La phrase de Zola

Étienne Brunet

***Chapitre 9 (numérique) de : Étienne BRUNET,* Tous comptes faits.**
Écrits choisis, tome III. Questions linguistiques*, Bénédicte PINCEMIN (éd.),*
Paris : Éditions Champion, sous presse (publication prévue en 2016).
Publié en ligne par la revue Texto ! Textes & Cultures, http://www.revue-texto.net
Volume XXI – n°2 (2016). Coordonné par Audrey MOUTAT.

La phrase de Zola[1]

Le corpus de Zola, emprunté au *Trésor de la langue française*, offre de grands avantages. Il s'appuie sur l'excellente édition de la Pléiade réalisée par Henri Mitterand[2]. Il offre au chercheur le cas assez rare d'un ensemble fermé, d'un corpus constitué délibérément et préalablement par l'auteur lui-même. De la même façon que les *Mémoires d'outre-tombe* ou la *Recherche du temps perdu*, les *Rougon-Macquart* se présentent comme un lot d'un seul tenant, comme un héritage indivis que l'auteur cède au chercheur en lui évitant l'embarras et l'arbitraire du découpage. Corpus naturel, les *Rougon-Macquart* ajoutent à la cohérence de l'ensemble, l'abondance, la variété, l'équilibre et l'autonomie des parties, ce qui facilite l'étude comparative. Et surtout Zola nous invite à un long voyage linéaire de plus de vingt ans, posant honnêtement chaque année son jalon, sans ces entorses à la chronologie qu'on relève si souvent dans la rédaction de Proust, Chateaubriand ou Hugo. Enfin, les *Rougon-Macquart* par leur étendue (près de trois millions d'occurrences) offrent le champ libre à la plus vaste monographie qui puisse être tentée à partir des données de l'Institut national de la langue française. Or les méthodes statistiques, ayant plus de puissance que de finesse, se complaisent dans les grands espaces.

La phrase de Zola ne semble pas receler quelque mystère particulier qui puisse justifier, autant que celle de Proust ou de Chateaubriand, de Rousseau ou de Giraudoux, l'examen approfondi que nous avons entrepris. Aussi bien notre recherche présente porte sur l'ensemble du vocabulaire des *Rougon-Macquart* et les conclusions les plus

1. NDÉ : Article paru dans B. Derval et M. Lenoble (éd.), *La critique littéraire et l'ordinateur*, Montréal, 1985, p. 111-157 (1985d).

2. Ou du moins sur les trois premiers tomes de cette édition car les deux derniers volumes n'étaient pas encore parus au moment de l'enregistrement des données à Nancy, il y a vingt ans. Il a donc fallu procéder à une transposition de 3000 pages afin que toutes les références renvoient au texte de la Pléiade.

intéressantes sont plutôt d'ordre thématique que stylistique, le contenu l'emportant apparemment sur la structure, et le lexique sur la syntaxe.

Tableau 1. Index synoptique des *Rougon-Macquart* (extrait)

L'objet premier de notre recherche est en effet de constituer un **Index** complet et synoptique des *Rougon-Macquart*[3], à l'image de l'extrait présenté dans le tableau 1.

Il faut avouer aussi que la phrase est faite de rapports entre les mots et que ces rapports échappent à la statistique lexicale qui n'étudie que les mots individuels, détachés de leur contexte immédiat. Peut-être des méthodes plus sophistiquées auraient pu être appliquées ici qui saisissent les liens syntaxiques, voire sémantiques, dont le discours est tissé. Des programmes existent qui abordent les combinaisons de mots, et s'engagent dans l'analyse de contenu. Mais si prometteuses que soient ces expériences, leur rentabilité est encore trop faible pour le traitement d'un si grand corpus. Ajoutons que les méthodes statistiques perdent de leur efficacité quand le nombre des combinaisons s'accroît à l'infini et que les effectifs correspondants s'amenuisent du même pas. Dans ce champ nouveau de recherche, les repères manquent enfin, et les comparaisons deviennent hasardeuses. Nous nous en tenons donc aux méthodes classiques de la lexicométrie qui peuvent donner quelque éclaircissement – parfois indirectement – au problème qui nous occupe.

– I –

1 – Ces méthodes donnent un accès direct à l'un des caractères de la phrase qui est simple à isoler et que nous explorerons d'abord : la longueur. La longueur, 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 :

132 114 (.) + 24 025 (!) + 10 665 (?) + 14 906 (...) = 181 710 phrases

qu'on rapproche de l'effectif des mots, soit 2 874 755, pour obtenir une moyenne de 15,82 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'*Émile* (27,71). Nous disposons d'autres points de repère – un peu moins sûrs toutefois, car les points de

3. 20 textes des *Rougon-Macquart* notre étude – mais non l'index – ajoute deux romans antérieurs, *Thérèse Raquin* et *Madeleine Férat*.

suspension, de siglaison ou d'abréviation n'y ont pas été distingués du point[4] :

Chateaubriand :	22,23
Corpus XIX-XX :	14,60
Prose littéraire de 1860 à 1907 :	12,58

En négligeant pareillement les ambiguïtés qui s'attachent au point, la phrase de Zola aurait en moyenne 13,59 mots, et sa longueur, inférieure à celle de Chateaubriand, rejoindrait la norme du corpus entier et celle de la prose littéraire de son temps. En réalité cette norme incluant le théâtre se prête mal à la comparaison. Si l'on pouvait y isoler le genre romanesque, nul doute que la phrase y serait plus longue que celle de Zola.

2 – Mais la segmentation du discours ne se limite pas aux ponctuations fortes. Si l'on envisage en outre le point et virgule, les deux points et la virgule, l'espace moyen entre deux délimiteurs[5] se réduit à 4,96 mots chez Zola, alors qu'il s'étend à 8,12 chez Proust, 9,53 chez Rousseau, 7,05 chez Chateaubriand, 6,05 dans l'ensemble du corpus XIX-XX et 5,35 dans la prose littéraire du temps de Zola. La segmentation est donc plus serrée chez Zola que chez les autres écrivains.

3 – Ainsi mesurée la segmentation est un effet global dont les composantes peuvent varier largement d'un auteur à l'autre. En faisant abstraction de l'étendue des textes, on peut mesurer la proportion (le pourcentage) de chaque signe dans l'ensemble des ponctuations relevées dans un texte, un auteur ou un corpus. La figure 2 donne la représentation graphique de cette répartition. On y constate la parenté de Rousseau et de Chateaubriand et le goût commun – partagé par leur époque – pour les signes moyens : point et virgule et deux points. On voit aussi que chez Proust et Giraudoux romancier la ponctuation tend à se simplifier et à se réduire à deux signes : le point et la virgule, Giraudoux préférant le point et Proust la virgule. Quant à Zola, le dosage (point 30,5 %, exclamation 4,1 %, interrogation 1,8 %, virgule 58,7 %, point et virgule 3,2 %, deux points 1,7 %) semble s'écarter moins violemment de celui qu'on observe dans l'ensemble du corpus.

---

4. En ce qui concerne le corpus XIX-XX on a tenté d'isoler le point et d'éliminer les impuretés, ce qui porterait la longueur moyenne de la phrase à 15,24 mots (et à 15,82 si l'on tient compte des noms propres).

5. Ou de deux ponctèmes, si l'on adopte la terminologie de Nina Catach. Sur cette question, voir le numéro 45 de *Langue Française* consacré à la ponctuation et notamment la mise au point de N. Catach, p. 16-27 (Larousse, février 1980).

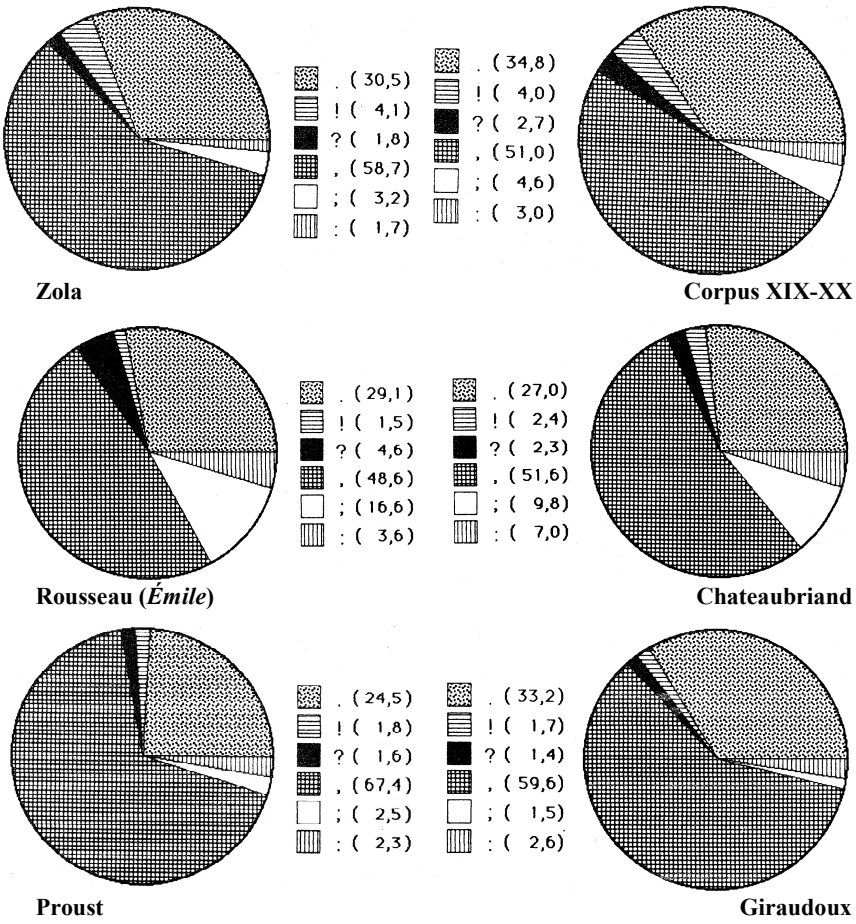

Figure 2. Répartition comparée des signes de ponctuation chez Zola et quelques écrivains (pourcentages)

Mais si l'on applique la mesure probabiliste en prenant en compte l'étendue du texte, les écarts sont largement significatifs, surtout en ce qui concerne l'excédent de la virgule.

	effectif observé chez Zola	effectif théorique	écart réduit
point	176 861	165 176	+29,36
exclamation	24 025	18 090	+36,47
interrogation	10 665	12 637	-17,91
virgule	340 479	242 170	+203,99
point et virg.	18 391	22 063	-25,24
deux points	9 673	14 131	-38,29

Zola participe donc à la simplification observée au XX^e^ siècle. Lui aussi s'éloigne des signes minoritaires (: ; ?) à l'exception du point d'exclamation dont il fait grand usage.

4 – Resterait à observer certains signes spécialisés pour certaines fonctions : guillemets, tiret de dialogue, italiques, trait d'union, parenthèses. Mais on se heurte ici à trop d'inconstance dans les usages et à l'arbitraire des éditeurs. Ainsi les guillemets qu'on trouve dans l'édition de la Pléiade au début de chaque passage dialogué ont été supprimés dans le texte enregistré, n'étant conservés que dans les citations. Bornons-nous ici à observer que Zola ignore la parenthèse (on n'en a relevé que trois dans le corpus) comme certains tirets qui jouent le même rôle, et qu'il n'y a pas chez lui ce second plan du discours auquel tirets et parenthèses donnent accès et où Proust engage si souvent son lecteur.

– II –

1 – Mais les conclusions sont plus sûres lorsqu'on s'enferme dans le même corpus, dans l'enceinte d'une édition complète et homogène, et qu'on procède à l'étude des variations internes. La constance des choix étant assurée, on peut suivre l'évolution de Zola, de *Thérèse Raquin* au *Docteur Pascal*. L'étude repose sur les données de base reproduites dans le tableau 3, qui compte plus d'un demi-million d'observations. Comme nous disposions de la localisation précise de chacun de ces signes, il nous a été possible de désambiguïser le point et d'isoler les points de suspension, dont le nombre n'est pas négligeable puisqu'ils représentent 10 % des points. Le tableau 3 n'est pas directement exploitable puisque les effectifs dépendent de la taille respective des différents textes. On peut toutefois établir un rapport entre les ponctuations fortes (point, suspension, exclamation et interrogation) et les ponctuations faibles (virgule, point et virgule, deux points). Or ce rapport, qui dans l'ensemble est de 0,5, est nettement plus fort au début (0,6) qu'à la fin (0,4) et l'on peut suivre cette progression dans la dernière colonne du tableau. On discerne cette évolution plus nettement encore au graphique E de la figure 4, où les deux séries sont représentées conjointement dans une courbe chronologique. Autour de la ligne médiane (marquée par le nombre 0) les points s'ordonnent selon la valeur, positive ou négative, de l'écart réduit, par quoi l'on mesure l'excédent ou le déficit de l'objet considéré dans chacun des textes. C'est dire que, ici comme dans nos précédentes études, la statistique s'appuie sur le schéma d'urne et la loi normale. Nous avons tenté ailleurs de justifier dans le domaine

linguistique l'usage de la statistique classique[6] et nous y reviendrons brièvement dans la suite de cet exposé. Il suffit ici de suivre des yeux le mouvement des courbes.

Tableau 3. Les signes de ponctuation chez Zola. Effectifs absolus

	. point	... susp	! excl	? inte	, virg.	; pév	: deux	— tir.	— trait	« guil	ital	a fort.	b faib.	a/b rapport
RAQ	3468	236	138	76	6195	691	168	285	543	56	83	3918	7054	0.56
FER	5131	362	236	245	8012	922	291	518	1115	174	83	5974	9225	0.65
FOR	6157	238	534	229	11346	936	362	711	1301	238	142	7158	12644	0.57
CUR	4958	443	357	232	11164	738	364	679	1200	234	73	5990	12266	0.49
VEN	5538	534	426	232	13234	1015	318	734	1201	250	42	6730	14567	0.46
CON	6292	943	756	512	11585	1285	402	1432	1916	162	74	8503	13272	0.64
FAU	6364	679	859	462	13948	634	377	1067	1564	72	165	8364	14959	0.56
ROU	7258	728	1002	496	14295	889	543	1469	1907	247	116	9484	15727	0.60
ASS	7820	769	1929	443	19244	1399	436	995	2272	76	311	10961	21079	0.52
PAG	6206	843	918	492	11953	647	347	1235	1409	107	80	8459	12947	0.65
NAN	7821	1007	1796	611	17881	1087	509	1523	1797	185	80	11235	19477	0.58
POT	6875	1045	1873	651	17044	912	675	1669	1935	78	162	10444	18631	0.56
BON	6916	808	1313	651	18402	972	642	1531	2114	88	513	9688	20016	0.48
JOI	5803	710	972	623	13917	602	420	1142	1590	40	19	8108	14939	0.54
GER	7944	606	1463	729	21388	908	621	1362	2083	46	117	10742	22917	0.47
OEU	5000	774	1692	668	18292	811	566	1107	1489	88	134	8134	19669	0.41
TER	6979	957	2307	796	23417	947	657	1681	2113	70	95	11039	25021	0.44
REV	3052	292	385	237	9551	345	230	416	650	78	101	3966	10126	0.39
BET	5484	601	929	557	18264	673	467	957	1721	46	24	7571	19404	0.39
ARG	5022	850	1235	569	19267	684	399	1049	1677	48	101	7676	20350	0.38
DEB	7440	860	1833	690	26482	832	564	1398	2197	68	64	10823	27878	0.39
PAS	4586	621	1072	464	15598	462	315	955	1218	50	44	6743	16375	0.41
tot	132114	24025			340479	9673		35012	2623			368543		0.49
		14906	10665			18391		23915	2501			181710		
corr.	-.72	+.20	+.55	+.61	+.90	-.82	+.11	+.26	-.08	-.72	-.19	-.20	+.90	

6. NDÉ : voir tome II, chapitre 5, « Le viol de l'urne » (1984a), p. 79-93.

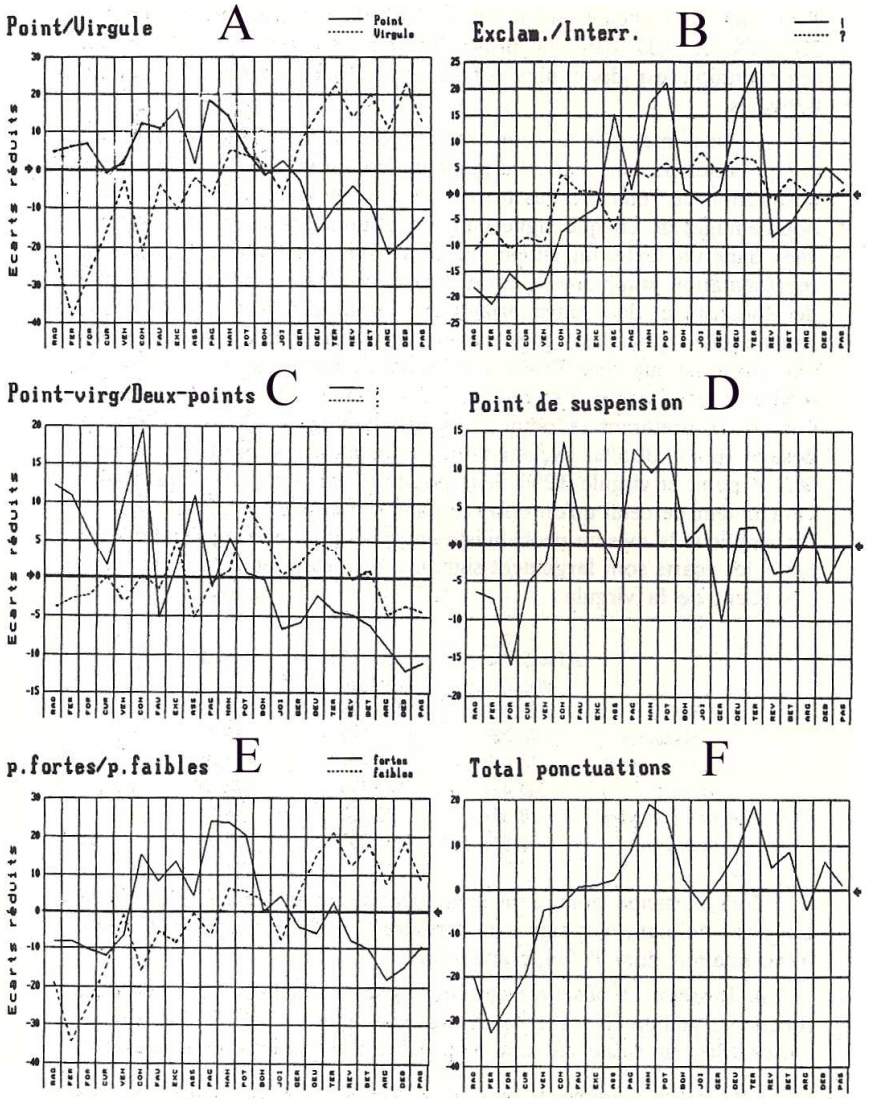

Figure 4. Courbes comparées des signes de ponctuation

On sera sensible tout d'abord au progrès d'ensemble des signes de ponctuation (Total ponctuations dans le graphique 4-F), ce qui implique que la segmentation de la phrase de Zola est de plus en plus courte. Nous avions reconnu là, dans cette segmentation serrée, un trait spécifique de Zola, comparé aux prosateurs qui l'ont précédé ou suivi. Zola durcit

donc, en vieillissant, les caractéristiques de son écriture[7]. Il ne faut point en conclure prématurément que la phrase de Zola se raccourcit au fil des années. Cela n'est vrai que dans la première moitié du corpus jusqu'à *Nana*. Mais à partir de *Nana* la phrase de nouveau s'allonge régulièrement. Ce mouvement contrarié, qui s'inscrit dans une courbe parabolique, échappe au coefficient de corrélation chronologique, qui n'est pas significatif (r = -0,20). Par contre le progrès des ponctuations faibles se marque dans la figure 4-E par une diagonale en pente raide (ligne en pointillé en haut) – ce que souligne un très fort coefficient de corrélation chronologique (+0,90).

Encore peut-on aller plus loin dans l'analyse. La même figure 4 décompose le système des ponctuations en ses éléments constituants. Les deux signes principaux (le point et la virgule) sont réunis sur la partie supérieure en A, le déclin du point (r = -0,72) accusant la progression de la virgule (r = +0,92). Mais le point a des substituts, point d'exclamation et point d'interrogation (graphique B), qui marquent un mouvement ascendant (respectivement +0,55 et +0,61). De la même façon, dans le clan des ponctuations faibles, le progrès de la virgule est atténué par le déclin du point et virgule (-0,82). Un seul parmi les six principaux signes n'atteint pas le seuil : les deux points (r = 0,20). Partout ailleurs la probabilité est inférieure à 0,01. C'est dire que la ponctuation de Zola est traversée de mouvements profonds, dont la complexité se simplifie quand on fait appel à l'analyse factorielle.

2 – L'analyse factorielle (figure 5) prend appui sur les sept premières colonnes du tableau 3, ou plus précisément sur leur transformation en écarts réduits[8]. Très visiblement le premier facteur[9] est lié à la chronologie : les 11 premiers textes occupent la moitié basse du graphique et les 11 derniers la partie supérieure. La suite linéaire du temps est à peu près respectée depuis *Thérèse Raquin* (quadrant inférieur gauche) jusqu'aux derniers textes des *Rougon-Macquart* (quadrant supérieur gauche), en passant par les textes intermédiaires (quadrant

7. Nous avons constaté chez Proust également une certaine tendance à accuser ses traits propres et à ne point céder aux critiques

8. Comme les textes sont de longueur inégale, une pondération est nécessaire qui donne à chaque ligne (à chaque texte) le même poids. L'écart réduit offre en outre l'avantage de pondérer les colonnes, c'est-à-dire les signes de ponctuation, afin d'éviter la domination absolue des deux plus fréquents, le point et la virgule.

9. Le premier facteur explique 58 % de la variance, le second 17 % et le troisième 13 %.

inférieur, puis supérieur de la moitié droite). On reconnaît la forme en croissant qui est caractéristique des données sérielles et qui parcourt dans un mouvement enveloppant les quatre coins du graphique.

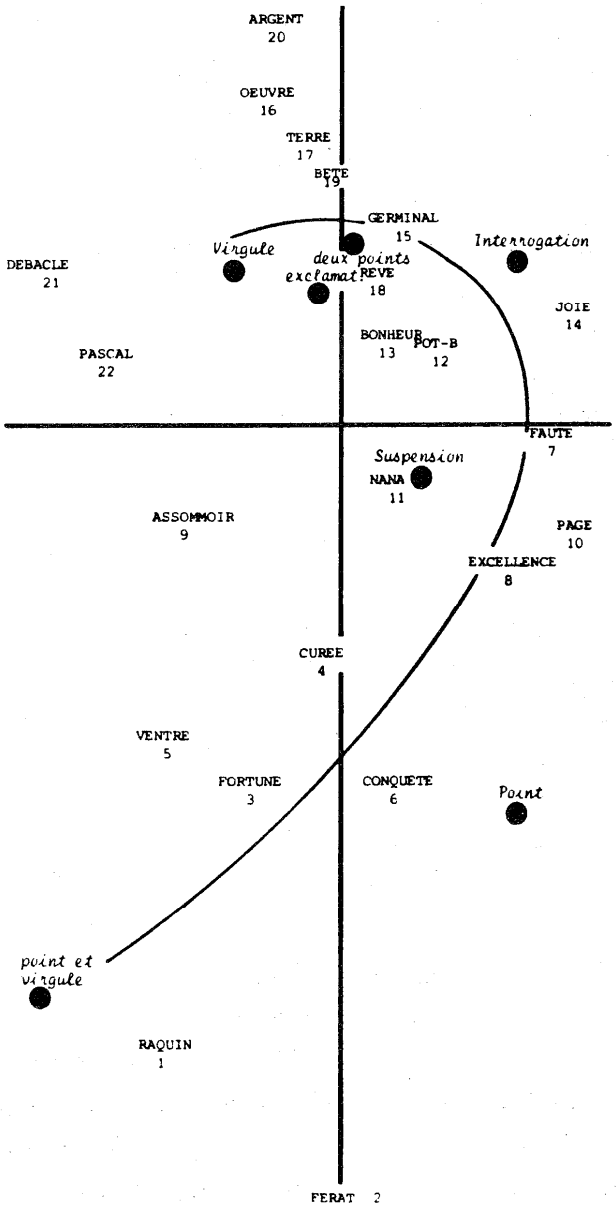

Figure 5. Analyse factorielle des signes de ponctuation

On ne s'interrogera pas longtemps sur le deuxième facteur qui lui aussi reflète la chronologie et oppose les termes extrêmes aux termes moyens : les textes de la période centrale (de la *Conquête* à *Germinal*) se portent sur la partie droite, où se trouvent les ponctuations fortes, les autres textes occupant la zone gauche, du côté des ponctuations faibles. Ce graphique 5 rend claires les conclusions auxquelles nous étions parvenu : d'une part un double mouvement de raccourcissement puis d'allongement de la phrase, d'autre part l'abandon du point et virgule au profit de la virgule et celui du point au profit des ponctuations affectives : interrogation et surtout exclamation.

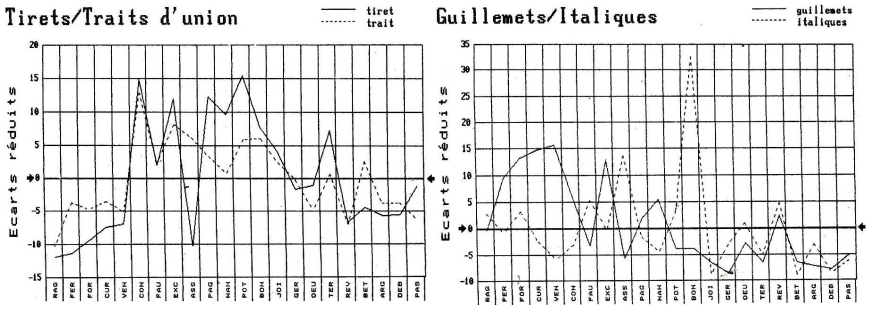

Figure 6. Courbe des signes complémentaires

3 – L'analyse s'enrichit si l'on incorpore certains signes de ponctuation dont la fonction est de signaler un changement de locuteur, qu'il s'agisse des guillemets ou des tirets de dialogue. On y ajoutera un signe assez rare dans les *Rougon-Macquart*, l'italique, et un symbole fort ambigu : le trait d'union, qui a une fonction lexicale dans les mots composés et un rôle syntaxique dans les tournures inversées et principalement dans les phrases interrogatives. Le graphique 6 montre le parallélisme du trait d'union et du tiret de dialogue, dont la courbe est pareillement parabolique avec un plateau sommital qui se maintient de la *Conquête de Plassans* à la *Joie de vivre* et un affaissement aux deux bouts de la chaîne chronologique. Cette courbe en cloche n'est pas sans évoquer celles du point et du point d'interrogation reproduites dans la figure 4. D'autre part la courbe des guillemets, nettement descendante (r = -0,72), montre l'abandon de la technique des propos enchâssés dans le discours, au profit de la transcription directe et non intégrée à la chaîne narrative. Zola utilise les guillemets chaque fois qu'il introduit des propos écrits, ou seulement pensés, ou bien encore prononcés antérieurement, chaque fois que s'établit une certaine distance avec la situation. Quand au contraire la parole est directement donnée au personnage, la réplique

intervient en début de ligne avec un tiret initial. Nous avons renouvelé l'analyse factorielle en intégrant ces nouvelles données. La figure qui en résulte se superpose à celle du graphique 5 sans en modifier l'économie. Le croissant chronologique subsiste, avec ses trois tronçons plus nettement détachés encore. Dans le premier se regroupent les premières œuvres jusqu'au *Ventre de Paris* et c'est là que les guillemets choisissent leur place. Le second réunit les textes qui s'échelonnent de la *Conquête* à la *Joie de vivre* et qui accueillent les signes d'une narration plus vive et plus riche en dialogue (point, point d'interrogation, point de suspension, tiret, trait d'union). Le troisième enfin commence avec *Germinal* et s'oriente vers un temps narratif plus large où la virgule tend à se substituer au point[10].

Un écart vaut toutefois la peine d'être noté. C'est celui de l'*Assommoir* qui déserte la place normalement assignée par le deuxième facteur, pour s'établir dans le clan opposé, là où se trouvent les grandes fresques au rythme large, du *Ventre de Paris* à la *Débâcle* en passant par *Germinal* et la *Terre*. À un moindre degré le *Bonheur des Dames* subit la même tentation et échappe en partie à son entourage chronologique. La loi chronologique n'est donc pas souveraine et Zola lui-même en avait conscience, qui distribuait ses textes « en grandes études sociales », du type de l'*Assommoir* et de *Germinal*, en « études d'histoire », comme *Son Excellence Eugène Rougon* et la *Conquête de Plassans* et en « efflorescences » plus psychologiques, comme la *Faute de l'abbé Mouret*, la *Joie de vivre*, l'*Œuvre*, le *Rêve* ou *Une page d'amour*[11]. Il y a donc, perturbant la chronologie, une volonté d'alternance, le goût de mêler les œuvres fortes, les grandes fresques descriptives dont le rythme est plus ample, et les œuvres mineures (Zola dit « en demi-teinte ») dont le cadre, narratif et psychologique, est plus traditionnel et dont l'action s'inscrit dans un rythme plus rapide. On retrouve ailleurs, par exemple dans l'étude des composants de la phrase, ces méandres de l'alternance qui brouillent le cours chronologique des *Rougon-Macquart*. Observons toutefois que dans le système des ponctuations ces méandres ont une

10. L'italique est un signe étranger au système, dont la répartition est très inégale. Zola ne s'en sert pas pour mettre en valeur un mot ou une expression mais seulement pour désigner un titre (de livre ou de chanson), ou le nom d'un magasin, d'un bistrot, d'un immeuble, d'une banque. Sous ce rapport *Au Bonheur des Dames* vient en tête, suivi de l'*Assommoir*, et dans les deux cas le titre lui-même aurait pu faire appel à l'italique.

11. Cette classification, proposée par Zola, est citée par H. Mitterand dans le tome 5 de la Pléiade, p. 1646.

faible amplitude et que la loi du temps reste prépondérante dans le rythme du discours zolien.

– III –

1 – On pourrait étudier ce rythme à un autre niveau que celui de la phrase, soit plus haut, si l'on envisage la succession des textes, soit plus bas, si l'on étudie la longueur du mot. La place nous manque ici pour pousser l'exploration de ce côté. Il nous suffira d'observer que de *Thérèse Raquin* au *Docteur Pascal*, les romans de Zola sont de plus en plus longs et que le coefficient de corrélation chronologique est significatif ($r = 0{,}65$, soit une probabilité inférieure à 0,01). Par contre la chronologie n'a pas d'effet sur la longueur du mot dont les variations tiennent au choix du sujet, les romans populaires (la *Terre* et l'*Assommoir*) privilégiant le mot court. Mais s'agissant de la phrase, notre étude ne serait pas complète si nous n'envisagions pas ce qui fait le propre de la phrase : son rythme. Et il ne suffit pas pour cela de relever le nombre de phrases dans un texte, encore faut-il en observer l'enchaînement. Car l'étude du rythme ne saurait être que dynamique.

Mais avant de descendre au niveau élémentaire où chaque phrase est étudiée dans ses relations avec celle qui la précède et celle qui la suit, nous procéderons par tranches de 5000 mots, soit 16 tranches dans *Thérèse Raquin*, 24 dans *Madeleine Férat*, 28 dans la *Fortune des Rougon* etc., et 704 au total[12]. Dans chaque tranche nous avons relevé le nombre de points (désambiguïsés). Voici, à titre d'exemple, les résultats observés dans le premier et le dernier textes du corpus.

	Thérèse Raquin	***Docteur Pascal***
nombre de tranches	16	29
nombre de points	3704	5207
nombre de points par tranche	222,38	175,03
écart-type de la distribution	19,34	26,96
coefficient de corrélation	-0,29	0,27

Effectifs des tranches

Thérèse Raquin	***Docteur Pascal***
200 232 250 237 264 221 231	172 148 135 177 172 201 215 195 152 78 166 193
207 197 224 223 192 205 238	179 186 164 167 161 180 176 178 195 183 167 209
208 229	150 167 195 219 196

12. On a négligé la tranche incomplète qui demeure en suspens à la fin de chaque texte. Précisons que la notion de « mot » ici est très extensive et comprend les noms propres et les signes de ponctuation eux-mêmes et que les points de suspension ont été incorporés aux effectifs du point (chaque triplet comptant pour une occurrence).

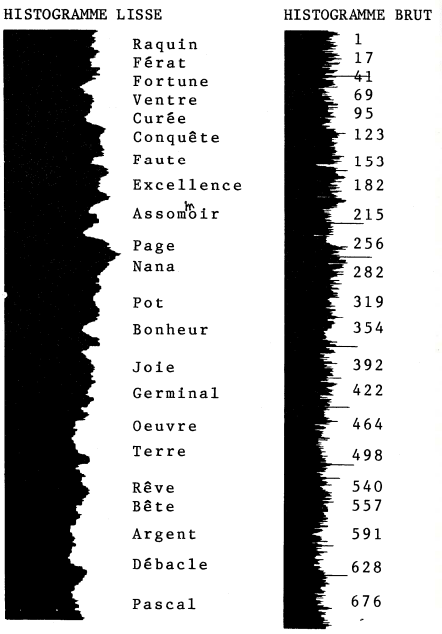

Figure 7. Histogramme du point, par tranches de 5000 mots

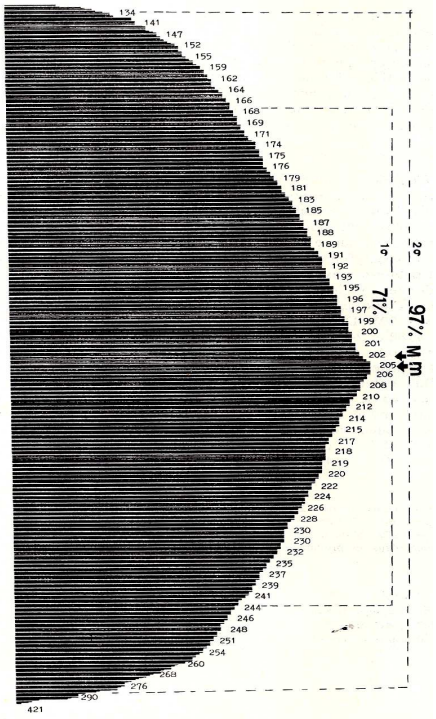

Figure 8. Courbe quasi-gaussienne de la distribution du point

Le nombre de points est en moyenne plus élevé dans *Thérèse Raquin* que dans le *Docteur Pascal* (moyenne de 222 et 175 respectivement)[13]. Étendue à l'intégralité du corpus, l'expérience donne l'histogramme de la figure 7 où les 704 tranches sont représentées (afin de reconnaître les textes dans ce continuum, on a marqué la première tranche de chacun). Comme il est difficile au regard d'apprécier un mouvement si vibratile, nous avons lissé la courbe par la méthode bien connue de la moyenne mobile. L'œil reconnaît alors (dans l'histogramme gauche) une pente descendante qui marque la raréfaction du point, ce que la figure 4 nous avait enseigné au niveau des textes. Au total on trouve une moyenne de 205,37 points par tranche, avec un écart-type de 38,347. La pente mesurée par le coefficient de Bravais-Pearson, est très significative (r = -0,527 pour 704 paires d'observations).

13. La phrase est aussi moins variée, la dispersion y étant moins forte (respectivement 19,34 et 26,96).

Avant d'en tirer d'autres conclusions, profitons de l'expérience pour apprécier la valeur de nos mesures et le bien-fondé du recours à la loi normale. Si nous ordonnons les 704 tranches de la plus pauvre en points à la plus riche, nous obtenons la figure 8, où les « bâtons » les plus longs sont ceux qui se rapprochent le plus de la moyenne, et les plus courts ceux qui s'en éloignent le plus (en moins ou en plus). La courbe n'est peut-être pas le modèle parfait de Gauss mais visuellement elle donne une impression de symétrie acceptable, avec une coïncidence presque exacte de la médiane (M = 203) et de la moyenne (m = 205). De plus à une distance de l écart-type – soit entre les valeurs 165,88 et 243,71 – on recueille 71 % des tranches, et 97 % si la distance est portée à 2 fois la valeur de l'écart-type. Ce sont des proportions proches de celles qu'aurait une population normalement distribuée. Bien entendu, les tests habituels (comme celui du signe) ne permettent pas ici de rejeter la normalité des observations. Le test du Chi2 par contre invite à conclure à la non-normalité. Pour les 704 observations la valeur du Chi2 est de 5040, ce qui représente une infime probabilité[14]. Mais on doit prendre garde qu'on se situe ici dans les grands nombres (N = 147 020) et qu'en de telles conditions le Chi2 est extrêmement sensible. Si on limite le calcul aux tranches d'un même texte et qu'on réduise en moyenne d'un facteur 20 les effectifs considérés, les valeurs du Chi2 sont beaucoup plus faibles et certaines n'atteignent pas le seuil de 5 % (*Thérèse Raquin*, *Conquête*, *Pot-Bouille*). On admettra donc que faute de mieux la loi normale reste une approximation acceptable et qu'en apportant une unité de mesure elle permet au moins la comparaison et les classements.

2 – Après ce détour méthodologique, revenons à nos tranches en envisageant un problème de critique interne : y a-t-il un mouvement dans chaque texte et ce mouvement est-il constant d'un texte à l'autre. Autrement dit, y a-t-il chez Zola une loi – ou une habitude – de

14. Les tables courantes du Chi2 ne permettent pas d'atteindre un degré de liberté aussi élevé et pour d.d.l. = 703 on doit recourir à une transformation en écart réduit (en l'occurrence z = 62,92). Mais si l'on reste dans le champ de validité des tables, nos données se situent dans les probabilités faibles :

	valeur Chi2 observée	***Chi2 pour p = 0,001***
10 premières tranches	41	31
50 premières tranches	195	89
100 premières tranches	533	153
200 premières tranches	1258	267

composition qui allongerait ou raccourcirait la phrase du début à la fin d'un texte. Les exemples du premier et du dernier textes du corpus précédemment exposés ne permettent pas de réponse assurée. Un coefficient de Bravais-Pearson a été calculé qui mesure la progression (ou la régression) du point : -0,299 dans *Thérèse Raquin*, +0,27 pour le *Docteur Pascal*. Le seuil à 5 % est pourtant atteint dans quelques cas : la *Curée*, le *Ventre de Paris*, la *Faute de l'abbé Mouret*, le *Rêve*, et à chaque fois dans le sens positif. Et là où le seuil n'est pas franchi, le signe est la plupart du temps positif. Il y a donc l'indication d'une tendance. Zola raccourcirait sa phrase à l'intérieur d'un même texte. Peut-être les nécessités de l'exposition imposent-elles au début du roman une phrase plus longue, parce que plus explicative et plus circonstanciée, alors que dans la suite des chapitres le rythme, plus dynamique, plus dramatique, s'accélère. Si l'on réunit les huit premières et les huit dernières tranches des 22 textes, on peut opposer les 35 815 occurrences du premier lot aux 36 750 du second et calculer un Chi2, qui est significatif (Chi2 = 12,05, soit une probabilité de l'ordre de 0,0005).

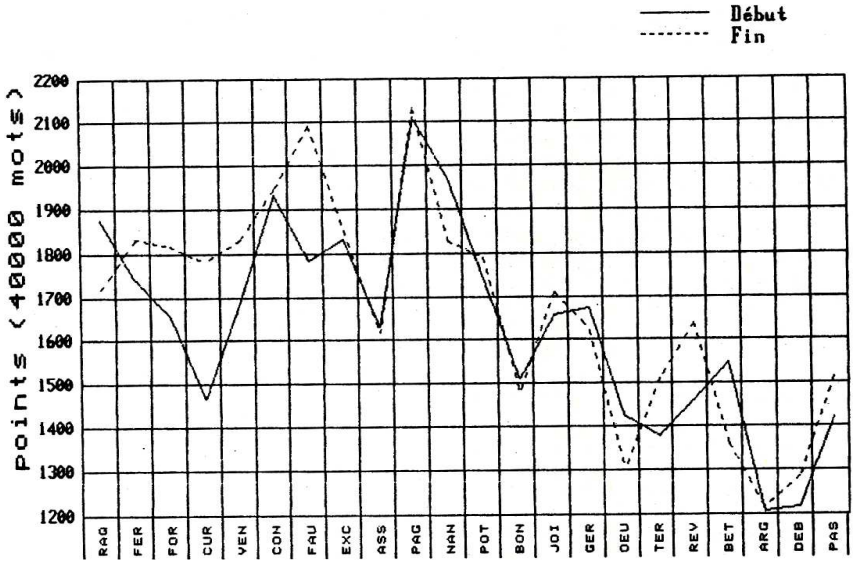

Figure 9. Comparaison du début et de la fin de chaque roman

Mais on observe que cette tendance est parfois contrariée, Zola désirant parfois surprendre son lecteur en le jetant *ex abrupto* dans le vif du sujet, au moyen d'une scène en mouvement : ainsi en est-il, dans l'*Œuvre*, de la rencontre inopinée, sous la pluie battante, de Claude et Christine. De même *Une page d'amour* s'ouvre par une scène dramatique

où la petite Jeanne se meurt devant une mère affolée. Les exordes de *Nana*, de l'*Argent*, de la *Bête humaine* manifestent de semblables motivations, qui sont moins rares à la fin des *Rougon-Macquart*, comme si Zola voulait se prémunir contre la routine menaçante. Si l'on oppose, dans les 22 textes du corpus, les 40 000 premiers mots aux 40 000 derniers, on obtient la courbe de la figure 9, où la ligne en pointillés – qui correspond aux derniers chapitres – s'élève au dessus de la ligne en traits pleins – qui symbolise les premiers chapitres –, et montre un plus fort pourcentage de points, au moins jusqu'à l'*Assommoir*[15].

3 – Reste enfin, renonçant à tout regroupement, à mener l'enquête sur chaque phrase individuelle et sur le lien qui l'enchaîne à celles qui l'entourent. Là encore, par simplification, on ne s'intéresse qu'au point. On a donc mesuré les 150 000 intervalles qui séparent les points et que l'on nomme « phrases ». Et chaque intervalle a été confronté à celui qui le précède et à celui qui le suit. Il s'agissait de savoir si le hasard préside à la succession des phrases longues et courtes, ou si au contraire une certaine aimantation s'exerce dans le discours zolien qui agglutinerait les phrases courtes en certains passages, et les phrases longues en certains autres, et qui donnerait à la succession des phrases l'apparence d'une houle, d'un rythme large dont on pourrait mesurer la longueur d'onde. Par quels moyens? Plusieurs formules s'offrent à qui traite des données sérielles, comme c'est ici le cas. Nous avons expérimenté chez Rousseau et Proust la technique des « différences quadratiques moyennes successives ». Nous utiliserons ici un procédé voisin, connu sous le nom d'autocorrélation (en anglais *serial correlation*). L'autocorrélation est un coefficient qui évolue entre -1 et +1 et mesure la corrélation entre les paires d'observations x_i et x_{i+h}. Si on donne au décalage h la valeur 1 (naturelle), on évalue la liaison établie entre toute phrase et celle qui la suit immédiatement. Rappelons la formule :

$$Rh = \frac{\sum xi\ x(i+h) - \bar{x} \sum xi}{\sum (xi - \bar{x})^2}$$

On reconnaît au dénominateur la variance de la longueur des phrases (ou plutôt de l'intervalle entre deux points), cette valeur servant de pondération. Voici les résultats obtenus dans les 1000 premières phrases de *Germinal* :

15. Bien entendu la raréfaction progressive du point tout au long des *Rougon-Macquart* incline vers le bas l'une et l'autre courbes.

nb phrases	100	200	300	400	500	600	700	800	900	1000
Rh observé	0,032	0,100	0,100	0,085	0,112	0,124	0,119	0,133	0,135	0,134
Valeur de Rh (seuil 0,02)	0,222	0,160	0,131	0,114	0,102	0,093	0,086	0,081	0,076	0,073

On voit que le seuil est atteint au bout de 500 phrases et largement dépassé à la 1000e. À plus forte raison le coefficient souligne-t-il la présence d'un rythme indubitable si les 12 000 phrases de *Germinal* sont passées au crible. Le signe (positif) du coefficient permet en outre de préciser de quel rythme il s'agit. Dans la succession des phrases trois hypothèses sont à envisager : ou bien l'écrivain ayant le souci de la variété s'applique à mêler intentionnellement phrases courtes et longues et le coefficient est alors négatif. Ou bien au contraire il donne un tempo plus régulier à sa phrase, utilisant des phrases de même type dans un même passage. Le texte obéit alors à une houle rhétorique, à un ample mouvement qui accumule les phrases courtes en certains endroits et les phrases longues en certains autres. Dans ce cas le coefficient est positif. La troisième hypothèse est l'hypothèse nulle et correspond aux valeurs (positives ou négatives) qui n'atteignent pas le seuil. Des tables permettent de savoir si le seuil est ou non dépassé, compte tenu du nombre d'observations (de phrases) envisagées. Comme ce nombre est ici très considérable, nous avons eu recours à la formule d'approximation qui fixe la limite à atteindre au seuil de 0,02.

$$\text{Rh (limite)} = \frac{2{,}326\ \sqrt{N-2} - 1}{N - 1}$$

Ainsi pour les 3704 phrases de *Thérèse Raquin* le seuil à 2 % correspond à la valeur 0,036 de *Rh*. Or l'observation fournit une valeur très supérieure : 0,130 et il en est ainsi de tous les textes (ou fragments) que nous avons réunis dans le tableau 10. La conclusion s'impose d'elle-même. La phrase de Zola – comme celles de Rousseau ou de Proust – recèle un rythme interne dont le hasard ne peut rendre compte.

4 – Reste une interrogation. Les phénomènes sériels ont parfois une phase, une périodicité. Ainsi l'étude des faits économiques doit tenir compte du rythme saisonnier, chaque mois étant lié non seulement à celui qui le précède et à celui qui le suit par le principe de la succession, mais aussi au mois correspondant des années précédentes par le principe du cycle. Peut-on faire apparaître cette phase dans la période oratoire? L'auto-corrélation propose un moyen de détecter la phase. En jouant sur la valeur du décalage (*h*), on peut mettre en relation chaque phrase non pas avec la suivante immédiate, mais avec celle qui se situe à une distance de *h* phrases. À côté de la valeur 1 que nous venons d'explorer,

nous avons essayé plusieurs valeurs : 2, 5, 9 et 17. Le résultat figure dans le tableau 10.

Tableau 10. Étude du rythme de la phrase zolienne. Autocorrélation

Texte	nombre points	nombre mots ou signes	Valeur de Rh décalage 1 phrase	décalage 5 phras.	décalage 9 phras.	décalage 17 phr.	Valeur Rh p<0,02
RAQ	3704	82832	0.130	0.064	0.013	0.016	0.036
FER	5493	121762	0.173	0.051	0.049	0.038	0.030
FOR 1	3129	74233	0.130	0.091	0.080	0.058	0.040
2	3266	73349	0.084	0.106	0.046	0.027	0.039
CUR 1	2135	57443	0.126	0.042	0.053	0.033	0.048
2	3266	75889	0.140	0.054	0.068	0.036	0.039
VEN 1	3244	76860	0.181	0.069	0.031	-0.018	0.039
2	2828	64829	0.202	0.095	0.058	0.033	0.042
CON 1	3443	71427	0.053	-0.012	0.015	0.006	0.038
2	3792	77273	0.112	0.017	0.017	0.016	0.036
FAU 1	3169	72987	0.240	0.100	0.130	0.082	0.039
2	3874	77228	0.195	0.060	0.055	0.049	0.036
EXC 1	3773	81050	0.111	0.023	0.021	-0.006	0.036
2	4213	84604	0.136	0.041	0.052	0.037	0.034
ASS 1	3069	73891	0.127	0.030	0.026	-0.014	0.040
2	3032	70713	0.111	0.035	0.027	0.028	0.040
3	2488	61363	0.102	0.042	0.020	0.037	0.044
PAG 1	3253	63362	0.169	0.045	0.073	0.030	0.039
2	3796	70829	0.162	0.046	-0.007	0.033	0.036
NAN 1	4408	90838	0.116	0.027	0.034	0.018	0.033
2	4420	95297	0.159	0.090	0.053	0.031	0.033
POT 1	4565	102029	0.114	-0.000	0.002	0.026	0.033
2	3352	76421	0.114	0.029	0.032	0.015	0.037
BON 1	4200	103210	0.166	0.054	0.057	0.016	0.034
2	3524	89312	0.199	0.074	0.068	0.070	0.037
JOI 1	4186	97475	0.150	0.047	0.022	0.028	0.034
2	2327	54970	0.137	0.057	0.051	0.047	0.046
GER 1	4798	118587	0.137	0.077	0.036	0.014	0.032
2	3752	95871	0.085	0.044	0.030	-0.011	0.036
OEU 1	3196	92639	0.182	0.073	0.040	0.008	0.039
2	2578	79236	0.130	0.034	0.030	0.031	0.044
TER 1	2859	78789	0.119	0.078	0.008	0.043	0.041
2	3146	87466	0.114	0.062	0.047	0.037	0.040
3	1931	50917	0.091	0.033	0.004	-0.017	0.050
REV	3344	87173	0.167	0.078	0.028	0.071	0.038
BET 1	3022	80374	0.152	0.057	0.074	0.067	0.040
2	3063	85760	0.166	0.037	0.039	0.047	0.040
ARG 1	2929	92757	0.185	0.091	0.082	0.041	0.041
2	2943	92954	0.184	0.119	0.030	0.023	0.041
DEB 1	2574	78760	0.154	0.060	0.022	0.024	0.044
2	2823	76717	0.112	0.045	0.039	0.013	0.042
3	2903	88836	0.153	0.053	0.020	0.046	0.041
PAS 1	2450	71460	0.358	0.211	0.185	0.092	0.045
2	2757	76189	0.126	0.086	0.046	0.004	0.042
seuil atteint n fois			44	31	23	14	

Alors que le coefficient est toujours significatif dans l'étude des phrases consécutives, la liaison est moins forte quand une phrase intermédiaire s'interpose (42 coefficients significatifs sur 44). Elle s'affaiblit quand le décalage est de 5 phrases (seuil atteint 31 fois), et tend à se diluer au delà (23 coefficients significatifs pour $h = 9$ et seulement 14 pour $h = 17$). Est-ce à dire que la périodicité n'existe pas? Certes non. Mais le retour du cycle n'a rien de mécanique : la phrase d'un écrivain est une musique dont la mesure change à chaque instant, ce qui désoriente le métronome dont nous nous servons.

Les faits de rythme sont loin de se réduire au tempo. Ils s'attachent aussi au retour des thèmes mélodiques et au rappel des teintes harmoniques. Pour aller au fond des choses, il ne suffit pas de compter les mots, encore faut-il les identifier et noter les répétitions, les variations, les modulations. L'étude des répétitions n'échappe pas tout-à-fait à l'ordinateur et nous l'avons entreprise chez Giraudoux dans le cadre restreint de la phrase. Chez Zola pareille étude serait certainement intéressante à condition d'envisager des unités plus larges. Car les fameuses répétitions de Zola se situent rarement dans la même phrase ou dans la même page. Les *leitmotive* de Zola agissent sur la mémoire du lecteur en vertu d' une rémanence des images, des mots et des formules qui se prolonge sur tout un chapitre et même sur un roman complet, voire sur le cycle entier des *Rougon-Macquart*[16]. Le rythme de la phrase zolienne ne peut donc être complètement isolé de la thématique.

Et de la même façon une telle étude ne saurait être séparée de celle de la structure de la phrase, et des habitudes syntaxiques propres à Zola. Nous ne pouvons ici qu'amorcer l'ébauche de cette question que nous traitons ailleurs. Bornons nous à constater le déficit des catégories grammaticales qui servent à structurer la phrase : coordonnants. subordonnants, relatifs, interrogatifs, adverbes de liaison. La phrase de Zola a donc une structure simple, sans ces multiples niveaux et recoins où se perd le lecteur de Proust. Rythme et syntaxe, longueur et structure en sont des aspects liés et complémentaires.

En outre on a de bonnes raisons d'estimer que ces aspects s'accordent aussi au contenu lexical, à cet anti-intellectualisme de Zola[17] qui lui fait choisir les mots concrets plutôt que l'abstraction et les suffixes, et qui l'entraîne à faire voir, à évoquer les êtres et les choses, plutôt qu'à les démonter ou les démontrer.

---

16. Le plus bel exemple de *leitmotiv* est celui de la dernière phrase des *Rougon-Macquart* qui s'achève sur l'évocation d'un nourrisson « qui tétait toujours, son petit bras en l'air, tout droit, dressé comme un drapeau d'appel à la vie » et qui reprend en l'orchestrant un thème amorcé trois pages auparavant : « il s'était mis à lever son petit bras en l'air, tout droit ainsi qu'un drapeau. » La Pléiade, tome V, p. 1217 et l220. On trouvera dans le même *Docteur Pascal*, un autre exemple presque obsessionnel présenté d'abord p. 1050 : « un cou délicat surtout, satiné et rond, ombré de cheveux follets sur la nuque » et repris textuellement à la page suivante avec la seule modification de l'article initial.

17. Cf. David Baguley, « L'anti-intellectualisme de Zola », *Cahiers Naturalistes*, numéro 42, 1971, p. 119-129.

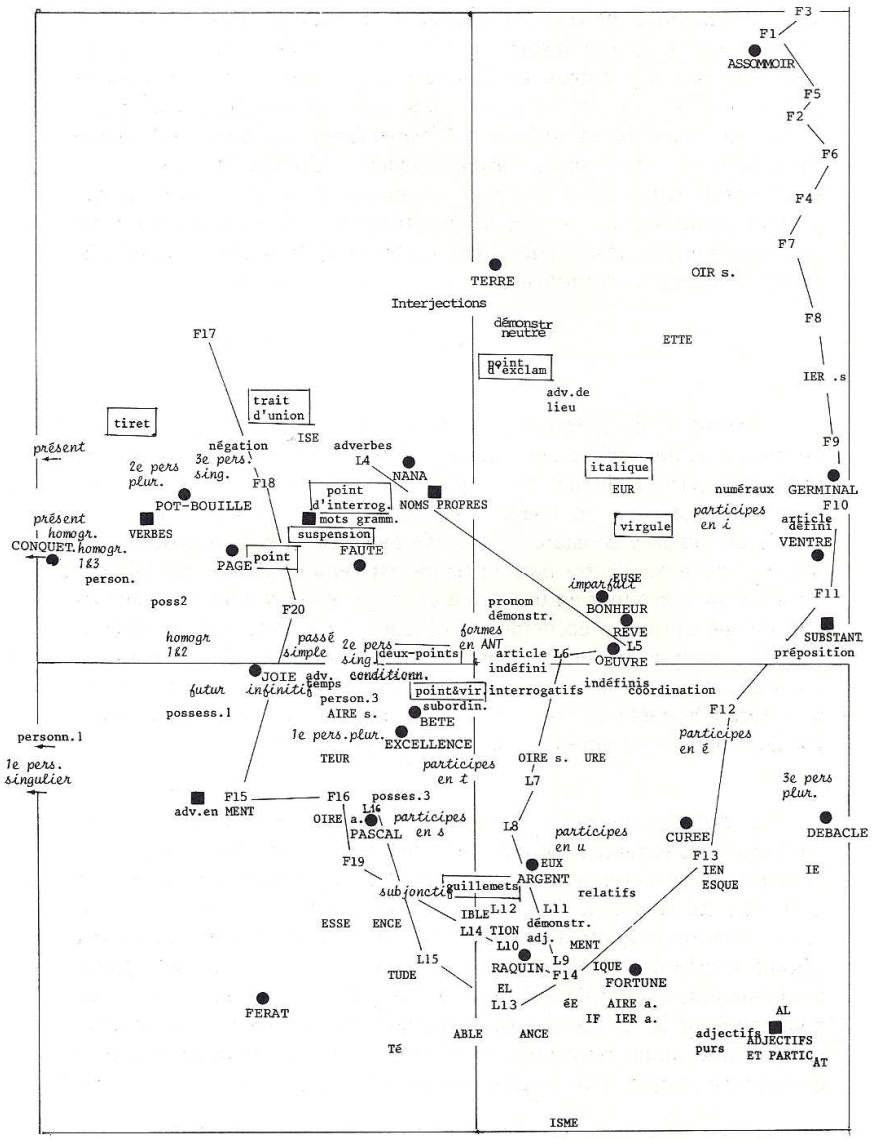

Figure 11. Analyse factorielle récapitulative des éléments quantifiables dans le corpus Zola

L'étude d'ensemble du vocabulaire de Zola permet d'accéder ainsi à une certaine cohérence, à une vision générale dont la présente étude n'est qu'un point de vue très fragmentaire. Nous en donnerons un aperçu dans la figure 11 qui présente une analyse récapitulative de toutes les données

quantitatives relevées chez Zola : parties du discours, espèces de mots grammaticaux, classes de fréquence, classes de longueur de mots, sous-classes d'adjectifs et de participes, temps, modes et personnes des verbes, classes de suffixes et enfin signes de ponctuation. On notera que la chronologie n'est plus le facteur dominant et que le genre impose sa loi, qui distribue à gauche les textes où le verbe l'emporte (avec les adverbes) et à droite ceux qui privilégient les catégories nominales, soit le substantif (quadrant supérieur), soit l'adjectif (quadrant inférieur). Dans le premier lot figurent les romans d'action, comme la *Conquête de Plassans*, qui recourent au dialogue, aux temps du présent, aux personnels et aux possessifs. Là le rythme est plus rapide et Zola multiplie le point (et aussi le point d'interrogation et les points de suspension). Dans le second groupe l'analyse place les grandes fresques qui requièrent les vertus descriptives du substantif et de l'adjectif (et des catégories associées : articles, numéraux, prépositions, coordination). Là le lexique sépare les romans populaires (l'*Assommoir*, la *Terre*, *Germinal*) dont le vocabulaire est plus concret et plus technique, et les romans bourgeois (la *Fortune*, la *Curée*, l'*Argent*) où s'accumulent les suffixes plus abstraits. De ce côté descriptif, un seul signe de ponctuation trouve refuge : la virgule.

La perspective chronologique que nous avait ouverte l'étude séparée du système des ponctuations est-elle donc un trompe-l'œil? Non, car nous la retrouvons dans la même analyse avec le troisième facteur qui range d'un côté de l'axe les 11 premiers titres et de l'autre les 9 derniers. La loi du temps agit dans les *Rougon-Macquart* mais son action est moins puissante que celle du genre, comme nous l'avons observé dans le grand corpus littéraire du *Trésor de la langue française*[18]. Pourtant dans ce corpus aussi la ponctuation se montrait surtout sensible à la chronologie, sans doute parce que le système des ponctuations n'est pas encore complètement stabilisé au XIX^{e} siècle. En particulier tout au long de ce siècle la virgule se développe pour atteindre son sommet à l'époque de Zola, en liaison avec le progrès de l'adjectif et du participe. Et le goût de Zola pour la description le pousse à surenchérir sur son époque et à développer certains traits que le naturalisme avait hérités du romantisme. Personne n'ira plus loin dans ce sens et après Zola la littérature française rebrousse chemin.

18. Cf. notre *Vocabulaire français de 1789 à nos jours*, Slatkine, (1981a), 3 vol.