[Projet MATH/INFO]

|  |  |  |
| --- | --- | --- |
| [Chikhani Charles  Zapfack Messiane jasen steve] | [Date] | Nom de l’enseignant : [HERVÉ Fournier] |

# Compréhension de l’article

[A. Compréhension de l’article 1](#_Toc376433041)

[1. Introduction : (Dans cette partie on parle du problème et de son côté historique et des hypothèses) 1](#_Toc1448912892)

[2.Methodologie & approche. 2](#_Toc901613044)

[Demo ( POUR MIEUX EXPLIQUER LEUR PROCESSUS ) 7](#_Toc318292550)

## 1. Introduction : (Dans cette partie on parle du problème et de son côté historique et des hypothèses)

Dans des articles publiés au XXème, une rumeur selon laquelle Pierre Corneille serait l’auteur des pièces de Molière, rumeur lancée par Pierre Louÿs, romancier né à Gand en Belgique en 1870, de nationalité française. Cette théorie ne repose sur aucune preuve concrète, Pierre Louÿs étant réputé pour être « un homme de supercherie extraordinaire », de plus qu’aucun des contemporains de Molière ne s’est permis de contester la paternité de ses pièces. Mais cette rumeur perdure du fait que Molière a eu une éclosion tardive, avait un prétendu manque d’éducation et de culture, son agenda chargé et qu’il y avait aucune preuve manuscrite qui pouvait réfuter directement la rumeur. Elle s’est éternisée tout le siècle, jusque dans les années 2000, où Cyril et Dominique Labbé, deux chercheurs affirment que Corneille a écrit pour Molière. Pour cela, ils définissent une « une distance intertextuelle » mesurant la différence de lexique entre les textes des auteurs. Ils déterminent ensuite un seuil, et si ce seuil n’est pas dépassé, on considère que les deux pièces sont écrites par le même auteur.

L’argument de Dominique Labbé était qu’à cette époque la plupart des pièces de théâtres était signées sous le nom de « comédien poète ». Environ 90% des comédies et la moitié des tragédies du XVIIe siècle étaient signées par des "comédiens poètes" en tant que prête-noms, selon Dominique Labbé. Cette pratique permettait aux auteurs de rester anonymes tout en permettant aux acteurs de promouvoir et de jouer leurs œuvres.

Mr Bernet, un chercheur, a constaté que la distance intertextuelle séparant Molière et Corneille était la norme de l'époque. En effet, il a trouvé que P. Corneille et son frère Thomas auraient également pu écrire la plupart des pièces du corpus de Mr Bernet. Cependant, les deux chercheurs ont objecté, reprochant à Mr Bernet de ne pas avoir respecté leur méthode.

Mr Schöch affirme que l’implémentation de C. et D Labbé « lisserait artificiellement les différences entre les auteurs ». En effet, dans leur étude, les deux chercheurs utilisent la distance euclidienne, ce qui augmente le poids des lemmes fréquents. Cela engendre la diminution de la disparité entre les fréquences observées de différentes formes.

La méthode de Vonfelt se concentre sur les 1-grammes de caractères pour évaluer la distance intertextuelle, mais elle ne prend pas en compte la sélection ou le traitement des textes. Marusenko et Rodionova ont utilisé une approche probabiliste pour attribuer des pièces de théâtre à des auteurs, en comparant les comédies en vers de Molière avec celles de Quinault et de Corneille, selon cinq critères grammaticaux. Cependant, cette méthodologie soulève des questions sur sa performance empirique, l'inégalité des corpus de comparaison et la logique de la procédure elle-même. Des approches plus qualitatives ont également été proposées pour résoudre le problème de l'attribution des comédies de Molière. La complexité de cette controverse tient au doute généralisé sur les auteurs officiels, ce qui rend impossible l'utilisation de méthodes supervisées pour identifier les styles d'au moins une partie des auteurs de l'époque. La versification et les règles de décence sont rigoureusement codifiées dans le théâtre classique, ce qui ajoute une difficulté supplémentaire à l'attribution des œuvres. Il y a également eu a cette période d’importantes inspirations entre les auteurs.

deux hypothèses sont examinées pour remettre en question la paternité des œuvres de Molière.

## 2.Methodologie & approche. (

Nos chercheurs pour montrer la véracité des 2 Hypothèses qui sont

* La première hypothèse, basée sur les travaux de Wouters et al. (23), suggère que Molière aurait fourni des brouillons à P. 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 P. Corneille (ou son frère), considérée comme une opération technique ne méritant pas de crédit explicite. Si cela était vrai, on pourrait trouver des similitudes de vocabulaire dans les pièces signées par Molière, mais les rimes, les mots de fonction, les affixes et les séquences morphosyntaxiques devraient être les mêmes que dans les autres pièces de P. (ou T.) Corneille
* La deuxième hypothèse, basée sur les théories du poète acteur de D. Labbé, suggère que Molière n'aurait pas écrit ni les intrigues ni les vers de ses pièces, et qu'il n'aurait été qu'un nom célèbre utilisé pour aider à promouvoir la pièce, pour satisfaire l'ego de l'acteur principal/metteur en scène et pour dissimuler le nom de l'auteur réel. Selon cette hypothèse, les sujets choisis dans les pièces de Molière, comme les Précieuses Ridicules, auraient été plus proches des intérêts habituels de P. (ou T.) Corneille et ne refléteraient aucune influence de Molière. Si cela était vrai, tous les indicateurs devraient montrer que le vocabulaire et le style de Molière n'existent pas, et les pièces de Molière devraient être confondues avec celles de P. Corneille selon chacun des six critères évalués dans cette étude.

Font une analyse des études caractéristiques qui sont de soient lexicaux (lemmes, lemmes de rime, formes de mots) ou grammaticaux (affixes, parties du discours, mots fonctionnels).

Pour arriver à une réfutation des deux hypothèses et conclut que ni P. Corneille ni T. Corneille (et incidemment, aucun des auteurs majeurs testés ici) n'auraient pu écrire les pièces signées sous le nom de Molière. Sans prouver définitivement que les œuvres de Molière sont les siennes - ce que seule une preuve historique pourrait faire, et c’est sur cette réfutation que nos chercheurs se base pour affirmer que Molière est bien l’auteur de ces œuvres.

Maintenant nous allons détailler, l’approche de nos chercheurs

Tout d’abord sur les 6 études caractéristiques nous allons parler en profondeur de 2, et vous donner un Example de comment cette caractéristique et traiter pour parvenir à nous donner un résultant qui est de démontrer la paternité d’un ouvre

Nous allons définir certain les caractéristiques

1. Lexicon : est un terme qui se réfère généralement à un vocabulaire spécifique à un domaine particulier ou à une langue donnée. Dans le contexte de la linguistique, un lexicon désigne l'ensemble des mots et des unités lexicales d'une langue, ainsi que leurs sens, leurs formes grammaticales et leurs relations. Il peut également faire référence à un dictionnaire ou à une liste organisée de mots avec leurs définitions. Dans le domaine de la psychologie, le lexicon peut se référer aux connaissances lexicales d'une personne, c'est-à-dire à son vocabulaire et à sa maîtrise des mots dans une langue example maison, chien, arbre. Ce lexicon répertorie des mots en français dans différents domaines sémantiques. Chaque mot représente un concept ou un objet spécifique.
2. Rhyme lexicon :   
   Un "rhyme lexicon" fait référence à un lexique spécifique aux rimes. Il s'agit d'une liste ou d'une base de données qui répertorie les mots et les expressions en fonction de leurs sonorités et de leurs similarités phonétiques, en particulier en ce qui concerne la dernière syllabe ou les sons finaux des mots. Un rhyme lexicon est souvent utilisé dans la poésie, la composition musicale ou d'autres domaines où les rimes jouent un rôle important. Il peut servir à trouver des mots qui riment entre eux, à générer des poèmes rimés ou à analyser les schémas de rimes dans une œuvre littéraire ou musicale. Exemple rat, chat, chapeau, bateau, plateau tous les mots se terminent par le son "-o" ou "-au", ce qui les rend appropriés pour des rimes entre eux.
3. Function Word (mots fonctionnels/mot-outils) : Les "function words" (mots fonctionnels) sont des mots grammaticaux qui ont principalement un rôle syntaxique ou grammatical dans une phrase plutôt qu'un sens lexical spécifique. Ils sont utilisés pour exprimer des relations grammaticales, des connexions logiques ou des fonctions grammaticales. Les mots fonctionnels sont souvent des prépositions, des conjonctions, des pronoms, des déterminants, des adverbes de liaison et des particules grammaticales. Comme les prépositions : de, à, dans, sur, sous

Nous ferons une démo explicative de cette caractéristique

Les mots fonctionnels sont est une caractéristique d’étude, très courant en stylométrie et est la plus efficace dans attribution des œuvres littéraire car elles se suivent dans un schéma plus naturel par les auteurs et varie dépendant du genre du textes et les thèmes d'œuvre

2)Processus :

Ils ont pris un set d’environs 5000 mots pour les auteurs avec au moins 3 comédies et ce set inclus au moins 12 auteurs. Ils sont un premiers résultant de ce set et constate que les ouvres de Molières forment un cluster distinct des autres œuvres, mais ce set contient des nombreuses perturbations causées par de nombreux phénomènes (co-écritures, plagiat, attribution incertaine etc ) et du coup suite à cela ils décident de prendre un set moins diversifier et plus contrôler de œuvres directement comparables à celle de Molières

Et donc cette limite à ces 5 auteurs T et P Corneille, Molière, Rotrou et Scarron, et les résultats sont pareils pour Molières, qui les que toutes les œuvres de Molières sont bien dans un même cluster, ils ont calculé un certain dégrée de pureté des clusters qui était 1 pour chacun qui signifie { YA LE CALCUL DU DEGREE DE PURITE PLUS BAS }

Ils ont aussi constaté que les mots fonctionnels ont donner des résultats qui permettait d'attribuer a 100% chaque auteurs son travail.

Pour revenir à nos 2 hypothèses du départ ,

La première hypothèse affirme que P. Corneille et Molière auraient pu collaborer, Molière fournissant des brouillons que P. Corneille aurait ensuite versifiés. Si tel était le cas, nous devrions observer une ressemblance dans le lexique des pièces signées par Molière, mais une dissemblance dans le vocabulaire des rimes, les affixes, la morphosyntaxe et les mots fonctionnels, qui devraient tous être plus proches de P. Corneille. Pourtant, rien de tel n'est observé dans les pièces signées par Molière.

Dans notre analyse finale cite plus tôt, toutes les pièces signées par Molière appartiennent au même groupe distinct des pièces de P. Corneille, quel que soit le type de caractéristique étudiée. À la lumière de nos résultats, ils considèrent donc cette première hypothèse comme réfutée.

La deuxième hypothèse stipule que les pièces de Molière auraient été entièrement écrites par un ou plusieurs auteur(s) de son époque, Molière étant seulement l'acteur interprétant le rôle principal. Si cela était vrai, des groupes mélangeant des pièces signées par P. Corneille (ou un autre auteur) et Molière devraient apparaître. Pourtant, nos analyses montrent une séparation nette entre toutes les pièces de Molière et celles des autres auteurs étudiés, y compris P. et T. Corneille. Cela confirme l'affirmation selon laquelle toutes les pièces signées par Molière n'ont pas été écrites par P. Corneille, ni par l'un des autres auteurs étudiés ici. Il reste la possibilité qu'elles soient dues à un autre auteur (ou plusieurs auteurs avec des styles très similaires) en dehors du cadre des principaux dramaturges de l'époque testés ici, mais cette hypothèse n'a jamais été avancée et semble peu plausible d'un point de vue historique.

Leur model a quelque limitation telle qu'il est impossible d’inclure les analyses des proses,

Ils ont rétreint leurs analyses des sets juste à des verses car ils voulaient un genre ou il aurait eu un assez grand nombre des œuvres de P. Corneille et de Molliere

Du a des raisons de statistique il ont préférées traiter de sets de données d’au moins 5000mots et exclure les auteurs de moins d'œuvre

Finalement après un processus de sélection il leurs reste 71 œuvres de 12 auteurs, pour les mots fonctionnels ils ont travaillé avec les 250 mots les plus fréquents et on enlever les pronoms, mais ont garder les formes verbales des verbes être et avoir

En stylométrie, la recherche de la caractéristique la plus fiable est très important.

Dans leur projet, pour augmenter la fiabilité de leurs analyses dans un set (corpus) de taille différente, ils ont décidé de se bases sur le niveau de confiance et la marge d’erreur qu’on peut obtenir de la taille minimale de notre set. Il le calcul avec cette formule statisticien,

n=p(1\_p)(z/e)^2 , ou n= la taille minimum du sample,

P = La probabilité moyenne des caractéristiques (fait référence à la probabilité moyenne ou à la moyenne des probabilités associées à une caractéristique particulière dans un contexte ou un ensemble de données donné. Dans le cadre de l'apprentissage automatique ou de l'analyse statistique, les caractéristiques sont des caractéristiques spécifiques ou des variables utilisées pour représenter ou décrire les données. La probabilité moyenne des caractéristiques donne une indication de la probabilité ou de la vraisemblance d'une certaine caractéristique dans le contexte de l'ensemble de données. Elle peut être calculée en faisant la moyenne des probabilités associées à cette caractéristique sur l'ensemble des échantillons ou des instances de données. La probabilité moyenne des caractéristiques peut être utilisée pour évaluer l'importance ou la pertinence d'une caractéristique dans un modèle prédictif ou pour comprendre sa distribution au sein de l'ensemble de données)

Est un estimateur de la probabilité de la caractéristique

z= niveau de confiance

e= marge d’erreur de l’estimateur p

Pour que cette formule soit valide faut que la distribution entre les sample des texte soit une gaussienne si cest n’est pas le cas on applique une autre formule, ensuite on pourra rentre p comme un estimateur non biaisée

Analyse statisticien dans l’attribution d’auteur grâce au machine Learning

Il utilise un unsupervised méthode , qui est Hierarchical Agglomératives Clustering, cet algo a une complexité de O(n^3) avec n étant la taille des matrix de fréquences de mot de notre caractéristiques

Cet algorithme ce base sur des points comme la notion de distance entre 2 texte , et le résultant est représenter par un dendrogramme, le critère d’union entre 2 cluster

La métrique utiliser s’appelle la distance de Burrow delta

(formule)

Qui est une métrique très efficace dans attribution d’auteur

La Ward agrégations est utiliser comme critère d’union entre les clusters

(Formule)

On calcule le la solidité de notre structure de cluster avec un coefficient de clustérisassions

La formule est la suivante

(formule)

Le cluster pureté est le pourcentage du nombre total d’oeuvre qui sont bien classe

(formule )

### Demo ( POUR MIEUX EXPLIQUER LEUR PROCESSUS avec photo)

On va prendre un set de données de plusieurs différents artistes et d’un même genre littéraire, par Example en format csv ou texte. En se basant sur les mots fonctionnels

On applique un algorithme de prétraitement, qui nous permettra d’avoir une matrice des fréquences de mots fonctionnelles

Complexité de O(nmlogm)

Sur cette matrice on applique l’algorithme de clusterisation hierachical

On a une complexité de O(n^2) avec n la taille des matrices de fréquences

**Interprétation du dendrogramme**