Analyse stylométrique du style de Gustave Flaubert à partir de trigrammes de caractères

Duhamel Océane - Etudiante Bachelor Erasmus

Semestre d'automne 2025-2026, Université de Genève

1 Introduction

Ce travail s'inscrit dans le cadre d'une initiation aux méthodes de la stylométrie et des humanités numériques. La stylométrie vise à mettre en évidence des régularités dans l'écriture à partir de mesures statistiques, afin de compléter l'analyse littéraire traditionnelle.

Le choix de Gustave Flaubert comme objet d'étude s'explique par l'importance accordée par l'écrivain à la forme de la phrase. Flaubert est notamment connu pour sa pratique du *gueuloir* qui consistait à lire ses phrases à voix haute afin d'en tester la sonorité, le rythme et l'équilibre. Cette attention portée à la phrase invite à s'interroger sur la possibilité de repérer, à l'aide de la stylométrie, des traces de ce travail stylistique.

L'objectif de cette étude est donc de déterminer si le style de Flaubert peut être mis en évidence par une analyse stylométrique, et plus précisément par une approche fondée sur les trigrammes de caractères. Contrairement aux analyses basées sur les mots, cette méthode permet de s'intéresser à la matérialité de la langue (les enchaînements de lettres, la présence des espaces et les structures syntaxiques récurrentes). Il s'agit ainsi de se rapprocher d'une analyse du rythme et de la construction de la phrase, en lien avec l'idée du *gueuloir*.

Les œuvres de Flaubert sont comparées à celles de plusieurs auteurs du XIX^{e} siècle, afin d'évaluer si ses textes présentent une cohérence stylistique identifiable et distincte.

2 Constitution du corpus

Le corpus a été constitué dans une perspective comparative. Il comprend plusieurs œuvres de Gustave Flaubert dont : *Madame Bovary*, *Un cœur simple*, *Hérodias* et *La Légende de saint Julien l'Hospitalier*. Ces textes ont été choisis car ils représentent différents moments et différents genres de l'œuvre de Flaubert (roman, conte, prose narrative), ce qui permet d'observer si son style demeure stable malgré ces variations.

Afin de situer le style de Flaubert par rapport à celui d'autres auteurs, le corpus inclut également des œuvres de Guy de Maupassant, d'Octave Mirbeau, d'Hector Malot, d'Alphonse Allais et de Jules-Amédée Barbey d'Aurevilly. Ces auteurs ont été retenus pour leur proximité chronologique et leur style, mais aussi pour la diversité de leurs pratiques d'écriture. Le corpus permet ainsi de comparer des styles narratifs proches, tout en testant la capacité de l'analyse

à distinguer des signatures formelles spécifiques et a voir si le style de Flaubert se retrouve dans celui d'autres auteurs de son époque.

Chaque œuvre est représentée par un fichier texte distinct au format `.txt`. Une phase de préparation du corpus a été nécessaire avant toute analyse, les fichiers non littéraires (fichiers de configuration, tables de résultats générées automatiquement par les logiciels, fichiers temporaires) ont été exclus, car leur présence biaise les calculs stylométriques. Cette étape de nettoyage a nécessité plusieurs vérifications manuelles afin de s'assurer que seuls les textes littéraires étaient pris en compte, il a fallu de plus nettoyer certains textes qui comportaient des caractères non reconnu par le logiciel.

3 Méthodologie

Les trigrammes ont été extraits à l'aide du package `quanteda`. Les textes ont d'abord été découpés en caractères, puis transformés en trigrammes. Une matrice de fréquences a ensuite été construite, indiquant la fréquence de chaque trigramme dans chaque texte, seuls les 3000 trigrammes les plus fréquents ont été conservés. Ce choix permet d'éviter que des formes très rares, peu représentatives du style global, n'influencent excessivement les résultats.

La mise en œuvre de cette méthodologie a donné lieu à plusieurs difficultés techniques. Le logiciel `stylo`, initialement envisagé pour l'analyse en trigrammes sur R, s'est révélé difficile à utiliser dans ce contexte précis. Plusieurs problèmes ont été rencontrés : inclusion involontaire de fichiers non littéraires dans le corpus, erreurs liées aux chemins d'accès et aux droits de lecture des fichiers, ainsi que des difficultés dans la génération automatique des graphiques. Lors de la mise en œuvre de l'analyse avec `stylo`, une erreur récurrente a été rencontrée lors de la génération des graphiques, signalée par le message *« objet `distance.name.on.graph` introuvable »*. Cette erreur semble liée à un dysfonctionnement lors de l'étape de visualisation, et non au calcul des distances ou à la préparation du corpus, malgré plusieurs ajustements (paramétrage manuel, modification du répertoire de travail, désactivation de l'interface graphique) le problème persistait. Afin de contourner cette difficulté, une solution alternative a été trouvée reposant sur l'extraction des traits stylistiques avec `quanteda` et le calcul des distances avec le package `proxy`. À partir de cette matrice de distances, un clustering hiérarchique a été réalisé puis visualisé sous forme de dendrogramme.

4 Résultats

Les résultats de l'analyse sont présentés sous la forme d'un dendrogramme hiérarchique qui représente les relations de proximité stylistique entre les textes du corpus, calculées à partir de la distribution des trigrammes de caractères.

5 Interprétation

L'observation du dendrogramme montre que les textes ont tendance à se regrouper par auteur, ce qui indique une cohérence stylistique propre à chacun. Les œuvres de Gustave Flaubert apparaissent ainsi proches les unes des autres, malgré des différences de genre ou

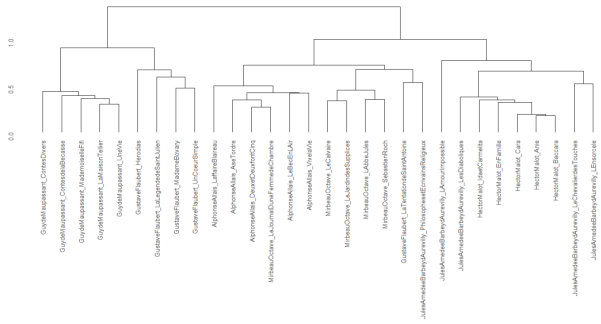

FIGURE 1 – Résultat trigramme

de sujet. Ce résultat suggère que l'analyse met en évidence des traits communs liés à la construction de la phrase et au rythme de l'écriture et peut être mis en relation avec la pratique du *gueuloir*, qui témoigne du travail de Flaubert sur la forme et la musicalité de la prose, indépendamment du contenu des textes.

La proximité entre certains textes de Maupassant et ceux de Flaubert peut s'expliquer par le fait qu'ils appartiennent à la même période et qu'ils partagent certains traits d'écriture, cette proximité n'empêche toutefois pas de distinguer leur style respectif. L'analyse montre ainsi qu'il est possible d'identifier des styles propres à chaque auteur, même à partir d'éléments très simples comme les enchaînements de caractères, tout en faisant apparaître des rapprochements entre certains écrivains.

6 Conclusion

Cette étude a montré que l'analyse stylométrique fondée sur des trigrammes de caractères constitue un outil pertinent pour l'étude du style littéraire, appliquée à l'œuvre de Gustave Flaubert, elle met en évidence une cohérence stylistique mesurable compatible avec l'idée d'un travail sur le rythme et la forme de la phrase.

Les difficultés techniques rencontrées au cours de l'analyse ont également permis de mieux comprendre les enjeux méthodologiques de la stylométrie, notamment l'importance de la préparation du corpus et du choix des outils. En se concentrant sur la forme de la langue plutôt que sur le vocabulaire, cette méthode apporte un point de vue quantitatif qui complète l'analyse littéraire traditionnelle et permet d'observer le style sous un autre angle.