

Qui a écrit Molière ?

CHIKHANI Charles ZAPFACK MESSIAN Jasen Steve

Encadrant : M. Hervé Fournier

Université Paris Cité

02 Juin 2023

Table des matières

1	Introduction	2
1.1	Présentation du sujet : la remise en question de l'attribution des pièces de Molière à l'auteur lui-même	2
1.2	Importance de l'analyse textuelle et statistique dans la résolution de cette question . .	2
1.3	Hypothèses et problématique	2
2	Compréhension de l'article de Florian Cafiero et Jean-Baptiste Camps	3
2.1	Les différents corpus utilisés	3
2.2	Les caractéristiques de l'étude	3
2.3	Choix de la caractéristique	3
2.4	Algortihme de <i>Clustering</i>	4
2.5	Métrique de l'algorithme	4
2.6	Dendogramme	4
3	Notre expérience	6
3.1	Description de l'expérience que nous avons menée	6
3.2	Choix des outils d'analyse textuelle et statistique utilisés	6
3.3	Méthodologie mise en place	6
3.4	Collecte et pré-traitement des données	13
4	Conclusion	13

1 Introduction

1.1 Présentation du sujet : la remise en question de l'attribution des pièces de Molière à l'auteur lui-même

Depuis plusieurs décennies, une question persiste dans le domaine de la littérature : l'attribution des pièces de Molière à l'auteur lui-même est-elle remise en question ? Cette controverse a été initiée par Pierre Louÿs, un romancier du XXe siècle, qui a suggéré que Pierre Corneille aurait pu être l'auteur véritable des pièces de Molière. Cependant, cette théorie repose sur des fondements fragiles et ne bénéficie d'aucune preuve concrète. Malgré cela, la rumeur persiste, alimentée par des éléments tels que l'élosion tardive de Molière en tant qu'auteur, son prétendu manque d'éducation et de culture, ainsi que l'absence de preuves manuscrites permettant de réfuter directement cette hypothèse.

1.2 Importance de l'analyse textuelle et statistique dans la résolution de cette question

Au début des années 2000, Cyril et Dominique Labbé, deux chercheurs, ont avancé l'idée selon laquelle Corneille aurait écrit pour Molière. Leur méthode consiste à mesurer une "distance inter-textuelle" qui évalue la différence de lexique entre les textes des deux auteurs. Si cette distance ne dépasse pas un certain seuil, les deux pièces sont considérées comme écrites par le même auteur. Ils se basent également sur le fait que de nombreux dramaturges de l'époque signaient leurs œuvres sous le nom de "comédien poète", ce qui permettait aux véritables auteurs de rester anonymes tout en bénéficiant de la promotion et de la représentation de leurs pièces par les acteurs.

Cependant, cette méthodologie a été contestée par d'autres chercheurs. Certains ont souligné que l'implémentation de la méthode de Cyril et Dominique Labbé pourrait "lisser artificiellement les différences entre les auteurs", en utilisant une distance euclidienne qui accorde trop de poids aux lemmes fréquents, réduisant ainsi la disparité entre les fréquences observées de différentes formes.

D'autres approches ont été proposées pour résoudre le problème de l'attribution des comédies de Molière. Certaines méthodes utilisent une analyse textuelle et statistique pour comparer les styles d'écriture, tandis que d'autres adoptent des approches plus qualitatives en examinant les intrigues, la versification et les sujets choisis dans les pièces.

1.3 Hypothèses et problématique

Dans ce rapport, nous examinerons deux hypothèses qui remettent en question la paternité des œuvres de Molière. La première hypothèse suggère que Molière aurait fourni des brouillons à Pierre Corneille, qui aurait ensuite versifié les pièces, peut-être avec l'aide de son frère. Selon cette hypothèse, Molière aurait créé les intrigues, mais la versification aurait été réalisée par Pierre Corneille, sans recevoir un crédit explicite. La deuxième hypothèse soutient que Molière n'aurait ni écrit les intrigues ni les vers de ses pièces, et qu'il n'aurait été qu'un nom célèbre utilisé pour promouvoir les pièces et dissimuler le véritable auteur.

Pour résoudre cette controverse, différentes méthodes ont été utilisées, chacune avec ses propres avantages et limitations. L'objectif de cette étude est d'évaluer ces approches et de fournir une analyse critique des résultats obtenus.

Pour ce faire, dans la section suivante, nous présenterons en détail l'article de Florian Cafiero et Jean-Baptiste Camps, qui examine cette question en utilisant une analyse textuelle et statistique. Nous aborderons les différentes perspectives et critiques soulevées par d'autres chercheurs, ainsi que les différentes approches méthodologiques qui ont été utilisées pour résoudre le problème de l'attribution des pièces de Molière.

Le rapport qui suit explorera en profondeur ces différentes approches et tentera de faire la lumière sur cette controverse persistante, fournissant ainsi une contribution significative à notre compréhension de l'œuvre de Molière et de son véritable auteur.

2 Compréhension de l'article de Florian Cafiero et Jean-Baptiste Camps

2.1 Les différents corpus utilisés

Il y a trois ensembles d'œuvres que nous appellerons des corpus. Un corpus désigne une collection importante et structurée de textes ou de documents utilisée pour l'analyse linguistique. Ces corpus peuvent généralement comprendre une variété de textes tels que des livres, des articles :

- Le premier corpus, appelé "corpus exploratoire", est constitué d'un large échantillon de comédies en vers. Cet échantillon comprend des pièces d'au moins 5000 mots pour les auteurs ayant écrit au moins trois comédies. Il inclut des pièces de théâtre de 12 auteurs.

- Le deuxième corpus, appelé "corpus final", est construit pour obtenir un résultat plus lisible et moins biaisé. Pour éviter les biais liés aux sous-genres, les chercheurs vont exclure les comédies héroïques et les courtes farces comiques. Afin d'éliminer le bruit ajouté par de nombreux phénomènes qui ne sont pas liés aux hypothèses présentées précédemment, ils choisissent de se concentrer uniquement sur cinq auteurs majeurs de l'époque. Ce corpus final comprend 37 pièces de T. et P. Corneille, Molière, Rotrou et Scarron.

- Le troisième corpus sert de test pour vérifier la précision de leur approche. Il est constitué de comédies en vers écrites après la mort de P. Corneille et Molière.

2.2 Les caractéristiques de l'étude

Sur chacun de ces corpus, les chercheurs ont appliqué les caractéristiques suivantes :

- **Lexicon** : un lexique désigne l'ensemble des mots et des unités lexicales d'une langue, incluant leurs sens, leurs formes grammaticales et leurs relations. Par exemple, "maison", "chien", "arbre".

- **Rhyme Lexicon** : fait référence à un lexique spécifique aux rimes. Il s'agit d'une liste qui répertorie les mots et les expressions en fonction de leurs sonorités et de leurs similarités phonétiques, notamment la dernière syllabe ou les sons finaux des mots. Par exemple, "rat", "chat", "chapeau", "bateau", "plateau" - tous les mots se terminent par le son "-o" ou "-au".

- **Affixes** : les affixes désignent les éléments qui peuvent être ajoutés aux mots pour en modifier le sens ou la fonction. Ils comprennent les préfixes (ajoutés au début du mot), les suffixes (ajoutés à la fin du mot) et les infixes (ajoutés à l'intérieur du mot).

- **Morphosyntactic sequences** : il s'agit de l'analyse des séquences morphosyntaxiques, c'est-à-dire l'étude des combinaisons de morphèmes et de structures grammaticales dans une phrase.

- **Mots fonctionnels / mots-outils** : ce sont des mots grammaticaux qui ont principalement un rôle syntaxique ou grammatical dans une phrase, plutôt qu'un sens lexical spécifique. Les mots fonctionnels comprennent souvent les prépositions, les conjonctions, les pronoms, les déterminants, les adverbes de liaison et les particules grammaticales. Exemples : "de", "à", "dans", "sur", "sous", etc.

2.3 Choix de la caractéristique

La sélection d'une caractéristique fiable et informative est une étape cruciale dans la réalisation d'une étude statistique. En effet, les caractéristiques ont pour but d'améliorer la fiabilité des analyses. La sélection d'une caractéristique s'effectue en fonction de la taille du corpus, le niveau de confiance et la marge d'erreur potentiel.

Grâce à cette formule, la taille minimale de l'échantillon notée n a été calculée en utilisant la formule :

$$n = p(1 - p)(z/e)^2$$

où p représente la probabilité moyenne de la caractéristique dans notre corpus, z le niveau de confiance et e la marge d'erreur de l'estimation de probabilité.

Dans l'étude, z a été fixé de manière à obtenir un niveau de confiance qui est supérieur à 90 et $e = 2s$ où s est l'écart-type de la caractéristique dans le corpus.

Toutes caractéristiques sélectionnées doivent suivre une distribution gaussienne.

2.4 Algortihme de *Clustering*

Dans le cadre d'une approche statistique pour l'attribution d'auteurs par le biais de l'apprentissage automatique, l'algorithme de clusterisation hiérarchique a été appliqué à chacun des corpus. Cet algorithme est un type spécifique d'algorithme de regroupement utilisé dans le domaine de l'apprentissage automatique et de l'analyse de données. Il s'agit d'une approche ascendante où chaque point de données est initialement considéré comme un cluster séparé, puis fusionné de manière itérative en fonction de leur similarité.

Un cluster, dans le contexte de l'analyse de regroupement, fait référence à un groupe de points de données partageant des similitudes ou présentant des schémas lorsqu'ils sont comparés à d'autres points de données. Ces points de données sont regroupés en fonction de certains critères tels que la proximité dans l'espace des caractéristiques ou la similarité dans les valeurs des attributs. Les clusters sont formés en se basant sur des mesures de similarité ou de dissimilarité utilisées dans l'algorithme de regroupement.

2.5 Métrique de l'algorithme

Le choix de la mesure de distance et du critère de liaison (liaison complète, simple ou moyenne) détermine la manière dont la similarité entre les clusters est évaluée lors du processus de fusion.

La métrique de distance utilisée dans cet algorithme permet de quantifier la similarité ou la dissimilarité entre les points de données ou les clusters. Elle détermine comment l'algorithme mesure la distance ou la dissimilarité entre les observations afin de former des clusters. Parmi les exemples de métriques couramment utilisées, on retrouve la *distance euclidienne*, la *distance de Manhattan* et la *distance cosinus*.

Dans l'étude, les chercheurs ont utilisé la distance de *Burrow's delta* et le *min-max*. Nous allons étudier la distance de Burrow. Cette distance calcule la distance de Manhattan entre les *z-scores* des fréquences de ces caractéristiques dans les textes de deux auteurs. Le *z-scores* est une notion statistique qui quantifie le nombre d'écart-types par lequel une observation ou un point de données s'éloigne de la moyenne d'une distribution. La distance de Borrow mesure donc la dissimilarité entre le style d'écriture des deux textes comparés, en prenant en compte les différences dans les fréquences normalisées de ces caractéristiques.

La formule de calcul de la distance de Burrow :

$$\delta(A, B) = \sum_{i=1}^n \left| \frac{(A_i - \bar{B}_i)}{\sigma_i} \right|$$

avec A_i et B_i des fréquences de mots dans le texte et σ_i la variance de l'utilisation du mot.

La métrique d'union (*linkage* en anglais) est une méthode utilisée dans le **regroupement hiérarchique** pour calculer la distance entre les clusters lors du processus de fusion. Elle est utilisée pour déterminer comment les clusters sont regroupés pour former des clusters plus grands. Différentes méthodes de linkage, telles que le ward linkage, le single linkage, etc., peuvent être utilisées.

La formule de calcul de distance entre les clusters utilisée dans la recherche se base sur la métrique d'union. Prenons l'exemple de deux clusters C_1 et C_2 , avec leurs centroides respectifs G_1 et G_2 , et les nombres d'individus dans les clusters n_1 et n_2 . La distance d entre les clusters, à minimiser, est définie par l'équation suivante :

$$d^2(C_1, C_2) = ((n_1 * n_2) / (n_1 + n_2)) * d^2(G_1, G_2)$$

L'objectif du regroupement hiérarchique est de minimiser la distance entre les clusters lors de la fusion, ce qui peut être réalisé en choisissant la méthode de linkage appropriée et en ajustant les paramètres en conséquence.

2.6 Dendrogramme

Après l'application de l'algorithme de *clustering* sur chacun des corpus. On a comme résultat un dendrogramme pour chacunes des caractéristiques :

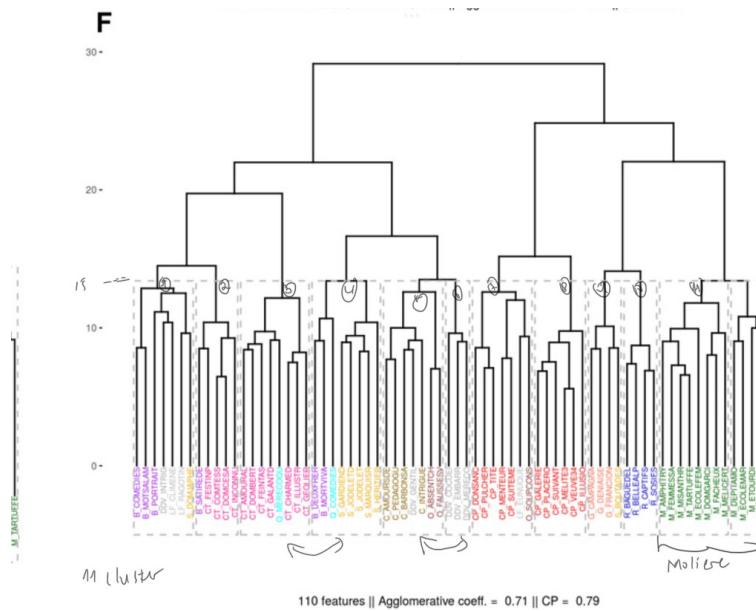


FIGURE 1 – Dendrogramme mots fonctionnels

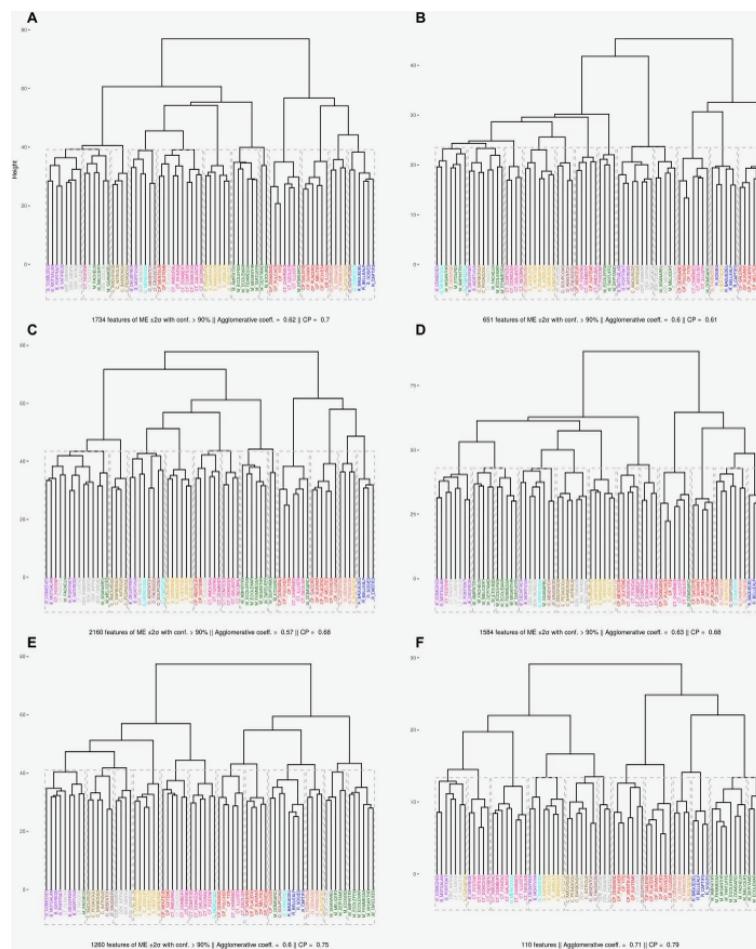


FIGURE 2 – Dendrogramme de toutes les caractéristiques

De cette interprétation, il en ressort une réfutation des hypothèses énoncées précédemment. En effet, comme indiqué sur la photo, on peut observer une distinction claire du cluster attribué à Molière par rapport aux autres auteurs. Cette distinction remet en question les hypothèses selon lesquelles Molière aurait pu emprunter les intrigues ou les vers d'autres auteurs, ou encore que son nom aurait été utilisé comme une simple marque de promotion dissimulant les véritables auteurs. Les résultats obtenus grâce à l'analyse statistique et à l'algorithme de clusterisation hiérarchique confirment ainsi l'unicité du style d'écriture de Molière.

3 Notre expérience

3.1 Description de l'expérience que nous avons menée

Dans le cadre de cette étude, nous avons mené une expérience visant à déterminer la paternité des textes de Molière. L'objectif principal était de développer une méthodologie pour identifier les caractéristiques distinctives du style d'écriture de Molière et les comparer à ceux d'autres auteurs de la même époque. Le second objectif était de mieux comprendre les enjeux du choix des caractéristiques. Pour cela, nous avons utilisé des techniques d'analyse textuelle et statistique pour extraire des informations à partir des textes et les utiliser comme base pour la classification.

3.2 Choix des outils d'analyse textuelle et statistique utilisés

Pour mener notre analyse, nous avons choisi plusieurs outils d'analyse textuelle et statistique. Nous avons opté pour la bibliothèque **NLTK** (*Natural Language Toolkit*) de Python pour ses fonctionnalités de prétraitement de texte, telles que la suppression des *stopwords*, la normalisation des mots et la lemmatisation. Cette bibliothèque nous a permis de nettoyer les données textuelles et de réduire les informations redondantes ou inutiles.

Nous utilisons également le modèle pré-entraîné **FastText** de Facebook pour la vectorisation des mots. En ce qui concerne l'analyse statistique, nous avons utilisé des techniques telles que l'analyse de fréquence des mots, l'analyse des *n-grammes* et l'analyse de similarité. Pour ces tâches, nous avons utilisé des bibliothèques *Python* telles que *scikit-learn*, qui offre des fonctionnalités avancées pour l'analyse de texte et la classification et la classification telle que la clusterisation hiérarchique et les K-means.

3.3 Méthodologie mise en place

Notre expérience s'est divisée en deux axes. Le premier axe consiste à déterminer comprendre et étudier le style d'écriture de Molière et de Corneille. Le second axe consiste à déterminer les clusters de textes qui se ressemblent le plus.

Grâce à la bibliothèque **NLTK** et **WordCloud**, nous avons pu faire un nuage de mot pour le corpus de Molière. Un nuage de mots est une manière de visualiser la fréquence des mots dans un corpus. Les mots les plus fréquents sont présentés sous forme de nuage, où la taille du mot est proportionnelle à sa fréquence. Cette représentation peut donner une vue d'ensemble rapide des termes les plus courants dans le corpus de Molière. Pour que l'analyse soit plus pertinente nous avons supprimé la prise en compte des noms propres, qui sont uniques aux œuvres.

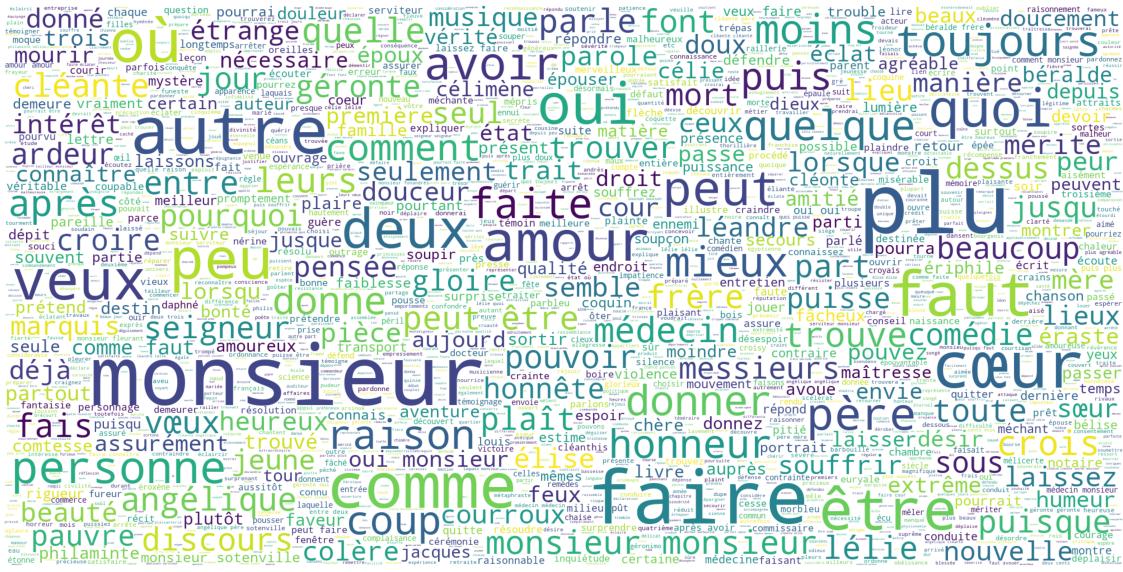


FIGURE 3 – Nuage de mots du corpus de Molière

Grâce au nuage de mots, nous avions remarqué que le mot *Monsieur* revenait très souvent. Cette grande importance du mot *Monsieur* dans le nuage de mots révèle le vocabulaire et les règles d'écriture de l'époque.

Afin de mieux comprendre le style d'écriture de Molière, nous avons également réalisé un histogramme des mots les plus fréquents dans le corpus de Molière. Pour rendre ce diagramme plus pertinent, nous avons décidé de ne pas prendre en compte les noms de personnages mais également le mot *Monsieur*.

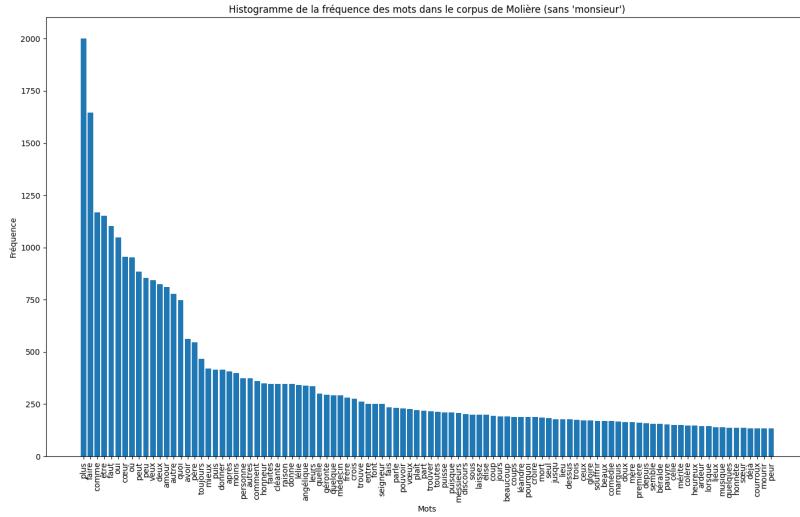


FIGURE 4 – Diagramme de mots du corpus de Molière sans le mot *Monsieur*

En répétant l'expérience sur le corpus de Corneille, nous trouvons des résultats assez similaires.



FIGURE 5 – Nuage de mots du corpus de Corneille

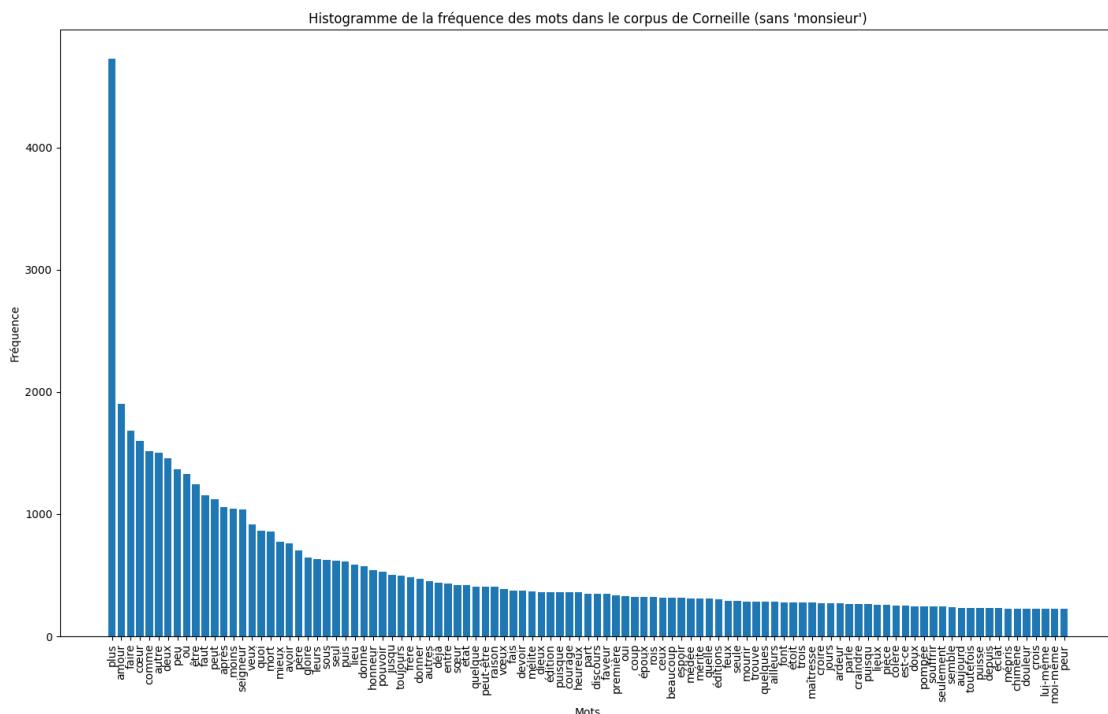


FIGURE 6 – Diagramme de mots du corpus de Corneille

Grâce à l'analyse des nuages de mots, nous constatons une utilisation quasi similaire du mot *plu, amour* ou encore *cœur*. Ces utilisations fréquentes des mots par nos deux auteurs peut rendre difficile l'attribution précise de l'auteur d'une œuvre en se basant uniquement sur la distance euclidienne inter-textuelle, comme celle proposée par D. et C. Labbé. Cela souligne l'importance d'utiliser des métriques et des méthodes d'analyse supplémentaires pour une identification plus précise de l'auteur.

Nous avons ensuite réalisé une analyse de bigrammes et trigrammes. En analysant les paires de mots (bigrammes) ou les groupes de trois mots (trigrammes) qui se produisent souvent ensemble dans les textes de Molière et de Corneille, nous avons pu comprendre davantage sur les expressions idiomatiques, les tournures de phrases et les habitudes stylistiques distinctes des auteurs.

Les résultats se présentent sous la forme d'un fichier texte répertoriant tous les bigrammes différents fréquemment associés ainsi que les groupes de trigrammes qui apparaissent ensemble avec leur fréquence respective.

Une observation intéressante est que le digramme (peut, être) est plus fréquent, avec 235 occurrences, dans les œuvres de Molière, tandis que chez Corneille, c'est le digramme (après, avoir) qui revient le plus souvent, avec 117 occurrences. Nous remarquons aussi peu de différences entre les fréquences de certains digrammes similaires, comme (puis, plus) qui se répète 44 fois chez Corneille et 33 fois chez Molière.

Au niveau des trigrammes, le triplet de mots le plus fréquent chez Molière est (monsieur, oui, monsieur), tandis que chez Corneille, c'est (plus, puis, faire). Nous ne remarquons pas de trigrammes similaires, en revanche le champs lexical est très proche, avec des mots comme monsieur, oui, puis, plus qui reviennent souvent.

Notre second axe d'analyse consiste à déterminer les clusters de textes qui se ressemblent le plus. Après avoir réalisé un prétraitement (où nous reviendrons en détails dans la partie 3.4) de compléxité globale $O(n^2)$, et avoir vectorisé nos textes, nous avons pu réaliser des dendrogrammes. Nous avons appliqué nos algorithmes de clustering, qui comprennent le k-means utilisant la distance euclidienne et la métrique du cosinus, ainsi que l'agglomerative Clustering utilisant la distance de Jaccard. Les mesures d'agrégation des clusters que nous avons utilisées sont "complete" et "average". En ce qui concerne la complexité de nos algorithmes, elle dépend de plusieurs facteurs tels que la taille des données et le nombre de clusters.

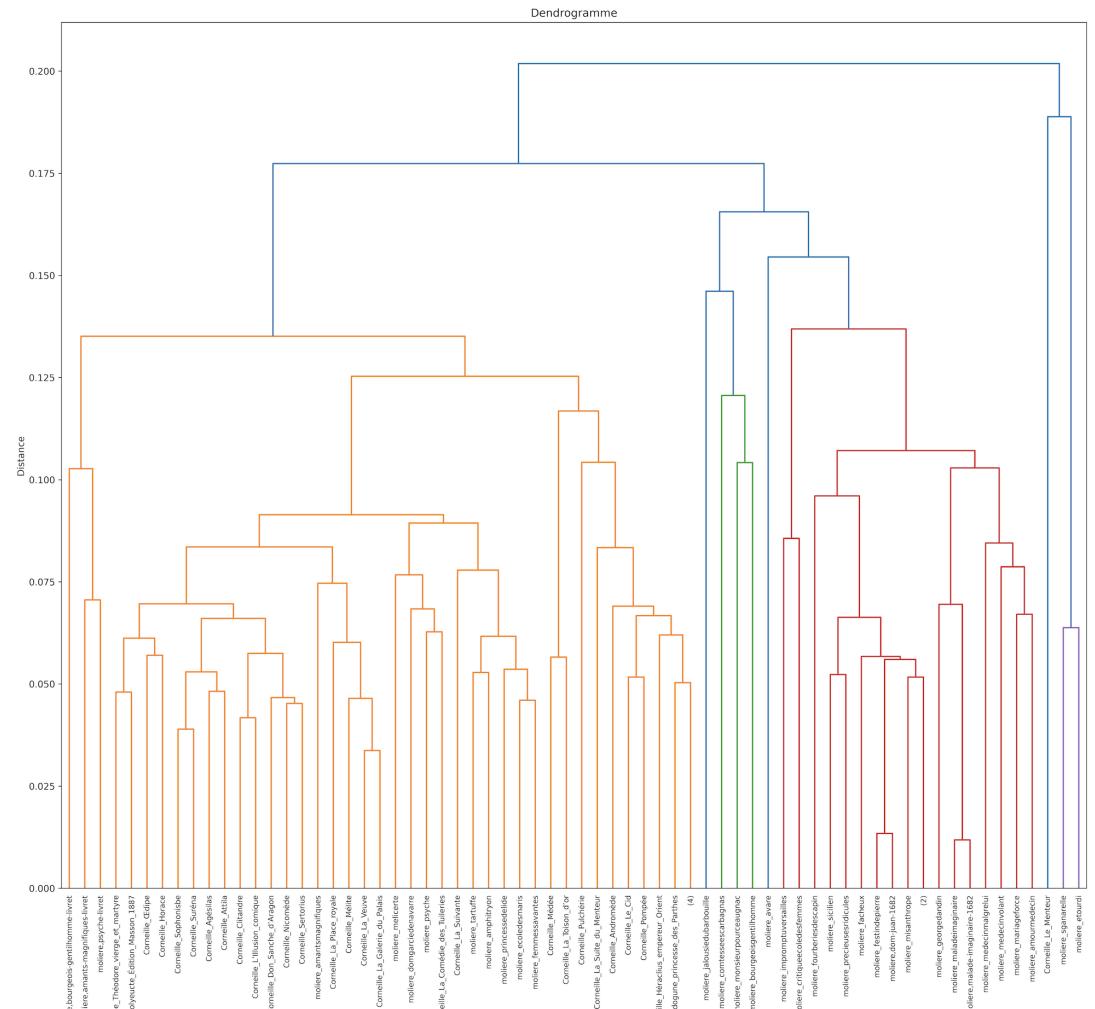


FIGURE 7 – Dendrogramme suivant la distance euclidienne

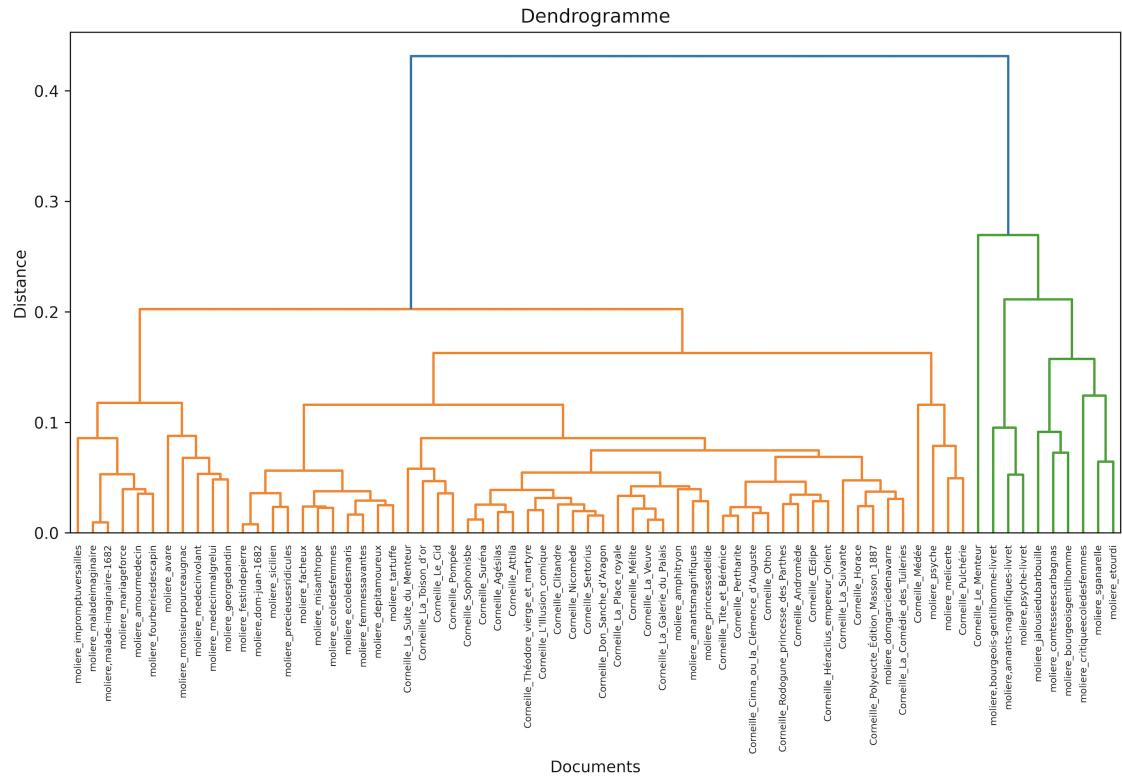


FIGURE 8 – Dendrogramme suivant la distance cosinus

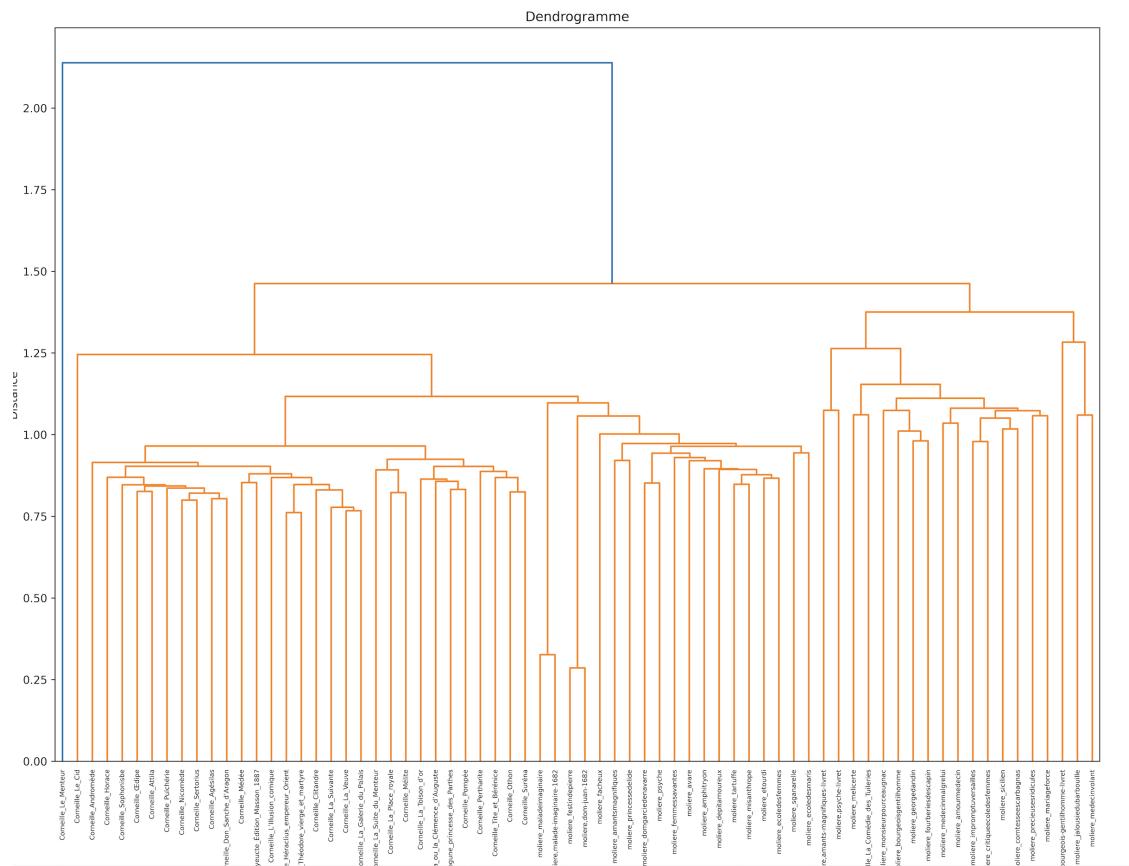


FIGURE 9 – Dendrogramme suivant la similarité de Jaccard

Lors de l'interprétation des résultats sur les dendrogrammes, nous constatons que certaines œuvres ne sont pas bien positionnées lorsque nous utilisons la distance euclidienne et la métrique du cosinus. En revanche, avec la métrique de Jaccard, nous observons des regroupements plus compacts, que ce soit pour les œuvres de Molière ou de Corneille. Cela suggère que la métrique de Jaccard est plus efficace pour capturer les similitudes et les différences entre les textes, grâce à son approche ensembliste qui le différencie des autres métriques. Ce qu'appuie notre étude, en considérons que l'approche et les choix des caractéristiques et des métriques sont plus pertinents dans la méthodologie développée par Mr Caffiero et Mr Camps, est qu'une étude stylométrique sur des textes littéraires si proches nécessite une approche plus fine et plus précise. La difficulté de l'attribution de l'auteur est d'autant plus grande que les auteurs sont proches stylistiquement.

3.4 Collecte et pré-traitement des données

Pour notre expérience, nous avons pris une collection d'œuvres de Molière, ainsi que des œuvres de Corneille. Ces données ont été recueillies à partir de sources disponibles en ligne. Pour Molière nous avons téléchargé des fichiers XML du GitHub de Mr Cafiero de 38 pièces. Pour Corneille nous avons téléchargé des fichiers PDF du site Wikisource pour un corpus composé de 34 pièces.

Le pré-traitement des données utilisées pour l'analyse de texte, c'est à dire, les bigrammes, les trigrammes et les mots les plus fréquents, a été réalisé en convertissant les fichiers PDF et XML en fichiers texte. Chaque œuvre est donc convertit en une liste de mots en supprimant les *mots vides* et la ponctuation. Les mots vides sont des mots qui n'ont pas de signification et ne sont pas pertinents pour l'analyse de texte.

Pour l'analyse de similarité, le pré-traitement des données a été légèrement différent. Comme nous avons fait le choix d'utiliser la caractéristique *Affixe*, nous avons, grâce à une fonction de la bibliothèque NLTK, pu supprimer les suffixes des mots. Les *mots vides* et la ponctuation ont également été retirés.

4 Conclusion

D'après nos différentes expériences, nous pouvons conclure que le style d'écriture de Molière présente une unicité remarquable. Les caractéristiques spécifiques que nous avons observées dans ses œuvres, telles que l'utilisation fréquente de certains bigrammes et trigrammes, ainsi que des motifs récurrents et la fréquence des affixes, nous ont permis d'identifier et de distinguer son style d'écriture par rapport à celui des autres auteurs. Ces résultats renforcent l'idée que Molière possède un style d'écriture unique et reconnaissable.