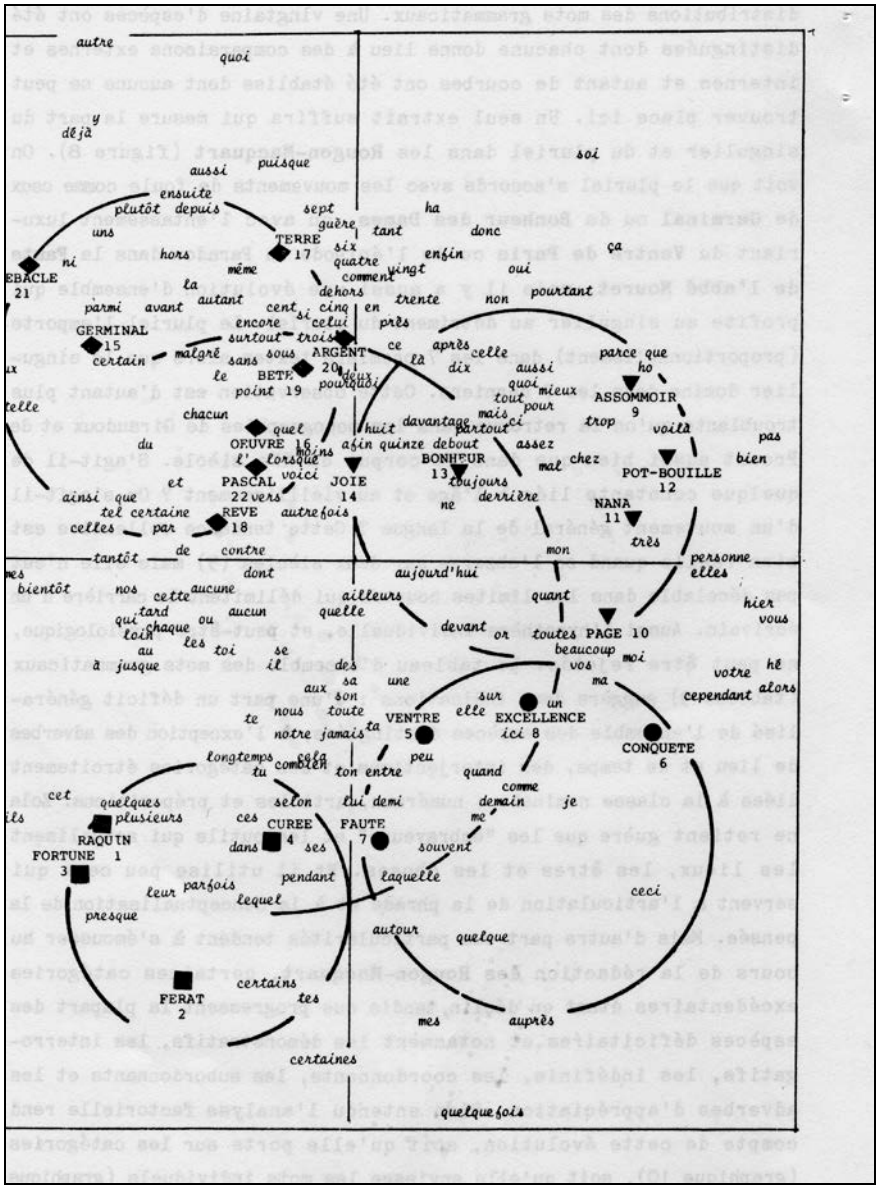
Et il utilise peu ceux qui servent à l'articulation de la phrase et à la conceptualisation de la pensée. Mais d'autre part ces particularités tendent à s'émousser au cours de la rédaction des *Rougon-Macquart*, certaines catégories excédentaires étant en déclin tandis que progressent la plupart des espèces déficitaires et notamment les démonstratifs, les interrogatifs, les indéfinis, les coordonnants, les subordonnants et les adverbes

6. Voir notre *Vocabulaire français de 1789 à nos jours.* Slatkine-Champion, Genève-Paris, 1981, tome 1, p. 393.

d'appréciation. Bien entendu l'analyse factorielle rend compte de cette évolution, soit qu'elle porte sur les catégories (graphique 9), soit qu'elle envisage les mots individuels (graphique 10).



Graphique 9. Analyse factorielle des mots grammaticaux (personnels et possessifs non compris)

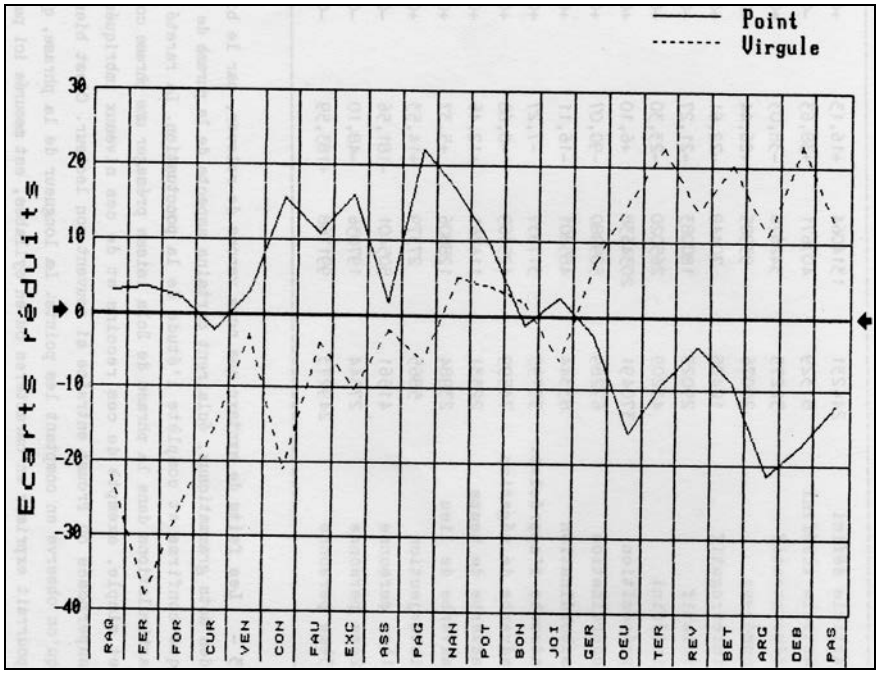


Graphique 10. Analyse factorielle des mots grammaticaux (pris individuellement)

3. Les faits de syntaxe que nous venons de relever, par le biais des mots grammaticaux, éclairent certains aspects de la phrase de Zola que confirme et complète l'étude de la ponctuation. La rareté des articulations dans la phrase de Zola laisse présager une phrase courte et simple, exempte de ces recoins et de ces niveaux imbriqués ou superposés où Proust entraîne si souvent son lecteur. C'est bien ce qu'on observe en comptant les points. La longueur de la phrase, qu'on pourrait exprimer en caractères ou en syllabes, est mesurée ici par le nombre de mots graphiques relevés dans l'intervalle de deux ponctuations fortes. Si l'on accepte de ranger parmi celles-ci le point, le point d'exclamation, le point d'interrogation, et les points de suspension, on calcule un effectif de 147.020 (.) + 24.025 (!) + 10.665 (?) + 14.906 (...) = 196.616 phrases, qu'on rapproche de l'effectif des mots, soit 2.874.755 pour obtenir une moyenne de : 14 mots par phrase. Dans les mêmes conditions les romans de Giraudoux ont une phrase nettement plus longue (de 20,64 mots) et la prose de Proust a une moyenne deux fois plus élevée (30,9), comme celle de Rousseau dans l'*Emile.* D'autres comparaisons faites avec Chateaubriand ou avec l'ensemble du corpus confirment la brièveté de la phrase de Zola.

Mais là encore l'évolution de Zola atténue ce trait spécifique de son écriture. Car la phrase de Zola s'allonge, surtout à partir de *Nana.* La courbe du graphique 11 qui reproduit conjointement la distribution du point et de la virgule dans le corpus zolien (respectivement 147.020 et 340.479 occurrences) montre deux mouvements contrastés : un déclin du point et un progrès de la virgule.

Certes les gains de la virgule viennent en partie des pertes du point et virgule, et le déclin du point est atténué par la montée des points d'exclamation et d'interrogation. Mais ces compensations sont de faible portée et dans l'ensemble, si la segmentation est de plus en plus serrée à cause de l'afflux des virgules, la phrase par contre gagne en ampleur à mesure que Zola progresse dans la rédaction de son oeuvre.



Graphique 11. Le point et la virgule

III. Le contenu

Il nous reste à aborder l'étude du contenu. Ici la prétérition sera plus allusive encore. Car le champ de recherche s'élargit quand on aborde ce domaine. Et la statistique est loin d'être l'arme souveraine lorsque compte le sens des mots et non plus seulement des faits de structure où l'identité des unités n'intervient pas. Le sens linguistique du lecteur est un peu démuni lorsqu'il s'agit d'apprécier, dans un grand corpus, la richesse ou l'originalité lexicale, ou même de mesurer de façon précise des faits de rythme ou de syntaxe que nous venons d'évoquer. Mais la lecture humaine est irremplaçable lorsqu'il s'agit d'évaluer l'importance relative des mots, le relief thématique d'un texte et la signification profonde d'une oeuvre. Ici les éléments quantitatifs ne peuvent que s'attacher à la surface, et laissent ignorer les transparences, les pudeurs, les déplacements, les échos, les antiphrases et toutes les feintes du discours où se cache et se dévoile la pensée. Ce reproche a été adressé depuis longtemps à la statistique linguistique et il restera fondé tant que l'ordinateur n'aura pas acquis le sens de l'humour.

Vocabulaire positif

+39.69	face	26.91	jupe	22.34	quartier
+37.67	abbé	26.87	tête	22.26	regard
+36.25	air	26.83	coron	22.20	fenêtre
+35.37	dame	26.49	geste	22.06	odeur
+35.24	voix	26.23	boutique	21.71	linge
+33.05	frisson	25.88	gorge	21.48	fond
+31.71	coup	24.99	vendeur	21.46	trou
+30.27	coin	24.85	comptoir	21.33	surprise
+30.14	milieu	24.62	veille	20.86	poing
+29.53	chaussée	24.60	rires	20.83	chaise
+29.34	trottoir	24.49	rue	20.70	bande
+28.89	épaule	23.49	bras	20.57	midi
+28.43	préfecture	23.12	docteur	20.39	querelle
+27.10	cou	22.81	bout		
+26.98	yeux	22.81	fichu		

Vocabulaire négatif

-19.02	siècle	-16.93	lundi	-15.74	nature
-18.84	ami	-16.87	dimanche	-15.73	juillet
-18.76	livres	-16.64	mars	-15.70	propos
-18.70	roi	-16.54	juin	-15.47	novembre
-18.62	loi	-16.54	peuple	-15.36	duc
-18.46	espèce	-16.52	anglais	-15.34	impression
-18.31	vie	-16.37	état	-15.30	théâtre
-18.30.	volume	-16.36	septembre	-15.21	journal
-17.76	sens	-16.08	octobre	-15.19	jeudi
-17.42	janvier	-15.99	août	-15.15	talent
-17.35	mer	-15.99	avril	-15.09	samedi
-17.31	vendredi	-15.96	littérature	-15.07	page
-17.24	auteur	-15.91	février		
-17.19	mai	-15.87	mardi		
-17.05	moyen	-15.81	caractère		

Tableau 4. Vocabulaire spécifique de Zola par rapport à son temps (substantifs)

1. Néanmoins certaines questions peuvent être posées à la machine dont les réponses ne sont pas toujours méprisables. On peut se demander quels mots sont caractéristiques d'un écrivain (c'est-à-dire plus fréquents sous sa plume), lesquels cernent le mieux le thème d'un texte, comment évolue le vocabulaire d'un auteur ou quels textes apparaissent comme les

plus spécifiques de sa manière. Tous ces problèmes sont abordés par le truchement de l'écart réduit, c'est-à-dire par le calcul d'un écart entre une fréquence observée et une fréquence attendue (eu égard aux dimensions respectives des deux ensembles comparés et en formulant l'hypothèse que les conditions et les situations du discours sont comparables). Il suffit de classer les mots suivant la valeur pondérée (ou réduite) de cet écart pour obtenir des listes « significatives ». Les deux listes du tableau 4 sont démesurément allongées puisqu'elles enveloppent 3391 mots dépassant le seuil dans le sens positif et 4420 dans le sens négatif. L'extrait très partiel que nous présentons n'en donne pas moins des indications suggestives : les mots familiers à Zola qui apparaissent dans la partie haute sont ceux de ses personnages (*abbé, dame, docteur, vendeur, mari*), du cadre qui les entoure (*chaussée, trottoir, préfecture, coron, boutique, comptoir, rue, quartier, fenêtre, magasin, cuisine, chambre*), de leur corps (*voix, épaule, cou, yeux, tête, gorge, poing, bras, lèvre*), de leurs habits (*jupe, fichu, linge, soie*), de leurs gestes et de leurs comportements (*air, frisson, geste, rire, regard, querelle, malaise, silence*). Le trait commun à toutes ces unités lexicales est leur caractère concret alors que le vocabulaire négatif, qui figure dans la seconde liste du tableau, laisse émerger, au contraire, les éléments qui échappent à la saisie directe des sens, comme l'âme, l'esprit et le temps, pour ne prendre que la tête de liste. Alors que la liste positive est ancrée dans l'espace et dans le monde des choses et des êtres palpables, celle des déficits est relative au monde de l'esprit et du temps. Si l'on inverse les deux listes en changeant les signes, c'est le visage de Proust qui se laisse deviner.

2. Si l'on s'enferme dans le corpus de Zola (il est assez vaste pour qu'on ne s'y sente pas à l'étroit), on peut s'interroger sur le profil spécifique de chaque texte. Cette fois c'est l'ensemble du corpus de Zola qui constitue la référence. Et sur cette toile de fond, le thème de chaque roman se détache avec une évidence telle que les calculs peuvent se passer de la preuve par 9. Des 22 profils obtenus ainsi, nous avons détaché celui de l'*Oeuvre* dont le peintre Claude (qui cache Cézanne) est le héros (tableau 5).

NOMS PROPRES		VERBES		SUBSTANTIFS			
90.07	Claude	22.18	peindre	51.72	tableau	6.16	corbillard
69.76	Sandoz	7.74	créer	42.16	peintre	6.05	figure
56.26	Fagerolles	6.58	exposer	32.60	toile	5.91	canot
54.66	Christine	5.64	travailler	31.58	peinture	5.86	série
46.65	Jory	5.62	gâter	26.77	art	5.85	facture
46.65	Mahoudeau	5.51	déjeuner	26.59	oeuvre	5.80	illusion
40.52	Dubuche	5.42	gratter	23.43	atelier	5.74	article
40.27	Gagnière	5.23	poser	23.38	artiste	5.57	lutteur
36.13	Bongrand	5.01	juger	21.37	jury	5.46	flambée
25.27	Irma	4.78	bâtir	19.82	quai	5.44	lumière
25.16	Chaîne	4.19	exécrer	19.61	sculpteur	5.44	nudité
23.37	Mathilde	4.15	aggraver	18.40	palette	5.43	butte
22.21	Naudet	4.14	indigner	17.97	institut	5.42	port
13.72	Seine	4.06	blaguer	17.21	ébauche	5.36	théorie
7.46	louvre	4.06	rhabiller	16.52	école	5.33	déchéance
6.82	Passy			15.42	étude	5.26	bibelot
6.68	Paris			14.43	modèle	5.26	envoi
5.26	Gaston			14.12	formule	5.26	voyou
4.86	Midi			13.82	chevalet	5.21	tourment
4.40	Alice			13.48	talent	5.18	enfance
				13.20	crétin	5.12	note
				12.61	veston	5.10	motif
				12.57	paravent	5.10	huée
				11.82	pont	5.08	bassesse
				11.77	succès	5.08	exagération
				11.56	pinceau	5.04	mépris
				11.38	séance	5.03	qualité
				11.16	génie	4.99	sein
				10.69	paysage	4.88	flèche
				10.17	divan	4.86	coussin
				9.90	poêle	4.85	cuisse
				9.62	musée	4.85	plâtre
				9.15	châssis	4.76	imbécile
				9.15	médaille	4.76	liste
				9.09	nature	4.76	mélancolie
				9.03	cité	4.76	silhouette
				8.39	bouquin	4.67	reste
				8.26	camarade	4.62	ami
				8.22	arche	4.62	contemplation
				8.17	statue	4.58	maître
				8.16	couleur	4.56	crâne
				8.09	jeudi	4.49	monument
				7.99	salon	4.48	buste
				7.81	élèves	4.48	enfilade
				7.81	berge	4.46	soleil
				7.63	cadre	4.37	réalité
				7.50	critique	4.37	révolutionnair
				7.46	vote	4.22	chair
				7.31	salle	4.20	impuissance
				7.17	échelle	4.20	rivière
				7.07	travail	4.13	abord
				7.05	architecte	4.10	bière
				6.95	avenue	4.10	chocolat
				6.85	brosse	4.10	génération
				6.80	création	4.09	baie
				6.74	dessin	4.06	application
				6.71	production	4.06	métier
				6.70	île	4.06	personnalité
				6.62	gloire	4.06	sculpture
				6.45	passion	4.05	espace
				6.44	ambition	4.04	influence
				6.41	jeunesse		
				6.39	morceau		
				6.21	farceur		
				6.18	littérature		
				6.16	bâtisse		

GRAMMATICAUX		ADJECTIFS+PART.	
12.06	ha	13.88	romantique
8.98	hein	12.67	amateur
7.95	tu	12.60	raté
7.87	ton	9.55	public
7.29	cette	7.51	peint
6.85	une	7.35	débutant
5.94	de	6.95	gaillard
5.57	tous	6.57	nue
5.46	quelle	6.24	exécrable
5.38	ce	6.15	bourgeois
5.21	toi	6.13	refusé
5.06	quel	5.54	infini
4.93	oui	5.43	original
4.42	des	5.36	célèbre
4.21	ho	5.26	incomplet
4.20	jadis	5.21	fuyant
4.09	où	5.20	manqué
4.09	y	5.16	vieux
		5.12	barbouillé
		5.11	enflé
		5.08	immortel
		5.01	vivant
		4.89	moderne
		4.85	fouetté
		4.75	comique
		4.69	lugubre
		4.66	malin
		4.62	ri
		4.59	glorieux
		4.46	sacré
		4.46	vitré
		4.14	triomphal
		4.14	visiteur
		4.10	somnolent
		4.06	gesticulant
		4.05	gueuse
		4.05	ravagé
		4.01	béant

PONCTUATION	
16.16	!
15.31	,
7.00	?
4.78	:

Tableau 5. Vocabulaire spécifique de Zola par rapport à son temps (substantifs)

Le nom de *Claude* est bien sûr le premier des noms propres, comme *peindre* est le premier des verbes et *tableau* le premier des substantifs. Il est difficile d'ignorer le sujet quand on a parcouru le début de ces listes

(*peindre, créer, exposer, travailler, gâter* parmi les verbes, *tableau, peintre, toile, peinture, art, oeuvre, atelier, artiste* parmi les substantifs, *romantique, amateur, raté, public, peint* parmi des adjectifs-participes). Mais on peut glaner aussi des indications sur le ton particulier et un peu autobiographique de ce roman, si l'on consulte la suite de la liste, où des faits inattendus se glissent parmi les évidences, ou si l'on s'attache à la ponctuation (excédent des points d'exclamation et d'interrogation) et aux mots grammaticaux : la fréquence des interjections (*ha, hein, ho*), le tutoiement (*tu, ton, toi*), la vivacité du dialogue et du sentiment (*quel, quelle*), l'immédiateté de la description (*cette, ce*) et la nostalgie du *jadis* sont autant de renseignements intéressants sur l'atmosphère du roman.

3. La succession des 22 profils peut donner une image en mouvement de l'évolution de Zola. Mais le relief de chaque image est trop accusé pour qu'on puisse espérer un fondu-enchaîné aisément déchiffrable. On aura donc recours à une autre méthode pour figer ce mouvement : le coefficient de corrélation chronologique, qui est calculé pour chaque mot et évolue entre deux limites - 1 (régression) et + 1 (progression). Notre index fournit cette information pour tous les mots dont la fréquence est suffisante pour donner un sens au calcul. Là encore il suffit d'appliquer le tri à l'ensemble des coefficients pour obtenir deux lots de mots, reproduits dans le tableau 6. Le premier (partie supérieure) regroupe les vocables qui gagnent en fréquence tout au long des *Rougon-Macquart* et le second (partie inférieure) recueille les laissés pour compte, ceux que la faveur de Zola a désertés. Comment n'être pas sensible à l'assombrissement de Zola quand on le voit utiliser de plus en plus les mots qui peignent la destruction (*catastrophe, désastre, déluge, tempête, débâcle, massacre, destruction, effondrement, anéantissement, déchéance, épave*), la détresse (*chagrin, plainte, peine, détresse, incertitude, fuite, menace, aguets*) mais aussi la résistance (*bravoure, enragement, obstination, effort, poursuite, travail*). Remarquons que la plupart des substantifs de cette liste sont très éloignés du concret, qu'ils représentent une situation, un rapport au monde, un sentiment ou une réaction, et qu'ils marquent un pas vers l'intériorisation et l'abstraction. Cela est plus net encore quand on parcourt la liste des abandons qui apparaît comme le paradis perdu de la femme, du corps, des biens et des jouissances. D'un monde coloré où l'on touche les choses, Zola semble avoir glissé à un univers plus sombre où l'on ne touche plus que l'ombre des choses, les idées.

NOMS PROPRES		VERBES		SUBSTANTIFS			
90.07	Claude	22.18	peindre	51.72	tableau	6.16	corbillard
69.76	Sandoz	7.74	créer	42.16	peintre	6.05	figure
56.26	Fagerolles	6.58	exposer	32.60	toile	5.91	canot
54.66	Christine	5.64	travailler	31.58	peinture	5.86	série
46.65	Jory	5.62	gâter	26.77	art	5.85	facture
46.65	Mahoudeau	5.51	déjeuner	26.59	oeuvre	5.80	illusion
40.52	Dubuche	5.42	gratter	23.43	atelier	5.74	article
40.27	Gagnière	5.23	poser	23.38	artiste	5.57	lutteur
36.13	Bongrand	5.01	juger	21.37	jury	5.46	flambée
25.27	Irma	4.78	bâtir	19.82	quai	5.44	lumière
25.16	Chaîne	4.19	exécrer	19.61	sculpteur	5.44	nudité
23.37	Mathilde	4.15	aggraver	18.40	palette	5.43	butte
22.21	Naudet	4.14	indigner	17.97	institut	5.42	port
13.72	Seine	4.06	blaguer	17.21	ébauche	5.36	théorie
7.46	louvre	4.06	rhabiller	16.52	école	5.33	déchéance
6.82	Passy			15.42	étude	5.26	bibelot
6.68	Paris			14.43	modèle	5.26	envoi
5.26	Gaston			14.12	formule	5.26	voyou
4.86	Midi			13.82	chevalet	5.21	tourment
4.40	Alice			13.48	talent	5.18	enfance
				13.20	crétin	5.12	note
				12.61	veston	5.10	motif
				12.57	paravent	5.10	huée
				11.82	pont	5.08	bassesse
				11.77	succès	5.08	exagération
				11.56	pinceau	5.04	mépris
				11.38	séance	5.03	qualité
				11.16	génie	4.99	sein
				10.69	paysage	4.88	flèche
				10.17	divan	4.86	coussin
				9.90	poêle	4.85	cuisse
				9.62	musée	4.85	plâtre
				9.15	châssis	4.76	imbécile
				9.15	médaille	4.76	liste
				9.09	nature	4.76	mélancolie
				9.03	cité	4.76	silhouette
				8.39	bouquin	4.67	reste
				8.26	camarade	4.62	ami
				8.22	arche	4.62	contemplation
				8.17	statue	4.58	maître
				8.16	couleur	4.56	crâne
				8.09	jeudi	4.49	monument
				7.99	salon	4.48	buste
				7.81	élèves	4.48	enfilade
				7.81	berge	4.46	soleil
				7.63	cadre	4.37	réalité
				7.50	critique	4.37	révolutionnair
				7.46	vote	4.22	chair
				7.31	salle	4.20	impuissance
				7.17	échelle	4.20	rivière
				7.07	travail	4.13	abord
				7.05	architecte	4.10	bière
				6.95	avenue	4.10	chocolat
				6.85	brosse	4.10	génération
				6.80	création	4.09	baie
				6.74	dessin	4.06	application
				6.71	production	4.06	métier
				6.70	île	4.06	personnalité
				6.62	gloire	4.06	sculpture
				6.45	passion	4.05	espace
				6.44	ambition	4.04	influence
				6.41	jeunesse		
				6.39	morceau		
				6.21	farceur		
				6.18	littérature		
				6.16	bâtisse		

GRAMMATICAUX		ADJECTIFS+PART.	
12.06	ha	13.88	romantique
8.98	hein	12.67	amateur
7.95	tu	12.60	raté
7.87	ton	9.55	public
7.29	cette	7.51	peint
6.85	une	7.35	débutant
5.94	de	6.95	gaillard
5.57	tous	6.57	nue
5.46	quelle	6.24	exécrable
5.38	ce	6.15	bourgeois
5.21	toi	6.13	refusé
5.06	quel	5.54	infini
4.93	oui	5.43	original
4.42	des	5.36	célèbre
4.21	ho	5.26	incomplet
4.20	jadis	5.21	fuyant
4.09	où	5.20	manqué
4.09	y	5.16	vieux
		5.12	barbouillé
		5.11	enflé
		5.08	immortel
		5.01	vivant
		4.89	moderne
		4.85	fouetté
		4.75	comique
		4.69	lugubre
		4.66	malin
		4.62	ri
		4.59	glorieux
		4.46	sacré
		4.46	vitré
		4.14	triomphal
		4.14	visiteur
		4.10	somnolent
		4.06	gesticulant
		4.05	gueuse
		4.05	ravagé
		4.01	béant

PONCTUATION	
16.16	!
15.31	,
7.00	?
4.78	:

Tableau 4. L'évolution des substantifs chez Zola

Il est paradoxal qu'on rejoigne ainsi Proust, que tout oppose à Zola et qui lui aussi se détache d'un univers plus poétique et plus lumineux, pour s'enfermer dans les profondeurs glauques du coeur et de la pensée

humaine. C'est aussi ce que nous avons observé chez Giraudoux qui ne partage guère avec Zola qu'un point commun, la haine de l'abstraction, et qui perd pied pourtant lui aussi, comme Zola, devant la marée montante de l'abstraction. Très éloignés au départ, ces trois écrivains maintiennent leur distance à l'arrivée. Mais quel parallélisme dans leur cheminement, qui va de la nature à l'homme, de la poésie à la morale et du concret à l'abstrait. Y aurait-il donc, au-dessus des tempéraments individuels, une sorte de pente commune, une gravitation universelle à laquelle les écrivains et les artistes seraient soumis ? C'est une hypothèse que le cas de Zola, après quelques autres, invite à envisager, et que Poirot-Delpech appelle un « secret de fabrique ». Qu'il s'agisse d'une simple hypothèse ou d'une clé authentique, il convient d'explorer les effets de l'âge, la force de la loi naturelle qui sécrète les fruits de l'âge mûr, les abstractions, comme des « excroissances », ou des « tavelures sur le dos des mains »[7].

7. Poirot-Delpech, « Secrets de fabrique », *Le Monde*, 30 décembre 1983.