

Les Rougon-Macquart à l'ère du numérique: enquête textométrique et scripturale sur la poétique zolienne avec Hyperbase

**Mohamed Seghir
BENOTMANE(*)**

Université Badji Mokhtar – Annaba,
B.P 12 Annaba 23000.

mohamedseghir.benotmane@univ-annaba.dz

Soumis le: 01/05/2026

Accepté le: 18/06/2026

Publié le: 30/06/2026

Résumé

Cet article propose une lecture croisée des apports de la textométrie quantitative et de l'interprétation psychocritique pour l'analyse de l'écriture d'Émile Zola dans le cycle des Rougon-Macquart. S'appuyant sur le logiciel Hyperbase, la recherche interroge les récurrences lexicales, les hapax et les réseaux de cooccurrences comme révélateurs d'un habitus scriptural singulier. Elle articule les données statistiques à une réflexion sur l'intentionnalité auctoriale et les mécanismes cognitifs qui sous-tendent la cohérence interne de l'œuvre zolienne, en accordant une attention particulière aux dimensions sensorielle, énonciative et corporelle du style naturaliste.

Mots-clés: *Textométrie, hapax, stylistique, habitus scriptural, Hyperbase.*

The Rougon-Macquart in the Digital Age: A Textometric and Stylistic Inquiry into Zola's Poetics with Hyperbase

Abstract

This article offers a cross-reading of the contributions of quantitative textometry and psychocritical interpretation to the analysis of Émile Zola's writing in the Rougon-Macquart cycle. Drawing on the Hyperbase software, the study examines lexical recurrences, hapax, and networks of co-occurrences as indicators of a distinctive scriptural habitus. It links statistical data to a reflection on authorial intentionality and the cognitive mechanisms underpinning the internal coherence of Zola's work, with a particular attention paid to the sensory, enunciative and corporal dimensions of the naturalist style.

Keywords: *Textometry, hapax, stylistics, scriptural habitus, Hyperbase.*

*Auteur correspondant: Mohamed Seghir BENOTMANE, mohamedseghir.benotmane@univ-annaba.dz



Introduction:

La question du style littéraire a longtemps relevé de l'appréciation qualitative, du jugement esthétique et de la lecture herméneutique. Depuis les travaux fondateurs de Charles Bally sur la stylistique générale (Bally, 1909) et ceux de Mikhaïl Bakhtine sur le discours romanesque (Bakhtine, 1978), les études littéraires ont progressivement intégré des outils formels pour objectiver les intuitions critiques. C'est dans ce sillage épistémologique que s'inscrit la textométrie, discipline née de l'alliance entre la linguistique de corpus et les sciences computationnelles, qui offre des instruments statistiques rigoureux pour l'exploration des grands ensembles textuels (Lebart & Salem, 1994).

L'œuvre d'Émile Zola, et plus particulièrement le cycle des *Rougon-Macquart* «cette histoire naturelle et sociale d'une famille sous le Second Empire», composée de vingt romans écrits entre 1871 et 1893, constitue un terrain d'investigation privilégié pour ces approches. La monumentalité du corpus, sa cohérence idéologique et la singularité stylistique de son auteur en font un objet d'analyse idéal pour qui cherche à articuler quantification lexicale et interprétation littéraire (Mitterand, 2002). Comme le souligne Mitterand, les «réalisations romanesques de Zola ne sont pas conformes à ses théories» et recourent à des «descriptions hallucinatoires et visionnaires, [à] un style poétique et [au] lyrisme» (Mitterand, 2002: p. 201). Déjà, dans *Le Roman expérimental*, Zola revendiquait pour le roman un statut de laboratoire: «J'ai simplement fait sur deux corps vivants le travail analytique que les chirurgiens font sur des cadavres» (Zola, 1880/1971: p. 321), définissant le naturalisme comme «l'esprit scientifique porté dans toutes nos connaissances». Ce paradoxe entre ambition scientifique affichée et débordement stylistique constant invite à se demander si l'écriture zolienne ne constitue pas, au-delà de la diversité de ses objets et de ses milieux, un système cohérent de préférences et d'habitudes scripturales que la variation thématique ne suffit pas à dissoudre.

Le présent article se propose d'examiner de manière systématique les apports de l'analyse textométrique conduite à l'aide du logiciel Hyperbase à la connaissance de la poétique zolienne. Il vise à montrer comment les récurrences lexicales, la distribution des hapax, les réseaux de cooccurrences et les stratégies énonciatives permettent de dégager des constantes stylistiques et de construire une image cohérente de l'écrivain comme sujet psycholinguistique une image que ni la lecture cursive ni l'intuition critique ne peuvent seules établir avec rigueur. L'étude s'appuie sur une perspective pluridisciplinaire articulant stylistique, psychocritique et psycholinguistique cognitive pour proposer une lecture renouvelée de la créativité zolienne: celle d'un style mesurable qui, à travers ses routines comme dans ses inventions ponctuelles, transcende les contraintes propres à chaque roman pour révéler la cohérence profonde d'une subjectivité créatrice singulière.

Afin de conduire cette investigation de manière rigoureuse et progressive, l'article s'organise en sept mouvements complémentaires. Une première partie pose les fondements épistémologiques de la textométrie et présente l'architecture fonctionnelle du logiciel Hyperbase comme outil d'exploration des grands corpus littéraires. La deuxième partie est consacrée à la constitution et à la caractérisation du corpus zolien, en définissant les *Rougon-Macquart* comme laboratoire stylistique et en exposant le protocole d'analyse retenu. La troisième partie examine les récurrences lexicales et les segments répétés comme marqueurs d'un idiolecte auctorial stable, en accordant une attention particulière au réseau lexical du corps. La quatrième partie propose une étude de cas centrée sur *L'Assommoir*, roman où les spécificités énonciatives, syntaxiques et argotiques du style zolien atteignent leur point de cristallisation maximale. La cinquième partie analyse la distribution et la nature des hapax à l'échelle du cycle, en distinguant hapax référentiels et hapax expressifs comme deux modalités complémentaires de la créativité lexicale. La sixième partie articule ces données quantitatives à une réflexion sur la dimension psychologique de l'écriture, en mobilisant les notions d'habitus scriptural, d'intentionnalité auctoriale et de profil psycholinguistique. Enfin, une septième partie



ouvre des perspectives méthodologiques vers une textométrie augmentée par les outils du Traitement Automatique du Langage et vers une poétique cognitive comparée.

1- La textométrie comme méthode: fondements épistémologiques

1-1- De la statistique textuelle à la textométrie:

La textométrie est définie comme «l'ensemble des méthodes quantitatives appliquées à des corpus textuels en vue d'en analyser les structures lexicales, stylistiques et sémantiques» (Lebart & Salem, 1994: p. 12). Née dans les années 1960 à la confluence de la linguistique computationnelle et de l'analyse des données, elle a connu un essor considérable avec la numérisation des corpus littéraires. Les travaux pionniers d'Étienne Brunet, concepteur du logiciel Hyperbase, ont particulièrement contribué à fonder cette discipline en France (Brunet, 1988), en proposant une méthodologie rigoureuse pour l'étude des fréquences lexicales et des cooccurrences dans les textes littéraires.

La textométrie se distingue de la simple statistique textuelle par son ambition interprétative: elle ne se contente pas de dénombrer les occurrences, mais cherche à dégager des significations à partir des régularités observées. Comme le précise Pincemin (2012, p. 7), «la textométrie est une discipline d'interprétation, non de simple description» Cette orientation herméneutique est essentielle pour une application aux textes littéraires, où la signification excède toujours la simple présence statistique des formes. C'est précisément cette tension productive entre quantification et interprétation qui fonde l'originalité de l'approche développée dans l'analyse du corpus zolien.

1-2-Hyperbase: architecture et fonctionnalités

Le logiciel Hyperbase, développé et diffusé par le CNRS et l'Université Nice Côte d'Azur, représente l'un des outils textométriques les plus complets dans la communauté francophone. Conçu initialement pour le traitement du français et du latin, il s'est progressivement étendu aux grandes langues européennes. Son architecture repose sur un tableau lexical croisé qui permet de mesurer la distribution des formes à travers les différentes parties d'un corpus, d'identifier les spécificités lexicales de chaque texte et de visualiser les réseaux d'associations entre les mots (Vanni, 2024).

Les fonctionnalités d'Hyperbase se déclinent en plusieurs modules complémentaires: l'analyse des fréquences et des spécificités lexicales, l'étude des cooccurrences et des corrélats, la détection automatique des hapax et des formes rares, et depuis 2018, l'intégration d'algorithmes de *deeplearning* permettant le repérage automatique de motifs linguistiques profonds. Dans le contexte de l'analyse du corpus zolien, ces fonctionnalités permettent d'opérationnaliser des hypothèses stylistiques et de les soumettre à la vérification empirique.

2- Le corpus zolien: construction et enjeux méthodologiques

2-1- Les Rougon-Macquart comme laboratoire stylistique et objet textométrique:

Le cycle des *Rougon-Macquart* offre un corpus d'une richesse exceptionnelle pour l'analyse textométrique du style littéraire. Composé de vingt romans rédigés entre 1871 et 1893, il constitue une somme stylistique unique dans l'histoire du roman français: un même auteur traverse, en deux décennies de production intense, des milieux sociaux radicalement hétérogènes: la bourgeoisie provinciale de *La Conquête de Plassans*, le monde ouvrier de *L'Assommoir* et de *Germinal*, le commerce parisien d'*Au Bonheur des Dames*, l'aristocratie de *Nana*, l'armée de *La Débâcle*, tout en maintenant une cohérence formelle et idéologique fondée sur le déterminisme héréditaire et le naturalisme scientifique (Mitterand, 2002). C'est précisément cette tension entre variation thématique et unité stylistique qui fait du cycle un terrain d'investigation privilégié: il permet de mesurer, sur une base quantitative, ce qui relève du style propre à chaque roman et ce qui constitue l'idiolecte profond et constant de l'auteur, traversant tous les milieux et tous les registres.

Car Zola ne change pas seulement de décor d'un roman à l'autre: il module son lexique, reconfigure ses réseaux sémantiques, adapte ses structures syntaxiques aux contraintes de chaque univers social. L'analyse textométrique rend visible ce travail de modulation stylistique



que la lecture cursive perçoit intuitivement sans pouvoir le mesurer. Les spécificités lexicales de chaque roman identifiées par les indices de fréquence et de spécificité d'Hyperbase, révèlent un écrivain qui ne réutilise pas mécaniquement son propre langage, mais le réinvente à chaque œuvre tout en conservant des structures profondes reconnaissables: la densité des substantifs concrets, la récurrence des métaphores biologiques, la progression des registres sensoriels.

Zola a lui-même théorisé ce travail stylistique sous les dehors d'une ambition scientifique. Dans *Le Roman expérimental* (1880), il s'inspire explicitement du physiologiste Claude Bernard pour définir le romancier naturaliste comme celui qui «montre dans l'homme et dans la société le mécanisme des phénomènes dont la science est maîtresse» (Zola, 1880/1971: p. 18). Mais c'est Mitterand (2002) qui déplace le regard vers ce que cette déclaration dissimule: les «réalisations romanesques de Zola ne sont pas conformes à ses théories» et font appel à des «descriptions hallucinatoires et visionnaires, [à] un style poétique et [au] lyrisme» (p. 201) que le programme naturaliste officiel ne saurait contenir. La textométrie opère ici une révélation décisive: en mesurant les fréquences, les cooccurrences et les hapax à l'échelle du corpus entier, elle confirme que le style zolien excède toujours le projet naturaliste, que l'écrivain déborde constamment le théoricien. L'analyse quantitative du corpus redouble ainsi, au niveau du métalangage critique, la tension constitutive de l'écriture zolienne elle-même entre l'ambition d'objectivité scientifique revendiquée et la puissance d'une subjectivité stylistique que les chiffres, paradoxalement, sont les mieux placés pour révéler.

2-2- Protocole d'analyse et grille de lecture:

La constitution d'une grille de lecture systématique est une étape fondamentale de toute démarche textométrique appliquée à un corpus littéraire. Dans l'analyse du corpus zolien conduite avec Hyperbase, cette grille s'articule autour de plusieurs paramètres: la segmentation du corpus en unités textuelles cohérentes (les romans du cycle), l'identification des formes graphiques et de leurs lemmes, la distinction entre vocabulaire commun et vocabulaire spécifique, et le traitement particulier des formes rares dont les hapax constituent le cas limite.

La méthodologie adoptée s'appuie sur les protocoles formalisés par Lebart et Salem (1994) et sur les recommandations spécifiques à l'analyse des corpus littéraires développées dans les travaux de Brunet (2004). Elle intègre également les principes de la sémantique interprétative tels que les a formulés Pincemin (2012), selon lesquels la signification ne réside pas dans les formes isolées mais dans les relations qu'elles entretiennent entre elles et avec leur contexte discursif.

Tableau1: Grille d'analyse textométrique du corpus des *Rougon-Macquart*

Paramètre	Opération	Outil / Méthode	Objectif stylistique
Segmentation du corpus	Découpage en 20 unités textuelles (un roman = une unité)	Hyperbase – tableau lexical croisé	Mesurer la variation et la constance stylistiques d'un roman à l'autre
Identification des formes	Recensement des formes graphiques et association à leur lemme	Lemmatisation automatique + vérification manuelle	Neutraliser les variations morphologiques pour comparer les fréquences réelles
Distinction lexicale	Séparation du vocabulaire commun (fréquent, partagé) et du vocabulaire spécifique (sur/sous-représenté par roman)	Calcul des spécificités (indice de Brunet)	Dégager l'idiolecte propre à chaque roman et les constantes de l'idiolecte auctorial



Traitement des formes rares	Identification et classement des hapax (<i>legomena</i>) par roman et pour l'ensemble du cycle	Détection automatique Hyperbase + analyse qualitative	Repérer les zones de tension créatrice maximale et les néologismes stylistiques
Analyse des cooccurrences	Cartographie des associations préférentielles entre formes dans une fenêtre contextuelle définie	AFC (Analyse Factorielle des Correspondances) + CDH	Révéler les réseaux sémantiques obsédants (corps, mine, foule, lumière...)
Interprétation stylistique	Mise en relation des données chiffrées avec les hypothèses littéraires	Sémantique interprétative (Pincemin, 2012)	Articuler quantification et herméneutique pour dépasser la simple description statistique

Source: PINCEMIN, B (2012). «Sémantique interprétative et textométrie», *Texto !*, Volume XVII, n°3.

La grille est enrichie de six paramètres au lieu de quatre, en intégrant l'analyse des cooccurrences et l'interprétation stylistique finale: deux étapes absentes de la version originale mais indispensables dans la pratique réelle d'Hyperbase. Afin d'éviter une liste purement technique, chaque entrée est structurée selon un enchaînement logique en trois temps: l'opération à conduire, l'outil ou la méthode mobilisée, et la finalité stylistique visée. C'est cette troisième colonne qui constitue la véritable valeur ajoutée de la grille révisée: en ancrant chaque paramètre dans la problématique centrale de l'article, elle rappelle en permanence que l'objectif n'est pas de produire des statistiques, mais de comprendre et d'interpréter le style de Zola, sa cohérence, ses variations, ses tensions créatrices, à partir des régularités que le calcul permet de mettre au jour.

3- Récurrences lexicales et constantes stylistiques dans l'écriture zolienne:

3-1- Le paradoxe lexical de Zola:

La première révélation que réserve l'approche quantitative concerne la richesse lexicale de Zola. Étienne Brunet, dans une étude pionnière soumettant l'ensemble du cycle au calcul statistique, formule ce constat contre-intuitif: «Le vocabulaire de Zola est plus pauvre que celui de Proust» (Brunet, 1985: p. 36). En recourant à l'indice *w* de mesure de la variété lexicale et à la loi binomiale, il établit que *Germinal* (172 410 occurrences, 7 775 vocables), *L'Assommoir* (165 017 occurrences, 7 383 vocables) et *La Débâcle* (195 340 occurrences, 7 825 vocables) affichent des taux de richesse lexicale systématiquement inférieurs à ceux de *La Prisonnière* de Proust, malgré des longueurs comparables (Brunet, 1985: p. 36-37).

Ce paradoxe s'explique par une donnée structurelle fondamentale: le lexique zolien fuit l'abstraction. L'analyse des suffixes révèle un déficit statistiquement significatif des suffixes conceptuels (*-tion*, *-té*, *-isme*: écart réduit respectivement de -44,55, -20,78, -14,01) et un excédent marqué des suffixes désignant des agents et des instruments concrets (*-ier*, *-oir*, *-ette*) (Brunet, 1985, p. 38). Zola ne construit pas un lexique de la pensée, mais un lexique de la chose. Son catalogue, note Brunet (1985), «fait penser à une encyclopédie non à un dictionnaire de langue» (p. 38). Cela s'observe dans *Au Bonheur des Dames*, où la liste des étoffes produit cet effet d'accumulation nominale: «Des satins et des velours, des failles et des gros de Tours, des moires et des damas s'étaient, débordaient, noyaient les comptoirs sous leurs flots pressés» (Zola, 1883/1980: p. 107). La prédominance des substantifs concrets et des adjectifs attributifs correspond précisément à ce que l'analyse factorielle des correspondances de Brunet (1985) visualise: les textes à dominante descriptive se regroupent dans le quadrant supérieur du plan factoriel.

3-2- La stratégie des segments répétés:



L'un des apports les plus spectaculaires de la textométrie appliquée aux *Rougon-Macquart* est la mise en évidence d'une écriture fondée sur la répétition systématique de segments. Charnois, Delente et Legallois (2013), en appliquant l'analyse par segments répétés, définis comme des «suites récurrentes de mots, devenues facilement détectables par des outils simples» comme le concordancier AntConc, révèlent que Zola use d'une stratégie d'écriture où des segments allant du syntagme au paragraphe entier se trouvent reproduits à plusieurs centaines de pages de distance, y compris entre des romans différents du cycle. Ce phénomène, «parfois aperçu par la critique, mais dont l'ampleur n'a jamais été mesurée ni appréciée» (Charnois et al, 2013), confère au cycle une cohérence interne qui dépasse la seule logique narrative.

La mine dans *Germinal* en offre une illustration paradigmatique: «Le Voreux, au fond de son trou, avec son tassement de bête mauvaise, semblait mâcher davantage» (Zola, 1885/1978: p. 34). Cette métaphore anthropomorphe du puits-dévorateur revient sous des formes légèrement reconfigurées à plusieurs reprises dans le roman, fonctionnant comme un *leitmotiv* syntaxique identifiable par la fouille de motifs séquentiels. La méthode, qui détecte des associations récurrentes d'éléments non nécessairement contigus portant simultanément sur des formes, des lemmes et des catégories grammaticales, permet de formaliser ce que Charnois et al. (2013) appellent les «caractéristiques autoriales» qui émergent des spécificités de ces motifs.

3-3- Le corps comme réseau lexical:

Une littérature récente a élargi cette approche textométrique à la question du corps dans les *Rougon-Macquart*, en soumettant les vingt romans à la Classification Descendante Hiérarchique (CDH) et à l'Analyse Factorielle des Correspondances (AFC). Cette démarche a permis d'identifier les segments textuels à forte concentration de lexique corporel et de cartographier les «rapports géométriques, de proximité ou de distance» entre les mots et les textes sur un plan factoriel bidimensionnel (Filopone 2025: p. 1). L'AFC révèle notamment que les romans à dominante populaire (*Germinal*, *L'Assommoir*, *La Terre*) partagent un réseau lexical corporel dense centré sur la fatigue, la brutalité et la déformation physique, tandis que les romans bourgeois déplacent ce réseau vers la séduction et la maladie.

Dans *Nana*, le corps de l'héroïne devient ainsi un nœud lexical autour duquel gravitent les champs sémantiques du désir, de la corruption et de la mort, comme en témoigne cette description dont les termes-clés (*chair, or, animal*) sont statistiquement sur-représentés dans l'AFC: «C'était Vénus qui naissait de la mer, n'ayant pour voile que ses cheveux» (Zola, 1880/1977: p. 49). Cette cartographie quantitative rejoint et valide la thèse de Mitterand selon laquelle «le jeu linguistique de la description romanesque accorde un caractère magique à toutes ces oppositions sociales» (cité dans Brunet, 1985).

4-L'Assommoir: étude de cas textométrique

4-1- Le discours indirect libre comme signature énonciative:

L'Assommoir (1877), septième volume du cycle, représente un tournant stylistique radical dans l'œuvre de Zola et dans l'histoire du roman français. Zola lui-même revendique dans sa *Préface* ce geste philologique singulier: «La forme seule a effaré. Mon crime est d'avoir eu la curiosité littéraire de ramasser et de couler dans un moule très travaillé la langue du peuple» (Zola, 1877/1960, cité dans Marosvári, 1990: p. 5).

Sur le plan énonciatif, l'analyse génétique et textométrique du discours indirect libre (DIL) constitue un terrain de recherche particulièrement fécond. Les travaux récents en génétique textuelle montrent que Zola, dès la parution en feuilleton, modifie les occurrences du DIL selon trois opérations principales: la substitution, la suppression et l'ajout (Genesis, 2024: p. 3). Ces corrections tendent à «atténuer la subjectivité des personnages» tout en faisant apparaître «le point de vue [du narrateur] sous la forme d'une tournure objectiviste, parfois teintée d'ironie» (Genesis, 2024: p. 6). L'extrait suivant illustre ce dispositif, où la voix de Gervaise et celle du narrateur se fondent sans couture: «Ah ! que le travail était dur, et que la vie serait dure si ça continuait ! Elle avait mal partout déjà» (Zola, 1877/1978: p. 56). Cette technique, désignée par



les stylisticiens comme une polyphonie contrôlée, est mesurable textométriquement par le repérage automatisé des marqueurs du DIL (imparfait, modalisation, deixis temporelle) à l'échelle du corpus entier.

4-2- Les syntagmes péjoratifs et la syntaxe populaire:

L'une des signatures lexicales les plus saillantes de *L'Assommoir*, détectable par fouille de motifs séquentiels, est le syntagme de la forme [démonstratif + terme d'injure + *de* + *nom propre*]. Léonard (1974) en recense plus d'une centaine d'occurrences dans le seul roman: «*ce sacré sournois de Lantier*» (p. 312), «*cet animal de Boche*» (p. 251), «*cette grande carcasse de Virginie*» (p. 46). Ce paradigme est exceptionnellement riche et «tous les personnages du roman paraissent liés à ce syntagme» (Léonard, 1974: p. 45). D'un point de vue textométrique, ce moule syntaxique constitue un marqueur auctorial mesurable qui opère comme un opérateur d'uniformisation entre la voix narrative et les paroles des personnages.

Deux pronoms fonctionnent également dans *L'Assommoir* comme opérateurs textuels à très haute fréquence: *ça* et *on*. Le pronom *ça*, «degré zéro du syntagme nominal de type inanimé, de forme essentiellement parlée» (Léonard, 1974: p. 51), constitue l'indice le plus fiable du style indirect libre dans le roman. Son rôle est double: anaphore ou cataphore, il structure des phrases disloquées caractéristiques de la syntaxe affective populaire, comme dans: «*Ça lui semblait trop dur, de ne plus pouvoir se remuer chez elle*» (Zola, 1877/1960: p. 354, cité dans Léonard, 1974: p. 52). Quant au pronom *on*, il «permet à Zola de neutraliser différentes oppositions sur lesquelles le récit habituellement se fonde» (Léonard, 1974: p. 53), fonctionnant tantôt comme équivalent de *ils*, tantôt comme pronom inclusif qui associe imperceptiblement le lecteur au groupe ouvrier.

4-3- L'uniformisation progressive: vers le monologue intérieur

L'originalité lexicale de *L'Assommoir* se manifeste surtout dans la *progression narrative* des formes argotiques. Léonard (1974) a mis en évidence la courbe d'effacement du discours direct au fil des chapitres: son taux est maximum au chapitre I, minimum au chapitre XII. À mesure que Gervaise sombre, la parole directe disparaît au profit d'un «ruminement intérieur» (Léonard, 1974: p. 50) qui fonde les soliloques des derniers chapitres. La langue argotique, d'abord marqueur de dialogues, devient alors le tissu même de la conscience narrante, comme dans ce passage du chapitre XII: «*Vrai, elle le trouvait trop rossard, cet entripaillé, elle l'avait où vous savez, et profondément encore !*» (Zola, 1877/1960: p. 443-444, cité dans Léonard, 1974: p. 50). Dubois analyse ce mouvement comme une «prose doublement artiste» où «l'indirect libre remplit une fonction narrative: la narration d'auteur s'oralise. La stylisation est inscrite au service de la production d'un effet de réel» (cité dans Marosvári, 1990: p. 5).

5- Les hapax zoliens: singularité lexicale et créativité stylistique

5-1- Définition et enjeux théoriques de l'hapax:

Le terme *hapax legomenon* (du grec, «dit une seule fois») désigne une forme lexicale n'apparaissant qu'à une seule occurrence dans un corpus donné. Dans le domaine de la stylistique quantitative, les hapax constituent un objet d'étude particulièrement riche, car ils se situent précisément à l'intersection de la norme linguistique et de l'écart stylistique. Leur présence signale une volonté de singularisation lexicale, une recherche d'expressivité qui ne se satisfait pas des ressources ordinaires de la langue (Kozak, 2018). L'étude systématique des hapax dans le corpus zolien innove en proposant une analyse quantifiée à l'échelle du cycle entier, permettant de dégager des tendances générales et de situer les singularités lexicales dans leur contexte stylistique global.

L'analyse des hapax dans le corpus des *Rougon-Macquart* révèle une double logique à l'œuvre. D'une part, certains hapax répondent à une nécessité descriptive: la volonté de rendre avec précision des réalités techniques, professionnelles ou sociales propres aux milieux décrits oblige Zola à forger ou à emprunter des termes spécialisés. Ces hapax «référentiels» témoignent de la conscience documentaire de l'auteur, de son souci d'exactitude ethnographique que les carnets d'enquête conservés à la Bibliothèque nationale attestent (Mitterand, 1986). D'autre part,



certain hapax obéissent à une logique purement stylistique: ils constituent des inventions lexicales par lesquelles Zola cherche à saisir des sensations ou des états affectifs que le vocabulaire ordinaire ne suffit pas à nommer. Ces hapax «expressifs» sont les marqueurs les plus sûrs de la créativité littéraire de l'auteur, les points où son langage se tend au maximum de ses possibilités inventives — que ce soit dans la violence pulsionnelle de *La Bête humaine* ou la déchéance physiologique de *Nana*.

5-2-Hapax dans les *Rougon-Macquart*: vue d'ensemble

Le corpus des *Rougon-Macquart* contient, 13442 hapax au total, c'est-à-dire, 13442 vocables employés une seule fois dans l'ensemble du cycle. Ces hapax se répartissent de manière très inégale selon les romans, et leur distribution révèle les zones de tension créatrice maximale dans l'écriture zolienne.

Les trois romans les plus riches en hapax (par rapport au corpus du *Trésor de la Langue Française*) sont:

Tableau2: récapitulatif des trois romans les plus riches en Hapax, dans les *Rougon-Macquart*, enrichie par les données de l'étude textométrique d'Étienne Brunet

Roman	Hapax (par rapport au TLF)	Remarque
<i>L'Assommoir</i>	38	Quasi ¼ du total — record absolu
<i>La Terre</i>	25	Lexique agricole et paysan rare
<i>Germinal</i>	13	Vocabulaire minier spécialisé
<i>Le Rêve</i>	16	Lexique religieux et médiéval
<i>La Faute de l'abbé Mouret</i>	15	Vocabulaire botanique exotique

Sources: Hyperbase version 9.0

Dans *L'Assommoir*: hapax argotiques

L'Assommoir domine de loin avec 38 hapax que l'on ne trouve nulle part ailleurs dans la littérature française recensée par le TLF. Zola était lui-même conscient de cette singularité: il proposa à ses lecteurs un glossaire explicatif annexé au roman. Ces hapax relèvent essentiellement du vocabulaire argotique et populaire parisien que Zola voulait «couler dans un moule très travaillé» formes comme:

- 1- *entripaillé* (insulte populaire désignant un gros homme ventru)
- 2- *rossard* (fainéant, terme argotique)
- 3- *ruminement* (néologisme de Zola pour désigner le monologue intérieur de Gervaise)

Dans *Germinal*: hapax techniques miniers

Zola forge ou emprunte des termes spécialisés pour décrire le monde de la mine, absents du vocabulaire bourgeois littéraire standard:

- 1- *haveur* (mineur qui abat le charbon à la hache)
- 2- *accrochage* (niveau d'arrêt des cages dans le puits)
- 3- *briquet* (repas du mineur, sens argotique spécifique)

Dans *La Terre*: hapax ruraux

La densité de vocabulaire agricole et paysan crée des hapax liés aux pratiques rurales:

- 1- Termes d'outils agraires spécifiques à la Beauce
- 2- Dénominations de techniques de labour et de moisson

Dans *Son Excellence Eugène Rougon*

Un exemple noté par Brunet: le terme *vespétro* (boisson alcoolisée rare) apparaît sous forme d'hapax.

Comme déjà signalé en amont les hapax zoliens obéissent à une double logique

- 1- **Hapax référentiels**: nécessité documentaire (vocabulaire technique de la mine, de l'agriculture, du commerce) Zola enquête sur le terrain et ramène des termes spécialisés



2- Hapax expressifs: invention stylistique purenéologismes forgés pour saisir des sensations ou états affectifs que la langue ordinaire ne peut nommer (paroxysmes de violence dans *La Bête humaine*, déchéance dans *Nana*)

Ce paradoxe est au cœur de la poétique zolienne: malgré un lexique globalement moins varié que celui de Proust, Zola atteint une originalité lexicale maximale dans les milieux populaires et extrêmes, là où son écriture sort de ses routines pour inventer des formes nouvelles.

6- Dimension psychologique de l'écriture zolienne:

6-1- Mémorisation des formes et habitus scriptural:

La notion d'*habitus scriptural*, empruntée à la sociologie de Pierre Bourdieu (1980), permet de rendre compte des régularités stylistiques observées chez Zola comme le produit d'une expérience langagière accumulée, de pratiques d'écriture répétées qui ont fini par constituer une «seconde nature» créatrice. Dans cette perspective, le style n'est plus seulement un «choix» ponctuel de l'auteur, mais le résultat d'un système de dispositions incorporées qui orientent les décisions linguistiques sans passer par la conscience réflexive. Les récurrences lexicales dégagées par l'analyse textométrique ne sont dès lors plus de simples faits de surface: elles deviennent les traces d'une histoire individuelle d'apprentissage de la langue, sédimentée au fil des romans dans des routines d'écriture relativement stables.

Les modèles psycholinguistiques de la production langagière proposés par Levelt (1989) permettent de préciser ce que l'on entend par mémorisation des formes. Selon ces modèles, la sélection lexicale répond à un processus largement automatisé, où des représentations mémorisées sont activées en fonction de paramètres tels que la fréquence et la récence. Plus une forme (lexème, tournure syntaxique, schéma phrastique) a été employée, plus son seuil d'activation est bas, et plus elle a de chances d'être réutilisée. Appliquée à la production littéraire, cette perspective suggère que les récurrences textométriques observées chez Zola qu'il s'agisse de lexèmes, de syntagmes ou de constructions entières, reflètent en partie les «patterns d'activation» d'une mémoire scripturale: l'auteur puise, souvent sans en avoir pleinement conscience, dans un répertoire de formes qui se sont stabilisées comme solutions privilégiées.

C'est précisément ici que la textométrie apporte un gain décisif: elle offre les moyens de cartographier empiriquement cet habitus scriptural. Les profils de fréquence établis sur l'ensemble des *Rougon-Macquart* montrent par exemple que certains noyaux lexicaux — ceux du corps, de la foule, de la chaleur, de l'enfouissement, reviennent avec une constance remarquable d'un roman à l'autre, indépendamment des milieux représentés. De même, les réseaux de cooccurrences, visualisés par l'Analyse Factorielle des Correspondances, mettent en évidence des constellations stables (par exemple l'association récurrente de termes de chaleur, de profondeur et d'organicité dans *Germinal*) qui témoignent d'un imaginaire lexical structuré, bien plus qu'ils ne reflètent les seules contraintes de la situation narrative.

Les segments répétés, enfin, constituent sans doute la matérialisation la plus tangible de cette mémoire d'écriture. L'analyse séquentielle menée sur le cycle a montré que Zola réemploie des séquences entières (allant du syntagme à la phrase complète) à plusieurs centaines de pages de distance, parfois d'un roman à l'autre. Dans *Germinal*, la métaphore du Voreux comme «bête» ou «gueule» dévorante resurgit sous des variantes minimales, fonctionnant comme un véritable bloc phraséologique mémorisé. Dans *L'Assommoir*, le syntagme péjoratif de type «ce/cet/cette + injure + de + nom propre» («ce sacré surnois de Lantier», «cet animal de Boche», etc.) forme un moule syntaxique extrêmement productif, que l'on peut interpréter comme une unité de style préfabriquée réactivée à chaque fois que le récit requiert un jugement populaire sur un personnage. La forte fréquence de ces patrons, leur distribution régulière et leur capacité à traverser les frontières entre discours direct, style indirect libre et narration montrent que l'on a affaire à des solutions d'écriture stabilisées, et non à de simples coïncidences locales.

On peut ainsi proposer une définition opératoire de l'*habitus scriptural* zolien: un système de routines lexicales et syntaxiques, historiquement constitué par la pratique d'écriture et



cognitivement stabilisé, que l'analyse textométrique met au jour sous forme de récurrences, de cooccurrences et de segments répétés, et qui assure la cohérence profonde du style au-delà des variations de sujets, de milieux et de genres. Autrement dit, ce qui fait qu'*L'Assommoir*, *Germinal* ou *Nana* peuvent sembler, par leurs univers, radicalement différents et pourtant immédiatement reconnaissables comme des textes de Zola, c'est précisément cette mémoire formelle mesurable: la textométrie en fournit la carte, la psycholinguistique en propose le mécanisme, et la notion bourdieusienne d'habitus en donne le statut théorique.

6-2- Intentionnalité auctoriale et psychocritique:

La psychocritique, fondée par Charles Mauron (1963), se donne pour tâche de dégager «des réseaux d'associations d'images obsédantes, et probablement involontaires, qui se répètent» d'un texte à l'autre et qui constituent le «mythe personnel» de l'écrivain (Mauron, 1963: p. 9). Dans le cadre du corpus zolien, les outils d'Hyperbase offrent précisément la possibilité de repérer ces réseaux obsédants à partir de données quantitatives, en objectivant ce que Mauron élaborait par superposition qualitative des textes (réseaux d'images liées au sang, à la souillure, à la fécondité, à l'enfouissement. La figure d'Émile Zola se dessine ainsi à travers un profil d'écrivain singulier dont les choix lexicaux révèlent une cohérence interne et une subjectivité créatrice profonde.

6-3- Profil psycholinguistique de l'auteur:

La construction d'un profil psycholinguistique de l'auteur à partir des données textométriques prolonge les deux dimensions précédentes: l'habitus scriptural (6.1) et l'intentionnalité inférée (6.2). Il ne s'agit plus seulement de repérer des régularités lexicales ou métaphoriques, mais de les rapporter à des fonctionnements cognitifs stables décrits par la psycholinguistique, afin de caractériser la manière dont Zola traite et organise le langage dans l'acte d'écriture. Un tel profil ne prétend pas saisir la «psychologie intime» de l'écrivain, mais la logique de ses opérations verbales telle qu'elle se laisse reconstruire à partir des traces statistiques du corpus.

Les travaux de Galbraith (1999) sur l'écriture comme processus de constitution de connaissances montrent que rédiger implique des boucles de planification, de mise en texte et de révision. Dans le cas de Zola, la convergence entre dossiers préparatoires, programme théorique (*Le Roman expérimental*) et résultats textométriques suggère un fonctionnement fortement marqué par la systématisation: l'auteur organise son matériau en matrices récurrentes (scènes de foule, descriptions de corps, dispositifs de regard) qui servent de cadres générateurs. La forte présence de segments répétés dans les *Rougon-Macquart* traduit cette tendance à réécrire selon des schèmes déjà éprouvés, parce que l'écriture y trouve des formes stables d'orientation de la pensée.

Sur le plan lexical, le profil qui se dessine est celui d'un écrivain dont le vocabulaire privilégie des noyaux concrets, sensoriels et situés plutôt que les abstractions conceptuelles. Le paradoxe relevé par Brunet: un lexique globalement moins «riche» que celui de Proust, mais d'une densité référentielle exceptionnelle, prend ici une portée psycholinguistique: Zola apparaît comme un auteur dont le système d'activation favorise les liens courts entre mots et perceptions (corps, objets, lieux, matières). La prééminence des substantifs concrets, des adjectifs descriptifs et des verbes de perception confirme une pensée par images et scènes plutôt que par notions, où la conceptualisation passe par la saturation sensorielle du texte.

Cette sensibilité perceptive s'accompagne d'une forte tendance à la structuration sérielle et oppositive. Les analyses de cooccurrences montrent que le lexique se cristallise en réseaux binaires (pureté/souillure, lumière/ombre, chaleur/froid, haut/bas) qui organisent la description en systèmes de contrastes. La poétique zolienne des «séries» (de boutiques, de corps, de gestes, de machines) correspond ainsi à un fonctionnement cognitif qui privilégie les structures tabulaires et les progressions graduées, ce que la statistique met en évidence sous forme d'alignements lexicaux récurrents.



Enfin, le croisement de ces tendances avec les résultats psychocritiques fait apparaître une inclinaison à la dramatisation affective: les zones de forte concentration d'hapax et les pics de cooccurrences autour de nœuds lexicaux (corps souffrant, foule en crise, catastrophe mécanique) marquent les points où le système cognitif habituel de Zola est poussé à la limite. Les segments où les hapax se condensent (descriptions paroxystiques, crises de violence, bascules dans la déchéance) signalent des moments où l'auteur sort de ses routines d'activation pour inventer des formes nouvelles à la hauteur d'expériences extrêmes.

En synthèse, Zola se caractérise par une écriture à la fois hautement routinisée et puissamment inventive: un habitus scriptural très structuré (routines descriptives, matrices de scènes, motifs lexicaux) coexiste avec des zones de tension maximale (hapax, métaphores paroxystiques) qui forcent le système à se réinventer. La répétition n'y est jamais pure redondance, mais espace de variation interne: d'un roman à l'autre, les mêmes schèmes d'activation (corporels, sensoriels, sériels) se réajustent aux contraintes du milieu représenté. La textométrie documente ce mode de fonctionnement, la psycholinguistique le modélise et la psychocritique en éclaire les enjeux imaginaires, donnant consistance à l'idée d'un profil d'auteur qui n'est ni une simple projection biographique ni l'addition de tics de style, mais la forme que prend, dans la langue, une manière singulière de penser et de percevoir le monde.

7- Perspectives méthodologiques: vers une textométrie augmentée

7-1- Le Traitement Automatique du Langage comme prolongement:

Les perspectives dégagées à partir de l'analyse textométrique du corpus zolien convergent vers une intégration des méthodes du Traitement Automatique du Langage (TAL). Les techniques récente (analyse de sentiments, modélisation thématique (*topic modeling*), représentations vectorielles des mots (*word embeddings*)) ouvrent des possibilités inédites pour l'analyse des grands corpus littéraires (Moretti, 2013). En particulier, les représentations vectorielles de type Word2Vec ou BERT permettent de saisir les relations sémantiques entre les mots avec une finesse que la statistique lexicale classique ne peut atteindre. Appliquées au corpus zolien, elles permettraient de cartographier les espaces sémantiques propres à chaque roman et d'identifier les transformations de sens que subissent les mots clés en traversant les différents volumes du cycle.

7-2- La textométrie comparée et la poétique cognitive:

La textométrie comparée offre un autre horizon prometteur. En mettant en relation les données statistiques extraites des *Rougon-Macquart* avec celles d'autres corpus contemporain (l'œuvre de Flaubert, de Maupassant ou de Huysmans) il serait possible de situer plus précisément la singularité stylistique de Zola dans le paysage littéraire du XIXe siècle (Becker, 1993). L'articulation des résultats textométriques à une approche psychocritique et cognitive ouvre par ailleurs la voie à une lecture renouvelée de la poétique zolienne, en interrogeant la dimension mémorielle, intentionnelle et affective de la création littéraire. La psychologie cognitive de la création (Galbraith, 1999) apporte une dimension complémentaire en s'intéressant aux processus mentaux qui sous-tendent la production textuelle: la planification, la mise en texte et la révision, dont les traces sont lisibles dans les structures syntaxiques privilégiées et les patterns de cohérence textuelle.

Conclusion:

La recherche conduite sur le corpus zolien à l'aide du logiciel Hyperbase illustre de manière exemplaire les potentialités de l'approche textométrique appliquée aux grands corpus littéraires, en révélant la trame profonde d'une écriture dont la cohérence excède de loin la simple addition des romans. En articulant la rigueur de l'analyse quantitative à la profondeur de l'interprétation psychocritique et cognitive, elle dégage une image renouvelée de la poétique zolienne: non plus seulement comme monument esthétique ou document historique, mais comme la trace dynamique d'un sujet créateur dont les choix lexicaux, les réseaux d'images et les récurrences obsessionnelles dessinent un véritable habitus scriptural.



L'écriture zolienne apparaît ainsi comme un système de contraintes et de préférences, où la répétition des formes, la persistance des motifs (corps, foule, mine, machine, dégénérescence) et l'invention ponctuelle des hapax participent d'une même logique de figuration du réel et de l'imaginaire. Le paradoxe lexical mis en évidence par Brunet (un vocabulaire apparemment pauvre, mais d'une densité sémantico-référentielle redoutable) rejoint les analyses énonciatives du discours indirect libre et les stratégies de segments répétés pour dessiner un style qui fonde sa puissance non sur l'étendue de son vocabulaire, mais sur la densité de ses réseaux lexicaux et la virtuosité de ses dispositifs narratifs. Car si l'hapax n'est pas un mot trouvé par hasard, il est forgé ou choisi délibérément, en réponse à une exigence expressive que le vocabulaire ordinaire ne saurait satisfaire, son caractère ponctuel est tout aussi décisif: cette invention lexicale ne surgit pas de manière systématique, mais en un point précis du texte, là où les routines scripturales de Zola atteignent leur limite, dans un moment de tension narrative maximale (scène paroxystique, état de crise, chute dans la déchéance). L'hapax marque ainsi l'exact endroit où l'écrivain, contraint de sortir de ses habitudes, crée une forme inédite utilisée une seule fois avant de retourner à ses structures d'activation privilégiées. C'est précisément cette coexistence du répété et de l'unique (de l'écrivain profondément routinier et de l'inventeur ponctuellement débordé par son objet, que ce soit dans la violence de *La Bête humaine*, la dégradation de *L'Assommoir* ou l'effondrement de *Germinal*) qui définit l'originalité profonde de la poétique zolienne. Ces pistes permettront de suivre, au plus près, les déplacements de l'imaginaire verbal de Zola, les inflexions de sa voix narrative, les recompositions de ses structures descriptives, en restituant l'ampleur d'une écriture à la fois systématique et débordante: cette démesure rigoureusement construite qui fait la modernité et la vitalité toujours actuelle des *Rougon-Macquart*.

Bibliographie:

- 1- Bakhtine, M. (1978). *Esthétique et théorie du roman* (D. Olivier, trad.). Edition Gallimard. (Œuvre originale publiée en 1935)
- 2- Bally, C. (1909). *Traité de stylistique française*. Edition Klincksieck.
- 3- Barthes, R. (1968). La mort de l'auteur. Edition *Manteia*, 5, 12–17.
- 4- Becker, C. (1993). *Lire le réalisme et le naturalisme*. Edition Dunod.
- 5- Bourdieu, P. (1980). *Le sens pratique*. Edition Minuit.
- 6- Brunet, É. (1985). *Le vocabulaire de Zola*. Edition Slatkine-Champion.
- 7- Brunet, É. (1988). *Le vocabulaire de Proust: concordances, index et relevés statistiques* (3 vol.). Edition Slatkine-Champion.
- 8- Brunet, É. (2004). *Hyperbase, version 6.0: logiciel hypertextuel pour le traitement documentaire et statistique de corpus*. CNRS – Université de Nice.
- 9- Charnois, T., Delente, É., & Legallois, D. (2013). Analyse automatique des segments répétés dans les *Rougon-Macquart*. *Corpus*, 12. <https://doi.org/10.4000/corpus.2312>
- 10- Ferrer, D. (2011). *Logiques du brouillon: modèles pour une critique génétique*. Seuil.
- 11- Filippone, P. (2025) *Les mots du corps dans l'univers fictionnel de Zola: une approche textométrique*.p1.Thèse.fr.
- 12- Galbraith, D. (1999). Writing as a knowledge-constituting process. Dans M. Torrance & D. Galbraith (Éds.), (1999) *Knowing what to write: Conceptual processes in text production* (pp. 139–160). Amsterdam University Press.
- 13- Genesis (2024). Génétique du discours indirect libre dans *L'Assommoir*. *Genesis – Revue internationale de critique génétique*, 58, 1–12.
- 14- Hay, L. (1989). *La naissance du texte*. Corti.
- 15- Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.
- 16- Kozak, A. (2018). *Dictionnaire des hapax dans la poésie archaïque, d'Homère à Eschyle* (Thèse de doctorat, Université Lumière Lyon 2). HAL theses. <https://theses.hal.science/tel-02082531v2>
- 17- Lebart, L., & Salem, A. (1994). *Statistique textuelle*. Edition Dunod.
- 18- Léonard, M. (1974). *Le style indirect libre dans L'Assommoir de Zola*. Presses universitaires de France.
- 19- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.



- 20- Marosvári, M. (1990). La langue populaire dans *L'Assommoir*. *Études zoliennes*, 12, 1–18.
- 21- Mauron, C. (1963). *Des métaphores obsédantes au mythe personnel: introduction à la psychocritique*. Edition Corti.
- 22- Mitterand, H. (1986). *Zola: l'histoire et la fiction*. Presses universitaires de France.
- 23- Mitterand, H. (1990). Poétique inconsciente et poétique explicite chez Zola. *Romantisme*, 70, 45–58. Presses universitaires de France.
- 24- Mitterand, H. (2002). *Zola* (3 vol.). Edition Fayard.
- 25- Moretti, F. (2013). *Distant reading*. Edition Verso.
- 26- Pincemin, B. (2012). Sémantique interprétative et textométrie. *Texte ! Textes et Cultures*, 17(3), 1–21.
- 27- Vanni, L. (2024). Hyperbase Web. (Hyper)Bases, Corpus, Langage. *Corpus*, 25. <https://doi.org/10.4000/corpus.8770>
- 28- Zola, É. (1971). *Le Roman expérimental*. Garnier-Flammarion. (Œuvre originale publiée en 1880)
- 29- Zola, É. (1871–1893). *Les Rougon-Macquart: Histoire naturelle et sociale d'une famille sous le Second Empire* (20 vol.). Edition Charpentier.

Note de fin d'article:

- 1- CDH: Classification Descendante Hiérarchique
- 2- AFC: l'Analyse Factorielle des Correspondances
- 3- TLF: Trésor de la langue française

