Corpus romanesque et lexicométrie

Le vocabulaire de Maupassant

Ouafae BENZINA

Faculté des Lettres et des Sciences Humaines Université Moulay Ismail-Meknès, Maroc

La présente contribution s'inscrit dans le domaine de la lexicométrie et a pour objet l'étude du lexique romanesque de l'auteur normand Guy de Maupassant en appliquant des instruments d'investigation scientifiques, à savoir, l'informatique et la statistique. L'analyse statistique assistée par ordinateur, ou la lexicométrie permet de lire les textes autrement. Charles Muller (1979) affirme que :

> La statistique lexicale, dans la majorité de ses applications, est orientée vers l'analyse stylistique des œuvres littéraires ; elle cherche les différences plutôt que les faits constants, la variété plutôt que l'homogénéité. Mais un écart ne se mesure que par rapport à une moyenne ou une norme. Si la stabilité d'un lexique et la répartition aléatoire d'un vocabulaire ne se réalisent guère dans les textes, ce sont cependant les seuls fondements d'une théorie linguistique du lexique.

Parmi les outils informatiques qui traitent des œuvres littéraires, nous avons opté pour le logiciel Hyperbase (version 7.1 pour Windows), un logiciel de statistique lexicale, élaboré par Etienne Brunet en 1989 et régulièrement mis à jour (la version 10 est disponible depuis 2018). Le traitement statistique des données avec le logiciel Hyperbase permet d'étudier la structure du vocabulaire d'un corpus, étude qui s'intéresse en général à la distribution des fréquences, à l'étude des hapax, à la richesse lexicale, ainsi qu'à l'accroissement lexical. En plus de l'étude de la structure d'un corpus, Hyperbase permet de déduire des conclusions également au niveau du contenu lexical en étudiant la distance lexicale, le vocabulaire spécifique ainsi que l'évolution d'un même auteur au cours de la période durant laquelle il a produit son œuvre.

À l'aide de ce logiciel, nous avons constitué une base que nous avons nommée maupassa.exe, à partir de six romans de l'auteur normand : *Une Vie*, *Bel-Ami*, *Mont-Oriol*,

Pierre et Jean, Fort comme la mort et *Notre cœur*. Le corpus a été constitué avec les textes mis gracieusement sur Internet par Thierry Selva[1].

Dans cette contribution, nous nous intéresserons à l'étude du vocabulaire spécifique des romans de Maupassant, en prenant comme référence le corpus *Frantext*. Dans Hyperbase, les données du TLF sont insérées en tant que norme et servent de base de calcul en indiquant la différence entre notre corpus et celle de *Frantext*. L'analyse de ce vocabulaire en champs sémantiques a pour but de mettre en lumière quels thèmes sont traités dans le corpus et ceux qui en sont absents. Le calcul du coefficient de corrélation chronologique permet d'analyser l'évolution du vocabulaire. Par la suite, nous verrons dans quelle mesure les résultats quantitatifs obtenus sont en adéquation avec les analyses stylistiques classiques sur l'évolution de l'écriture romanesque de l'auteur normand. Cependant, les résultats auxquels nous accéderons sont des données brutes qui ne prennent de signification qu'une fois interprétées.

1 Vocabulaire spécifique

Le vocabulaire spécifique est déterminé avec le logiciel Hyperbase, en relevant le vocabulaire excédentaire et le vocabulaire déficitaire dans le corpus, en calculant l'écart réduit de chaque forme dans chacune des parties du corpus. En effet, si le vocable en question est significativement sur-employé par notre auteur, nous pourrions en déduire que l'écrivain éprouve une certaine attirance pour ce vocable. Il s'agit donc de repérer quels sont les mots préférés de Maupassant. Dans son étude du vocabulaire de Proust, Étienne Brunet explique cette méthode comme suit :

> Les mêmes méthodes que nous avons utilisées pour mettre en relief les structures syntaxiques ou rythmiques propres à Proust peuvent servir à décanter les spécificités proustiennes dans la masse des mots lexicaux ou sémantiques. On exploite là encore les vertus de l'écart réduit, et la norme de référence est comme précédemment constituée soit par l'ensemble du corpus du T.L.F, soit par un sous-ensemble de ce corpus. *(Brunet, 1983 : p. 138)*

Notre étude s'intéresse au vocabulaire spécifique, positif et négatif, de Maupassant. Par vocabulaire positif/négatif, nous entendons les vocables particulièrement sur-employés et sous-employés qui nous donnent une idée des thèmes traités ou négligés. D'après Charles Muller :

> On peut ainsi, pour un fragment d'un texte, calculer (par les moyens mécaniques de préférence, vu le nombre des opérations) l'écart réduit. On obtiendra ainsi, en tête de liste, le *vocabulaire caractéristique positif* du fragment, l'ensemble des vocables dont la sous-fréquence est plus élevée que la fréquence dans le texte ne le fait prévoir ; en fin de liste, le *vocabulaire caractéristique négatif* du fragment. *(Muller, 1968 : p. 204)*

Dans ce sens, la lecture des listes des mots présentées par le logiciel Hyperbase pour chaque texte, nous amène à constituer les champs associatifs qui nous permettront d'interpréter ce vocabulaire.

[1] [http://maupassant.free.fr/], consulté le 20 août 2021.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

1.1 Vocabulaire positif

Il représente le vocabulaire excédentaire qui, après une comparaison avec le corpus de TLF, a un écart réduit positif.

Les résultats obtenus par l'application de la fonction *Spécificités* du logiciel sont présentés dans les listes suivantes (réduites aux vingt premiers rangs) :

Tab. 1 : Vocabulaire spécifique positif

Écart	**TLF**	**Texte**	**Mot**
- 37.67	1540412	4007	Que
-34.20	1252134	3239	Je
-31.04	986675	2506	Est
-29.20	477329	854	Me
-28.36	504761	984	A
-21.35	433468	1066	Nous
-20.97	309477	650	Mon
-19.56	132026	132	Mes
-18.59	218211	427	Ma
-18.54	336060	837	Y
-18.04	159577	254	Sont
-15.39	4074626	16913	De
-15.03	109395	172	Ont
-14.42	75675	83	Nos
-14.05	818177	2955	Vous
-13.53	977554	3654	Qui
-12.61	248484	732	Ai
-12.39	741472	2736	Pas
-11.59	52437	64	Avais
-11.49	57290	80	Vos

À partir de la liste du vocabulaire spécifique, nous avons classé les mots selon leur appartenance catégorielle ; ainsi, nous avons pu remarquer que les mots de liaisons sont majoritaires (95.341 occurrences), que les substantifs apparaissent en deuxième position (28.665 occurrences) ; les verbes en troisième (22430 occurrences) et les adverbes en quatrième position (9432 occurrences). Viennent ensuite, les adjectifs avec 8331 occurrences et enfin, les noms propres avec 3195 occurrences. Le graphique suivant illustre cette répartition :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

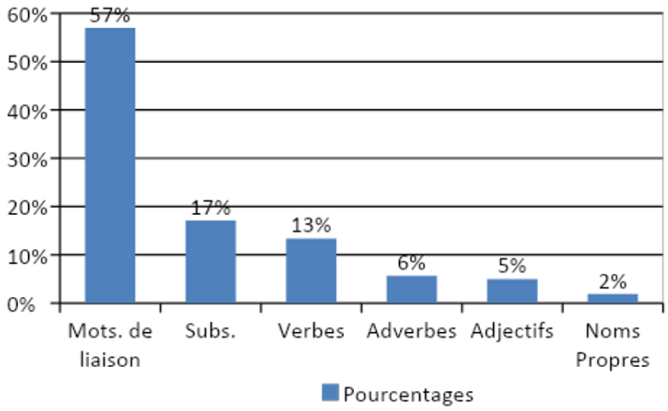


FIG. 1 : Les pourcentages des appartenances catégorielles du vocabulaire spécifique

Étienne Brunet (1981) avait déjà remarqué que la littérature du XIX^{e} siècle se caractérise par l'abondance des substantifs. La richesse du substantif dans l'œuvre romanesque de Guy de Maupassant confirme donc les observations de Brunet.

Nous remarquons que les mots qui se trouvent en tête de la liste du vocabulaire positif sont, évidemment, ceux des personnages des romans de Maupassant : *Christiane (M-O), Gontran,* frère de Christian *(M-O), Forestier (B-A), Walter (B-A), Bertin (F), Jeanne (V), Rolan (P-J), Annette (F), Rosalie (V), Julien (V), Paul (V), Georges (B-A), Madeleine (B-A), Roy (B-A), Olivier (F), Rival (B-A), Louise (P-J).* Viennent ensuite les verbes de parole (*murmurer, balbutier, demander, répondre, déclarer…*), les verbes de sentiment (*sentir, sembler, aimer, penser,…*), les verbes de perception (*apercevoir, regarder, écouter, voir…*), les verbes d'action (*s'asseoir, aller, prendre, chercher, venir, rentrer, marcher…*), les noms de sentiments (*sensation, désir, émotion, tendresse, angoisse, chagrin, affection, cœur…*), les noms de métier (*docteur, peintre, médecin, artiste*), du genre féminin (*femme, fille, mère*), des parties du corps (*chair, lèvre, jambe, figure, main, joue, gorge, yeux, doigt, cheveux, pied…*). Dans cette liste, nous trouvons aussi des substantifs appartenant au registre de la nature (*pierre, arbre, mer, eau*).

Le registre de l'argent est illustré en bas de liste avec les mots : *argent, riche, fortune, pièce.*

Examinons maintenant les listes du vocabulaire négatif du corpus.

1.2 Vocabulaire négatif

Par vocabulaire négatif ou déficitaire, nous désignons les formes qui n'apparaissent pas beaucoup dans le corpus et dont la présence par rapport au TLF est insignifiante. Ces formes ont un écart réduit négatif.

La liste ci-dessous (réduite aux vingt premiers rangs)regroupe le vocabulaire déficitaire du corpus Maupassant par rapport à la base *Frantext :*

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Tab. 2 : Vocabulaire négatif

Écart	TLF	Texte	Mot
- 37.67	1540412	4007	Que
-34.20	1252134	3239	Je
-31.04	986675	2506	Est
-29.20	477329	854	Me
-28.36	504761	984	A
-21.35	433468	1066	Nous
-20.97	309477	650	Mon
-19.56	132026	132	Mes
-18.59	218211	427	Ma
-18.54	336060	837	Y
-18.04	159577	254	Sont
-15.39	4074626	16913	De
-15.03	109395	172	Ont
-14.42	75675	83	Nos
-14.05	818177	2955	Vous
-13.53	977554	3654	Qui
-12.61	248484	732	Ai
-12.39	741472	2736	Pas
-11.59	52437	64	Avais
-9.87	75875	169	Dieu

Nous remarquons, d'après ces listes, la rareté de la conjonction *que* qui se trouve en tête de la liste du vocabulaire déficitaire, avec un écart négatif de -37,67. Ce déficit de la conjonction *que* signale la rareté des phrases complexes dans le corpus par opposition à la dominance des phrases simples.

Les verbes *être* et *avoir* avec plusieurs de leurs formes fléchies sont également déficitaires (*est, sont, a, ont…*). Ce n'est pas un constat surprenant, d'autres études ayant remarqué que « le verbe *avoir* et le verbe *être* appartiennent au vocabulaire significatif négatif des romans, au point même que ce seul critère suffit à désigner un texte comme roman. » (Brunet, 2009 : p. 270).

Le mot *Dieu* apparaît au début de cette liste du vocabulaire déficitaire, avec un écart réduit négatif de -9,87. Dans son étude du vocabulaire religieux, Brunet a remarqué que : « le roman est très déficitaire, en ce qui concerne le vocabulaire religieux » (2009 : p. 337). En analysant la distribution du terme *Dieu,* dans ses trois formes

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

(*dieu*, *dieux, Dieux*) à travers les genres littéraires, Brunet conclut également que : « *la divinité refuse d'apparaître dans le roman* » (2009, 340).

Après cette étude du vocabulaire positif et négatif, nous examinerons l'évolution thématique dans l'œuvre romanesque de Guy de Maupassant. Nous préciserons donc quels termes y sont en déclin et quels autres en progrès.

2 Évolution du vocabulaire

L'évolution du vocabulaire interroge l'ensemble du vocabulaire et mesure l'évolution de chaque terme, par le coefficient de corrélation de Bravais-Person. Avec ce calcul réalisé par le logiciel Hyperbase, nous saurons quels sont les mots que Guy de Maupassant emploie de plus en plus et ceux qu'il abandonne progressivement dans son œuvre romanesque.

La formule du coefficient de corrélation de Bravais-Pearson est :

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$$

Fig. 2 : Calcul de r

Le calcul se déroule, d'après Étienne Brunet, de la façon suivante :

> Un programme de tri ordonne alors tous les mots traités selon la valeur du coefficient, de -1 à 0,50. La première liste correspond aux mots dont l'emploi tend à baisser, la seconde rend compte de ceux qui progressent. Et dans les deux listes apparaissent en tête les mots dont le mouvement (à la baisse et à la hausse) est le plus accentué. L'arrêt à la valeur 0,50 du coefficient s'explique par le souci de maintenir le seuil de 0,05. *(Brunet, 1981 : p. 686).*

Ce coefficient examine une éventuelle relation entre une distribution et la chronologie. Il évolue entre deux limites : +1 (pour une progression) et -1 (pour une régression), comme l'indique Brunet : « la valeur du coefficient est négative ou positive selon que le mot est en régression ou en progression. » (1988 : p. 391).

Le calcul de corrélation nous permet de suivre, globalement, la trajectoire des mots dans les différents textes du corpus. Le tri des résultats permet de constituer la liste (décroissante) des mots en progression et la liste (croissante) des mots en régression.

Observons maintenant les listes qui montrent le vocabulaire progressivement moins employé (à gauche du tableau) et celui qui est de plus en plus recherché par l'auteur (à droite du tableau). Il convient de rappeler que les textes qui constituent notre corpus s'échelonnent dans le temps et sont classés par ordre chronologique.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Tab. 3 : Évolution chronologique du vocabulaire

Mots en régression			**Mots en progression**		
Coeff.	**Fréq.**	**Mot**	**Coeff.**	**Fréq.**	**Mot**
- 0.968	77	grosse	+ 0.975	422	car
- 0.956	164	brusquement	+ 0.965	21	donnent
- 0.953	19	avança	+ 0.964	56	art
- 0.945	31	levait	+ 0.957	321	cet
- 0.932	42	suivit	+ 0.957	15	émouvoir
- 0.931	1464	puis	+ 0.951	21	ressemblaient
- 0.925	2105	au	+ 0.948	18	poétique
- 0.925	20	bourse	+ 0.947	66	goût
- 0.922	427	bras	+ 0.943	30	devinait
- 0.916	15	baissa	+ 0.932	15	tromper
- 0.915	72	immobile	+ 0.930	2313	avait
- 0.907	36	élança	+ 0.921	2805	ce
- 0.905	459	tête	+ 0.919	18	redoutable
- 0.901	92	blanche	+ 0.918	18	animée
- 0.899	64	nouvelles	+ 0.909	41	artistes
- 0.896	98	balbutia	+ 0.908	56	aimé
- 0.896	20	reprenait	+ 0.906	3654	qui
- 0.895	32	furieux	+ 0.906	939	peu
- 0.892	29	chérie	+ 0.905	21	contemplait
- 0.891	105	laisser	+ 0.902	975	même

Dans la liste du vocabulaire en régression, nous relevons :

- le lexique des parties du corps : *bras, dos, face, lèvre, front, pieds, mains, nez, moustaches.*
- le lexique de l'argent : *francs, argent, bourse, affaire, dépenser, économie, dette.*
- les chiffres : *vingt, cent, cinq, deux, six.*
- les verbes d'action : *avancer, suivre, élancer, baisser, laisser, descendre, passer, aller, rouler, revenir, jeter, fermer, boire, rester, rire, trouver, approcher, saisir, entrer, arriver, donner, emplir, courir, relever, commencer, accomplir, renverser.*
- les verbes de parole : *demander, balbutier, appeler, dire*

La liste des mots en progression regroupe :

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

- les pronoms personnels : *je, il, elle, vous.*
- le lexique de l'art : *art, artiste, poétique, célèbre, œuvre,*
- les mots de la séduction et de la beauté : *coquet, séduction, séduire, coquetterie, plaire, admirer, attirer, beauté, jolie, agréable, exquise, flatter.*
- le vocabulaire des sentiments et des préoccupations morales et spirituelles : *Dieu, émouvoir, aimer, soulager, seul, émotion, énervement, amoureux, souffrir, crises, tendresse, liaison, amitié, jaloux, plaisir, mélancolique, vanité, nerfs, envie, besoin, ardeur, âme, sincère, sens, plaire, ému, impression, intolérable, penser, espérer, spirituel, tourmenter, jalousie, remords, dégoûter, cœur, ennuyer, malaise, désirs, profondément, douter, remords.*

Nous remarquons qu'il y a un renouvellement lexical chez Maupassant. L'auteur passe du lexique des parties du corps, des verbes d'action et de l'argent (termes concrets) au vocabulaire des sentiments et des préoccupations morales (termes abstraits). Alain Quesnel note que :

> Les romans de Maupassant, s'ils offrent une unité stylistique certaine, témoignent d'une évolution surprenante pour qui les lirait d'affilée, évolution qui tient certes à la variété des thèmes abordés, mais plus encore à des choix d'écriture.
>
> *(Quesnel, 1999 : p. 11).*

Le graphique ci-dessous illustre la distribution du terme *francs* dans notre corpus :

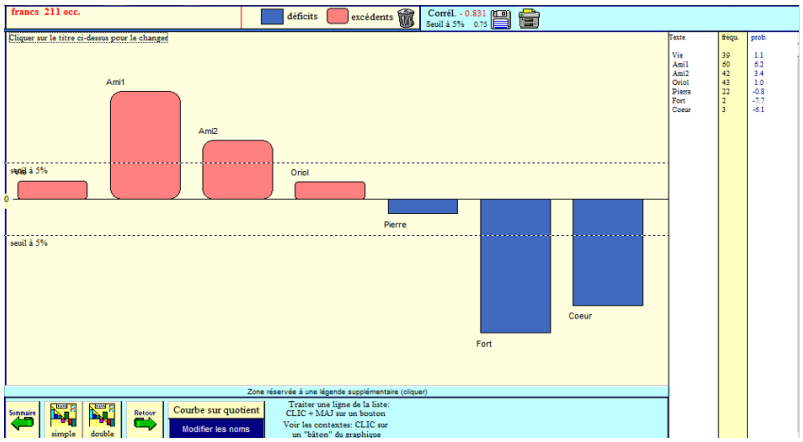


Fig. 3 : Distribution du mot « francs » dans l'œuvre romanesque de Maupassant

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

On voit que le thème de l'argent est dominant dans les quatre premiers romans de Maupassant, et disparaît complètement dans ses deux derniers romans, *Fort comme la mort* et *Notre cœur*. À partir de *Pierre et Jean*, le romancier observateur devient romancier psychologue. C'est ce qu'affirme André Vial : « Le romancier d'analyse, le romancier psychologue naissait donc, dans *Pierre et Jean*, du romancier observateur et du romancier de mœurs. » (Vial, 1954 : p. 404).

En supprimant tout problème matériel à ses personnages et en les faisant évoluer dans des milieux favorisés, Maupassant estime mieux s'attarder sur leur personnalité. Dans *Fort comme la mort*, à une seule exception près, on ne trouvera aucune somme d'argent numériquement définie ; il en va de même dans *Notre cœur* : « En somme, en écrivant *Fort comme la mort*, Maupassant se serait rendu à une méthode dans le vent, abandonnant sa chère 'objectivité' au profit de la formule 'psychologique' : c'est-à-dire, plus précisément, la psychologie à la Bourget. » (Harris, 1991 : p. 13).

Ainsi, dans ses derniers romans, Maupassant entre dans l'intimité de l'être. Il explore les arrière-pensées des héros, leurs sentiments secrets ou inconscients, d'où la progression du vocabulaire des sentiments à la fin de son œuvre romanesque.

3 Conclusion

L'étude des spécificités du vocabulaire de Guy de Maupassant dans son œuvre romanesque nous a permis de dégager le lexique spécifique de l'auteur par rapport au corpus du *Trésor de la langue française*. Nous avons donc remarqué que la liste positive est ancrée dans le monde des personnages, de leurs actions, de leurs sentiments et de leurs corps. La liste négative confirme les conclusions tirées par E. Brunet sur le vocabulaire romanesque, à savoir : le déficit des mots grammaticaux, des verbes *être* et *avoir*, et du vocabulaire de la religion.

L'observation de l'évolution du vocabulaire de Maupassant telle qu'elle est présentée par le logiciel Hyperbase, à la lumière du coefficient de corrélation chronologique de Bravais-Pearson, sous forme de deux listes de vocabulaires significatifs, progressifs et régressifs, révèle que l'auteur tend à faire usage du lexique des sentiments vers la fin de son œuvre, à adopter le lexique de l'art, de la beauté, des sentiments et des soucis psychologiques et à abandonner le lexique de l'argent, des parties du corps, des actions. Cette évolution est attestée par les listes fournies par le logiciel ; Maupassant passe d'un extrême à l'autre : il abandonne le lexique concret pour se tourner vers le vocabulaire abstrait. L'analyse lexicométrique confirme ainsi les études stylistiques classiques sur l'œuvre romanesque de Maupassant et notamment ce changement, au cœur de l'œuvre romanesque de l'auteur normand. En effet, dans un intervalle de temps très réduit (1883-1890), Maupassant change de registre et s'intéresse aux nuances psychologiques de ses personnages en choisissant des milieux riches et oisifs. Comme eux, l'écrivain, qui a pu quitter les bureaux du ministère où il perdait sa vie, est devenu indépendant, riche et il fréquente les salons.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352