Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Le théâtre français du XVIIe siècle : une expérience en catégorisation de textes

Jacques Savoy
Institut d'informatique, Université de Neuchâtel, Suisse
Jacques.Savoy@unine.ch

RÉSUMÉ

La catégorisation de documents (attribution d'un texte à une ou plusieurs catégories prédéfinies) possède de multiples applications. Cette communication se focalise sur l'attribution d'auteur en analysant le style de vingt pièces de théâtre du XVIIe siècle. L'hypothèse que nous souhaitons vérifier admet que le véritable auteur est le nom apparaissant sur la couverture. Afin de vérifier la qualité de deux méthodes d'attribution, nous avons repris deux corpus additionnels basés sur des romans écrits en français et italien. Nous proposons une amélioration de la méthode Delta ainsi qu'une nouvelle grille d'analyse pour cette approche. Ensuite, nous avons appliqué ces approches sur notre collection de comédies. Les résultats démontrent que l'hypothèse de base doit être écartée. De plus, ces œuvres présentent des styles proches rendant toute attribution difficile.).

ABSTRACT

The French Theater of the 17th Century : An Experiment in Text Categorization

The automatic assignment of a text to one or more predefined categories presents multiple applications. In this context, the current study focuses on author attribution in which the true author of a doubtful text must be identified. This analysis emphasis on the style of 20 French comedies written in verse by 9 authors during the 17th century. The hypothesis we want to verify assumes that the real author is the name appearing on the cover (called the signature hypothesis). In order to validate the reliability of two attribution procedures, we used two additional corpora based on 200 extracts of novels written in French with 30 authors and 140 Italian novels authored by 40 persons. After this verification, we propose an improvement of the Delta method as well as a new analysis grid for this model. Finally, we applied these approaches to our French comedy corpus. The results demonstrate that the signature hypothesis must be discarded. Moreover, these works present similar styles making any attribution difficult to support with a high degree of certainty.

MOTS-CLÉS : Classification automatique, humanités numériques, apprentissage automatique, attribution d'auteur. .

KEYWORDS: Text Categorization, digital humanities, machine learning, authorship attribution.

1 Introduction

Depuis plus de 150 ans, l'identité du véritable auteur d'Hamlet (écrit en 1601) ou de Roméo et Juliette (1594) a fait l'objet de quelques 4 000 livres et articles (Michell, 1999). Plusieurs noms ont été proposés comme F. Bacon, C. Marlowe, E. de Vere et pour les plus récents, J. Florio (Tassinari, 2009). Face à ces nombreuses propositions, les Stratfordiens récusent toute autre attribution que celle

de Shakespeare. Ce débat n'est pas clos et toute apparition d'une œuvre pouvant être attribuée à Shakespeare fait renaître la discussion (Kreuz, 2023).

Pour le XVIe ou XVIIe siècle, toute attribution s'avère complexe en raison de l'absence d'un véritable droit d'auteur et par la possibilité d'avoir un (ou deux) co-auteur(s) pour une pièce (Craig & Kinney, 2009). Dans le monde francophone, le débat s'est focalisé sur le tandem Molière – P. Corneille. L'objectif de cette communication n'est pas de trancher cette question mais d'appliquer les méthodes d'attribution d'auteur les plus fiables sur les comédies du XVIIe siècle (en excluant celles attribuées à Molière ou P. Corneille).

En présence d'une pièce de théâtre, on peut supposer que son véritable auteur correspond au nom imprimé sur la couverture (affirmation que nous nommerons hypothèse de la signature). Pour vérifier cette thèse, nous avons repris vingt comédies en alexandrins écrites par neuf auteurs (J. G. deCampistron (1656—1723), Champmeslé (1642—1701), Chevalier (16..—1673), Hauteroche (1617—1707), Montfleury (1608—1667), P. Quinault (1635—1688), J. Racine (1639-–1699), Tristan (1601—1655), et T. Corneille (1625—1709)). Si cette hypothèse se vérifie, nous devrions distinguer neuf écrivains présentant des styles dissemblables. Dans le cas contraire, l'hypothèse de la signature doit être abandonnée et nous devrions identifier le véritable auteur pour chaque œuvre.

Dans cette étude, nous supposons que les comédies retenues ont été écrites par un seul auteur, ce qui est généralement le cas. On peut rencontrer de temps à autre des exceptions, comme Psyché (1671) rédigé par P. Corneille, Molière et P. Quinault (selon l'hypothèse de la signature). De plus, l'auteur est la personne qui a rédigé le texte et non celle qui a fourni l'intrigue, des dialogues, des scènes comiques ou financé la rédaction de l'œuvre.

Dans la suite de cet article, nous présenterons un survol des connaissances en attribution d'auteur (section 2). La troisième section décrit nos trois corpus. La quatrième explique l'attribution fondée sur une distance intertextuelle et la cinquième expose la méthode Delta et nos améliorations. Enfin, une dernière section applique ces deux approches sur les vingt pièces de théâtre français. Une conclusion dresse les principaux résultats de cette étude.

2 État des connaissances

En catégorisation de textes (Sebastiani, 2002), on distingue entre les modèles basés sur la sémantique (e.g., indexation automatique, filtrage, etc.) ou le style (Savoy, 2020), (Karsdorp *et al.*, 2021). Dans ce dernier cas, les applications en stylométrie couvrent un champ assez large, allant de l'attribution et profilage d'auteur, à la détection de faux ou de plagiat, support en criminologie (Olsson, 2018), voire la datation d'un document (Kreuz, 2023).

Les premières approches datant du XIXe siècle (Mendenhall, 1887) ont proposé des mesures simples comme la longueur moyenne des mots, ou le pourcentage de termes apparaissant une seule fois (*hapax*) afin d'identifier différents styles. Toutefois, l'instabilité de ces mesures face à des textes de longueur variable rend ces solutions inopérantes (Baayen, 2008).

En se limitant à la question de l'identification de l'auteur, trois contextes sont possibles. Premièrement, dans un environnement fermé, le véritable auteur est l'un des écrivains proposés et, pour chaque candidat un ensemble de textes est fourni. Deuxièmement, l'auteur peut être un des noms mentionnés ou un autre, encore inconnu (contexte ouvert). Troisièmement, le système doit répondre si deux

textes ont été rédigés par la même plume ou non. Dans ces trois situations, une réponse ne saurait se limiter à un simple nom, et une justification plus complète devrait être fournie. Au cours de ces trois dernières décennies, des approches plus performantes ont été proposées que l'on peut classifier selon les attributs stylistiques retenus d'une part et, d'autre part, les mesures de similarité (ou de distance) appliquées.

Dans une première grande famille, on peut regrouper les modèles s'appuyant sur un ensemble de vocables sélectionnés. Ces derniers peuvent être fréquemment employés de manière inconsciente et correspondent à des mots-outils comme les articles (le, des), pronoms (tu, nous), prépositions (sur, vers), conjonctions (et, mais) et des verbes auxiliaires et modaux (est, avait, fait) (Hughes *et al.*, 2012). Cette énumération correspond à un anti-dictionnaire (*stoplist*) usité habituellement par les moteurs de recherche. D'autres modèles suggèrent de tenir compte des termes relativement fréquents chez un auteur et peu ou pas employés par les autres (Burrows, 2007), (Craig & Kinney, 2009). Finalement, certains modèles se fondent sur l'ensemble du vocabulaire, parfois en éliminant certains vocables peu fréquents (Labbé, 2007) ou en laissant le modèle sélectionner de manière automatique les termes les plus pertinents.

Avec l'accroissement des capacités des ordinateurs et les campagnes d'évaluation CLEF-PAN sur ce thème (Rosso *et al.*, 2019), la distinction entre différents styles peut s'appuyer sur de brèves séquences de lettres (n-grammes, avec $n = 2$ à 6) (Kjell, 1994). Évidemment, la combinaison des mots et des chaines de lettres permet de fournir un grand nombre d'attributs stylistiques (Savoy, 2020).

La détermination du style propre d'un auteur peut également s'établir en considérant la syntaxe comme, par exemple, en se fondant sur de courtes séquences de mots, soit les plus fréquentes, soit celles respectant des patrons prédéfinis (e.g., adjectif-nom-nom) (Kocher & Savoy, 2019). Comme autre source d'information, la longueur moyenne des phrases ou leur distribution peuvent fournir des attributs complémentaires et discriminants.

Dès que chaque texte est représenté par une liste d'attributs, les modèles d'attribution peuvent se baser sur une fonction de distance (ou similarité) entre textes ou recourir à un modèle d'apprentissage automatique (e.g., SVM, les k plus proches voisins) voire par un apprentissage profond. Dans tous les cas, l'attribution s'établit selon la règle du voisin le plus similaire (ou de la distance la plus faible). Signalons toutefois que toute stratégie basée sur l'apprentissage par machine requiert un jeu d'entrainement, données qui ne sont pas toujours disponibles. De plus, une forte corrélation doit exister entre le jeu d'apprentissage et celui du test. Ces contraintes impliquent que le contexte d'attribution soit fermé (le véritable auteur doit être présent dans le jeu d'apprentissage). Dans le cas contraire (contexte ouvert), la réponse proposée risque d'être erronée.

3 Les trois corpus étudiés

Afin d'analyser quelques comédies du XVIIe siècle, nous avons recouru à trois corpus différents. Nos deux premiers corpus serviront à vérifier la qualité des deux méthodes retenues pour l'attribution d'auteur. La première collection se compose de 150 romans contemporains écrits en italien par 40 auteurs distincts. Ce corpus nommé PIC (*Padova Italian Corpus*) a été construit par un groupe de chercheurs à l'Université de Padoue sous la supervision du prof. Cortelazzo et du prof. Tuzzi (Tuzzi & Cortelazzo, 2018).

La table 3 (dans les annexes) indique le nom des auteurs, le sexe et le nombre de romans inclus

dans le corpus. On y retrouve 27 hommes et 12 femmes de même que le nom E. Ferrante dont on souhaite découvrir la véritable identité. Tous les romans correspondent au même genre, soit des textes pour adultes. Tous les éléments n'appartenant pas au récit ont été soigneusement éliminés (e.g., numérotation des pages, titre courant, etc.). Le roman le plus long comprend 196 914 formes (Faletti, *Io uccito*, 2002) et le plus bref seulement 7 694 formes (Parrella, *Behave*, 2011, le seul document ayant moins de 10 000 formes).

Dans l'évaluation d'une procédure d'attribution d'auteur, il est important de souligner trois contraintes. D'abord, chaque document doit être assez long. Dans ce corpus italien, chaque test contient 10 000 formes ou plus (à une exception près). Ensuite, l'orthographe a été vérifiée. Enfin, d'autres facteurs pouvant influencer le style doivent être réduits au strict minimum. Ainsi cette collection renferme des textes écrits dans la même langue, et rédigés durant la même période (de 1987 à 2016).

Le deuxième corpus nommé St-Jean comprend 200 extraits de 67 romans écrits en français par 30 auteurs différents (Labbé, 2017). La distribution entre les divers auteurs est indiquée dans la table 4 (voir annexes). Chaque document comprend approximativement 10 000 formes. La plage temporelle couverte par ce corpus s'étend sur tout le XIXe siècle de Chateaubriand (*Atala*, 1801) à Proust (*Les Plaisirs et les jours*, 1896). Ce corpus respecte également les contraintes citées ci-dessus.

Le troisième corpus comprend une sélection de vingt pièces de théâtre dont la table 5 décrit les principales caractéristiques (voir annexes). Ces œuvres sont toutes rédigées en français et correspondent à des comédies en vers. Elles contiennent en général plus de 10 000 formes et couvrent la période de 1651 à 1709.

4 La distance intertextuelle

Afin de déterminer si deux documents sont rédigés par le même auteur, nous comparons l'ensemble du vocabulaire (Labbé, 2007). Si les termes employés et leur fréquence s'avèrent proches, la distance intertextuelle sera faible. Dans le cas contraire, elle s'élèvera jusqu'à un maximum de 1,0 lorsque deux textes ne possèdent rien en commun comme un roman écrit en français et un autre en finnois. A l'inverse, si les deux documents sont identiques, la distance sera nulle.

Plus précisément, la distance intertextuelle entre le texte A et B (notée $D(A,B)$) est indiquée dans l'équation 1 dans laquelle n_A signale le nombre de mots (*tokens*) du texte A et tf_{iA} la fréquence absolue du terme i (pour $i = 1, 2, \ldots, m$) dans le texte A. La taille du vocabulaire est indiquée par m. Si l'on admet que le texte B est plus long que le texte A, nous devons réduire les fréquences des termes appartenant à B. Ces dernières (notées tf_{iB}) sont multipliées par le rapport des tailles comme présenté à droite la formule 1.

$$D(A,B) = \frac{\sum_{i=1}^{m} |tf_{iA} - \hat{tf}_{iB}|}{2 \cdot n_A} \qquad avec\ \hat{tf}_{iB} = tf_{iB} \cdot \frac{n_A}{n_B} \tag{1}$$

La machine calcule la distance entre toutes les paires de romans présents dans un corpus (soit 200 x199 / 2 = 19 900 pour la collection St-Jean). Pour chaque œuvre, on peut alors déterminer les k extraits les plus proches.

Avec le corpus St-Jean, le plus proche voisin de chaque extrait correspondait toujours à un passage d'un roman rédigé par le même auteur. Une distance intertextuelle inférieure à 0,2 constitue une très

forte évidence que les deux textes sont du même auteur (Labbé, 2007), (Savoy, 2018). Parmi les facteurs pouvant favoriser une faible distance on peut ajouter la présence de texte étant du même genre et écrit dans la même décennie. La précision est donc de 100 % avec la règle empirique de 0,2 ; en d'autres termes, si la distance est inférieure à 0,2, les deux documents sont de la même plume. Cependant, deux textes du même auteur peuvent présenter une distance supérieure à cette limite. Le taux de rappel n'est donc pas parfait.

Avec la collection PIC dans laquelle on compare tout un roman avec un autre, le résultat s'avère similaire sauf pour certains romans d'Elena Ferrante. Pour être précis, seulement les trois derniers romans d'Elena Ferrante présentent un appariement inférieur à 0,2 avec trois romans de D. Starnone (soit *Autobiografia, Lacci* ou *Schezetto*).

En analysant les autres écrivains italiens, on constate généralement que les romans d'un même auteur présentent des distances supérieures à 0,2 (Savoy, 2018). On peut avancer l'hypothèse qu'un écrivain cherche souvent d'autres perspectives ou souhaite aborder d'autres thèmes avec un ton quelque peu différent. Toutefois, notre corpus italien indique que pour trois auteurs (Carofiglio, Faletti ou Veronesi), les distances entre romans demeurent souvent inférieures à 0,2.

5 Delta

Afin d'identifier le style d'un texte, le modèle Delta tient compte des m vocables (ou lemmes) les plus fréquents employés par les auteurs du corpus étudié (Burrows, 2002). L'opérateur est libre de fixer la valeur de m, mais les plus courantes varient entre 50 et 500. L'idée sous-jacente consiste à tenir compte des mots fréquents étant peu ou pas porteurs de sens et utilisés de manière inconsciente par l'auteur. Avec une valuer de m inférieure à 300 ou 400, la large majorité d'entre eux sont des mots-outils qui s'avèrent indépendant des thèmes du texte et donc plus associés au style de l'auteur.

Chaque terme t_i sélectionné possèdera un poids dénoté $Zscore(t_{ij})$ correspondant à la différence entre sa fréquence relative dans le texte T_j (notée rtf_{ij}) et la moyenne ($\overline{rtf}_i$) pour ce terme t_i sur l'ensemble des textes du corpus. Afin de tenir compte de la variabilité sous-jacente, chaque différence est divisée par l'écart-type (s_i).

$$Z\ score(t_{ij}) = \frac{rtf_{ij} - \overline{rtf}_{ij}}{s_i} \tag{2}$$

Étant donné un texte Q dont l'attribution est incertaine et un texte A_j (écrit par l'auteur A_j), nous pouvons calculer la différence en valeur absolue entre les scores Z du texte Q et A_j et d'en calculer la moyenne (voir équation 3).

$$\Delta(Q, A_j) = \frac{1}{m} \cdot \sum_{i=1}^{m} |Z\ score(t_{iQ}) - Z\ score(t_{iA_j})| \tag{3}$$

dans laquelle m indique le nombre de termes sélectionnés et t_{iA_j} le ie terme dans le texte A_j (de même pour t_{iQ} et le texte Q). Dans nos expériences, la valeur de m a été fixée à 200. Ce choix s'explique par notre souci de baser notre similarité stylistique sur les mots-outils et non des termes

TABLE 1 – Distribution des probabilités dans un contexte ouvert (“O”) ou fermé (“F”)

	1	2	3	4	5	6	7	8	9	10
O	30,7%	16,0%	10,6%	8,0%	6,0%	4,8%	3,9%	3,3%	2,7%	2,3%
F	94,4%	1,8%	1,0%	0,6%	0,4%	0,3%	0,2%	0,2%	0,2%	0,1%

liés aux thèmes des pièces de théâtre. Comme m varie d'une application à l'autre, on ne peut pas calibrer les distances retournées comme la distance intertextuelle le permettait.

Souvent appliquée en stylométrie (Karsdorp *et al.*, 2021), cette approche permet d'estimer une distance entre des textes, avec la plus faible valeur indiquant l'auteur probable du document Q. Afin d'estimer une probabilité que la distance obtenue indique le véritable auteur, nous proposons d'appliquer la fonction $softmin()$ sur l'ensemble des distances calculées comme exprimée dans l'équation 4.

$$Prob(A_j) = \frac{e^{-c \cdot \Delta(Q, A_j)}}{\sum_{i=1}^{k} e^{-c \cdot \Delta(Q, A_i)}} \tag{4}$$

dans laquelle $Prob(A_j)$ signale la probabilité que le texte Q ait été rédigé par A_j. La constante c, fixée à 20 dans nos expériences, permet de mieux disperser les valeurs retournées par la fonction Delta.

Ce modèle admet implicitement que le véritable auteur se trouve parmi les écrivains proposés (contexte fermé). Si ce n'est pas le cas, Delta classera tout de même tous les auteurs présents selon leur similarité avec le document cible. Afin de vérifier les divergences entre les valeurs retournées dans ces deux contextes, nous avons repris nos extraits de romans écrits en français. Par itération, on a recherché l'auteur de chaque extrait. Dans une première expérience, le véritable auteur a toujours été éliminé (contexte ouvert). En observant les valeurs Delta retournées, on constate que les intervalles de distance entre les cinq premiers rangs demeurent relativement faibles.

En transformant ces distances en probabilités, nous pouvons extraire une distribution des probabilités pour le contexte ouvert (“O”) ou fermé (“F”) (voir la table 1). Dans le premier cas (étiquette “O”), on constate que l'attribution au plus proche voisin signale une probabilité moyenne de 30,7 % pour le premier rang, 16 % pour le deuxième et 10,6 % pour le troisième. En sommant ces trois valeurs, on obtient un total de 57,3 %. En considérant les rangs suivants, les probabilités décroissent lentement. Comme le véritable auteur est absent, on comprend que cette distribution de probabilités se disperse sur les dix ou quinze premiers.

Par contre, si le véritable auteur est présent (contexte fermé, ligne “F”), la probabilité estimée pour le premier rang s'élève à 94,4 %. La différence entre le premier et le deuxième rang s'avère très nette. Des histogrammes similaires peuvent être générés depuis le corpus italien.

Bien que l'application directe de la méthode Delta ne puisse apporter une attribution claire, le recours à une estimation des probabilités permet de résoudre ce problème.

TABLE 2 – Distance inférieure 0,2 entre des pièces d'auteurs différents

AuteurTitre	AuteurTitre	AuteurTitre	AuteurTitre	AuteurTitre
CamJal	MonFem			
ChaPar				
ChaRag				
CheCar				
ChePéd	HauMus			
HauAma	MonFem	MonCom	QuiMèr	TCoDom/Alb/Fes/Com
HauMus	ChePéd	MonCom	QuiMèr	TCoAlb/Fes
MonFem	CamJal	HauAma	QuiMèr	TCoDom/Amo/Gal/Alb/Fes
MonCom	HauAma	HauMus		TCoCom/Alb/Fes
QuiRiv				
QuiMèr	HauAma	HauMus	MonFem	TCoAmo/Alb/Com/Fes
RacPla				
TriAma				
TriPar				
TCoDom	HauAma	MonFem		
TCoAmo	MonFem	QuiMèr		
TCoGal	MonFem			
TCoAlb	HauAma	HauMus	MonFem	MonCom QuiMèr
TCoCom	HauAma	MonCom	QuiMèr	
TCoFes	HauAma	HauMus	MonFem	MonCom QuiMèr

6 Evaluation

Reprenons notre corpus de vingt comédies du théâtre français. Notre analyse stylométique s'effectuera pièce par pièce sans supposer que toutes les œuvres parues sous le même auteur sont effectivement écrites par la même plume.

Afin de pouvoir présenter nos résultats, nous désignons chaque pièce par les trois premières lettres de son auteur (e.g., "Hau" pour Hauteroche) suivies des trois premières lettres du mot principal dans le titre (e.g., "Ama" pour *L'amant qui ne flatte point*). Comme notre corpus contient six œuvres de T. Corneille et que ces dernières sont souvent proches des autres pièces, nous ne répèterons pas le nom de T.Corneille (ou "TCo") pour chaque comédie. Ainsi l'acronyme "TCoAlb/Fes" indique les deux pièces de T. Corneille *Le baron d'Albikrac* et *Le festin de Pierre*.

La table 2 indique les rapprochements stylistiques déduits avec la méthode de la distance intertextuelle. Dans cette table, les similarités provenant de comédies du même auteur sont ignorées. Si l'hypothèse de la signature se confirme, la table 2 doit être vide. Pour chaque texte, aucune forte similarité stylistique ne devrait exister avec une comédie écrite par un autre auteur. Dans la table 2, certaines comédies ne se rapprochent d'aucune autre comme les deux pièces attribuées à Champmeslé, une à Chevalier (*L'intrigue des carrosses*), une de P. Quinault (*Les rivales*), deux à Tristan et *Les plaideurs* de J. Racine. Selon la distance intertextuelle, ces pièces respectent l'hypothèse de la signature.

A l'inverse, les deux pièces de Hauteroche ou de Montfleury s'apparentent fortement à des comédies d'Hauteroche, Montfleury, P. Quinault et T. Corneille. Avec J. G. de Campistron, l'unique comédie retenue (*Le jaloux désabusé*) se rapproche d'une pièce de Montfleury (*La femme juge et partie*). Pour

la comédie *Le pédagogue amoureux* de Chevalier, on observe également une similitude stylistique avec une comédie signée Hauteroche (*Crispin musicien*).

Dans les grandes lignes, la table 2 signale une forte similarité entre les comédies de Hauteroche, Montfleury, T. Corneille et une pièce de Quinault. Est-ce que ce rapprochement provient des thèmes et partiellement du style ?

Afin d'exclure une forte influence des thèmes, nous avons appliqué l'approche Delta sur l'ensemble des 13 comédies présentant un rapprochement stylistique. Les résultats obtenus confirment une similarité stylistique entre ces pièces. Par exemple, la moyenne des probabilités associées au premier rang pour ces 13 comédies s'élève à 38,7% et à 21% pour la seconde position. La méthode Delta n'arrive pas à identifier une plume unique pour chacune de ces œuvres.

En considérant P. Quinault et Chevalier, on remarque que pour ces deux auteurs une comédie présente de forte similarité stylistique avec d'autres œuvres tandis que la seconde propose un style original. Cette constatation justifie une analyse texte par texte au lieu de travailler avec un profil d'auteur généré depuis l'ensemble de ses œuvres. Toutefois cette absence de similitude entre pièces peur aussi s'expliquer par le choix restreint de cette étude. Seulement vingt comédies ont été analysées. De plus, d'autres auteurs n'ont pas été repris.

7 Conclusion

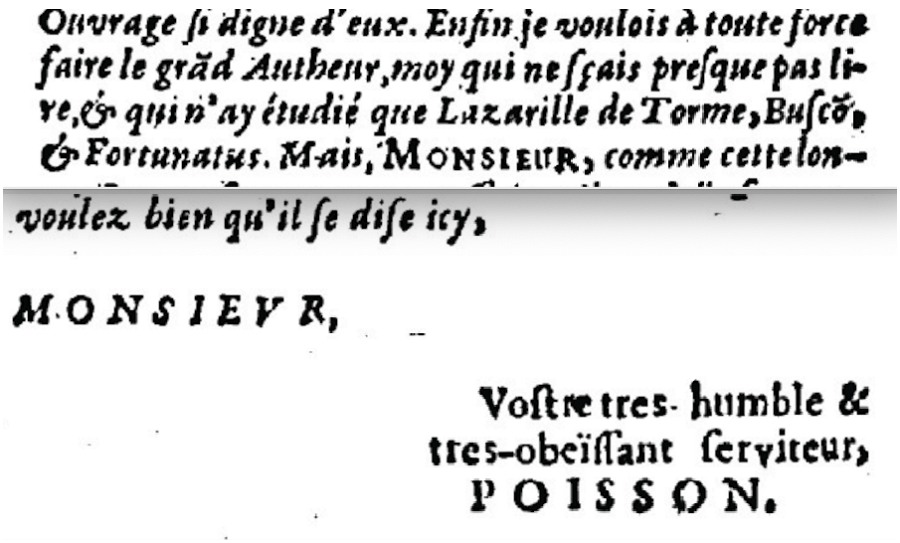
Cette étude explore les similarités stylistiques entre vingt comédies françaises écrites au XVIIe siècle par neuf auteurs distincts. Dans un premier temps, nous avons admis l'hypothèse de la signature ; le véritable auteur s'avère celui dont le nom apparait sur la couverture. Afin de vérifier cette affirmation, nous avons appliqué deux méthodes d'attribution d'auteur soit la distance intertextuelle (méthode non-supervisée) (Labbé, 2007) et l'approche Delta(Burrows, 2002).

Afin d'adapter cette dernière à un contexte ouvert (le véritable auteur demeure inconnu), nous proposons d'estimer une probabilité d'attribution sur la base des distances retournées par le modèle Delta. Avec cette estimation, nous observons clairement deux distributions distinctes de probabilités, à savoir lorsque le véritable auteur est présent ou absent du jeu d'entrainement. Avec cet ajustement, la méthode Delta s'applique à un contexte ouvert (le véritable auteur demeurant inconnu) et l'interprétation des distances s'en trouve simplifié.

Deuxièmement, nous avons testé l'hypothèse de la signature. Comme l'indique la table 2, cette dernière doit être écartée. Certes, pour certains écrivains (e.g., J. Racine, Tristan, Champmeslé), aucune similarité importante avec d'autres textes n'a été détectée. Pour d'autres, le style de certaines comédies s'apparente fortement au style d'œuvres parues sous le nom d'autres auteurs. En particulier, ce rapprochement touche des textes attribués à Hauteroche, Montfleury, P. Quinault et T. Corneille. Si nous n'avons pas quatre auteurs, combien en avons-nous réellement ? Doit-on pencher vers un auteur unique qui pourrait être T. Corneille ?

Avec les approches proposées, cette étude ne permet pas d'identifier sans l'ombre d'un doute le véritable auteur ou d'indiquer si certaines de ces comédies ont été écrites par deux (ou plusieurs) écrivains. Enfin, notre démarche s'appuie uniquement sur les textes et ignore les évidences externes (comme, par exemple, une étude de biographie comparée des auteurs, une concordance des dates, un livre de compte (Young & Young, 1977), ...) pouvant confirmer ou infirmer une possible attribution.

Un exemple d'évidence externe explicite est repris dans la figure 1. Dans cette préface, l'auteur (R. Poisson) indique qu'il sait "presque pas lire", un exemple complémentaire qui signale que l'hypothèse de la signature est infirmée.


Ouvrage ſi digne d'eux. Enfin je voulois à toute force
faire le grãd Autheur, moy qui ne ſçais preſque pas li-
re, & qui n'ay étudié que Lazarille de Torme, Buſcõ,
& Fortunatus. Mais, MONSIEUR, comme cette lon-
voulez bien qu'il ſe diſe icy,

MONSIEVR,

Voſtre tres-humble &
tres-obeïſſant ſerviteur,
POISSON.

FIGURE 1 – Extrait de la préface du *Le Poète basque* (1668) de R. Poisson

Remerciements

Le corpus Ferrante a été créé par les professeurs A. Tuzzi et M. Cortelazzo (Tuzzi & Cortelazzo, 2018) de l'Université de Padoue. Le corpus St-Jean a été construit par D. Labbé (Labbé, 2017).

Références

BAAYEN H. (2008). *Analysis Linguistic Data : A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

BURROWS J. (2002). Delta : A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**(3), 267–287.

BURROWS J. (2007). All the way through : Testing for authorship in different frequency strata. *Literary and Linguistic Computing*, **22**(1), 27–47.

CRAIG H. & KINNEY A. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press, Cambridge.

HUGHES J., FOTI N., KRAKAUER D. & ROCKMORE D. (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Science (PNAS)*, **109**(20), 7682–7686.

KARSDORP F., KESTEMONT M. & RIDDELL A. (2021). *Humanities Data Analysis. Case Studies with Python*. Princeton University Press, Princeton.

KJELL B. (1994). Authorship determination using letter pair frequency features with neural network classifier. *Literary and Linguistics Computing*, **9**(2), 119–124.

KOCHER M. & SAVOY J. (2019). Evaluation of text representation schemes and distance measures for authorship linking. *Digital Scholarship in the Humanities*, **34**(1), 189–207.

KREUZ R. (2023). *Linguistics Fingerprints. How Language Creates and Reveals Identity*. Prometheus Books, Guilford.

LABBÉ D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, **14**(1), 33–80.

LABBÉ D. (2017). *Une expérience d'attribution d'auteur. Le corpus St-Jean*. Rapport interne, Université de Grenoble.

MENDENHALL T. (1887). The characteristic curves of composition. *Science*, **214**, 237–249.

MICHELL J. (1999). *Who Wrote Shakespeare*. Thames and Hudson, London.

OLSSON J. (2018). *More Wordcrime. Solving Crime Through Forensic Linguistics*. Bloomsbury, London.

ROSSO P., POTTHAST M., STEIN B., STAMATATOS E., RANGEL F. & DAELEMANS W. (2019). *Evolution of the PAN Lab on Digital Text Forensics*, In *Information Retrieval Evaluation in a Changing World*, p. 461–486. Springer, Cham.

SAVOY J. (2018). Is Starnone really the author behind Ferrante ? *Digital Scholarship in the Humanities*, **33**(4), 902–918.

SAVOY J. (2020). *Machine Learning Methods for Stylometry : Authorship Attribution and Author Profiling*. Springer, Cham.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Survey*, **14**(1), 1–27.

TASSINARI L. (2009). *John Florio, The Man who was Shakespeare*. Giano Books, New York.

TUZZI A. & CORTELAZZO M. (2018). What is Elena Ferrante ? A comparative analysis of a secretive bestselling Italian writer. *Digital Scholarship in the Humanities*, **33**(3), 685–702.

YOUNG B. & YOUNG G. (1977). *Le Registre de La Grange 1659-1685*. Slatkine, Genève.

Annexe

TABLE 3 – Nom des auteurs, sexe (H/F), et nombre de romans

Nom	Sexe	Nombre	Nom	Sexe	Nombre
Affinati	H	2	Montesano	H	2
Ammaniti	H	4	Morazzoni	F	2
Bajani	H	3	Murgia	F	5
Balzano	H	2	Nesi	H	3
Baricco	H	4	Nori	H	3
Benni	H	3	Parrella	F	2
Brizzi	H	3	Piccolo	H	7
Carofiglio	H	9	Pincio	H	3
Covacich	H	2	Prisco	H	2
De Luca	H	4	Raimo	H	2
De Silva	H	5	Ramondino	F	2
Faletti	H	5	Rea	H	3
Ferrante	?	7	Scarpa	H	4
Fois	H	3	Sereni	F	6
Giordano	H	3	Starnone	H	10
Lagioia	H	3	Tamaro	F	5
Maraini	F	5	Valerio	F	3
Mazzantini	F	4	Vasta	H	2
Mazzucco	F	5	Veronesi	H	4
Milone	F	2	Vinci	F	2

TABLE 4 – Nom des auteurs, nombre d'extraits et nombre de romans

Nom	Extrait	Roman	Nom	Extrait	Roman
Balzac	13	6	Maupassant	10	5
Barbey	8	2	Musset	3	1
Bourget	6	1	Nerval	5	2
Chateaubriand	3	2	Proust	3	1
Daudet	6	1	Régnier	8	2
Dumas	10	3	Sainte-Beuve	6	1
Erckmann	4	1	Sand	10	3
Flaubert	11	6	Staël	6	1
France	8	2	Stendhal	6	2
Fromentin	5	1	Sue	10	1
Gautier	5	3	Vallès	4	1
Goncourt	5	2	Verne	4	2
Huysmans	4	2	Victor	8	2
Lamartine	7	2	Vigny	5	2
Loti	7	2	Zola	10	5

TABLE 5 – Nom des auteurs, titre des comédies, année de parution et longueur

Auteur	Titre	Année	Formes
Campistron	*Le jaloux désabusé*	1709	14 383
Champmeslé	*Le Parisien*	1684	18 091
Champmeslé	*Ragotin*	1684	15 596
Chevalier	*L'intrigue des carrosses*	1662	10 154
Chevalier	*Le pédagogue amoureux*	1665	16 977
Hauteroche	*L'amant qui ne flatte point*	1668	20 908
Hauteroche	*Crispin musicien*	1671	22 350
Montfleury	*La femme juge et partie*	1669	17 464
Montfleury	*Le comédien poète*	1673	21 771
Quinault	*Les rivales*	1653	18 680
Quinault	*La mère coquette*	1665	19 452
Racine	*Les plaideurs*	1668	10 063
Tristan	*Amaryllis*	1652	16 124
Tristan	*Le parasite*	1654	18 701
T. Corneille	*Dom Bertran de Cigarral*	1651	20 911
T. Corneille	*L'amour à la mode*	1651	20 819
T. Corneille	*Le galant doublé*	1659	21 152
T. Corneille	*Le baron d'Albikrac*	1667	20 558
T. Corneille	*La comtesse d'Orgueil*	1670	21 124
T. Corneille	*Le festin de Pierre*	1677	20 068