Réponse à Florian Cafiero et Jean-Baptiste Camps. Why Molière most likely did write his plays. Science Advances. 5: eaax5489. 27 November 2019.

Dominique Labbé

Pacte
Laboratoire des sciences sociales

Réponse à Florian Cafiero and Jean-Baptiste Camps. Why Molière most likely did write his plays.

Published in *Science Advances.* 5. 27 November 2019.

(19 décembre 2019)

D. Labbé
Pacte – Université Grenoble-Alpes
dominique.labbe@umrpacte.fr

Résumé

Dans cet article, publié en "open access", Cafiero et Camps prétendent apporter la preuve définitive que P. Corneille n'a écrit aucune des pièces présentées par Molière. Ils utilisent pour cela 6 caractères "features" (lemmes, formes, mots outils, rimes, affixes, n-grams) couplés avec deux calculs de distances (Burrows et MinMax) qui servent à des classifications automatiques. En fait, les auteurs fournissent peu d'informations précises sur ces méthodes et aucune donnée chiffrée. Les quelques informations, notamment dans les annexes en ligne, suffisent pour soulever beaucoup de doutes. Par exemple, la liste des "mots outils" comporte de nombreuses étrangetés qui ne peuvent s'expliquer simplement par des maladresses. De même, ils ont opéré un tri dans les pièces de Molière, retirant de l'expérience 24 des 33 pièces. Parmi ces pièces écartées : *Psyché* qui est une collaboration avouée entre Molière et Corneille, pour laquelle les passages censés écrits par chacun sont identifiés. Enfin, le détail des classifications (publié dans une annexe en ligne séparée de l'article) montre un échec total. Leurs méthodes se révèlent incapables de reconnaître : Boursault, Chevalier, Dancourt, Donneau de Visé, Gillet de la Tessonnerie, Pierre Corneille, Thomas Corneille, La Fontaine, Ouville, Quinault, Régnard, Rotrou… et Molière.

PLAN DU DOCUMENT

1. INTRODUCTION

L'article de Cafiero et Camps appelle trois remarques préliminaires.

Premièrement, il est paru dans un magazine généraliste en "open access" (accès libre) qui n'est pas reconnu dans les milieux scientifiques. Quoi qu'en dise la direction de ce magazine, les articles ne sont pas révisés par des pairs spécialisés dans le domaine. Ce serait d'ailleurs impossible puisque ce magazine publie de tout sur n'importe quel sujet. Les auteurs paient cher pour être publiés (4500$)[1], procédé condamné par les scientifiques. Autrement dit, ce faux journal scientifique est un prédateur qui vit de la crédulité d'universitaires en mal de publications, notamment dans le tiers monde. Aucun chercheur chevronné ne soumet à ce genre de revue et, dans la communauté scientifique, personne ne perd son temps à lire les articles qui y paraissent.

Cependant, le CNRS, l'Ecole des Chartes, l'Université de Paris, l'AFP… et même l'Université de Neuchâtel ont annoncé cette publication par des communiqués triomphalistes[2]. La grande presse, les médias s'en sont fait largement écho. Certains ont fait passer ce "junk journal" pour une "revue prestigieuse".

Il a donc fallu lire…

Deuxièmement, dès le début de cet article, il est clair qu'il est dirigé contre notre travail sur P. Corneille et Molière (sur les œuvres concernées, voir les annexes 1 à 3 à la fin de notre texte) et contre notre méthode d'attribution d'auteur (présentée en annexe 4). Comme l'indiquent les "Aknowledgments" à la fin de leur article, Cafiero et Camps ne nous ont pas contactés et ne nous ont pas communiqué leur article avant publication comme il est d'usage quand des chercheurs sont sérieusement mis en cause par les conclusions. S'ils avaient pris cette précaution, nous leur aurions communiqué confidentiellement nos remarques et objections. Leur manquement aux usages nous oblige à rendre ces commentaires publics. Au passage, cette analyse éclairera les curieuses pratiques de certains milieux littéraires. Elle aidera peut-être aussi à faire mieux comprendre la statistique appliquée au langage et les méthodes de la recherche scientifique.

Troisièmement, l'article de Cafiero et Camps contient très peu d'informations sur les mesures et pas de résultats chiffrés. En quelque sorte, l'article se contente de dire : "tous les chiffres sont bons", "c'est prouvé". Cette attitude n'est pas admise en matière de publication scientifique. Mais les auteurs

---

[1] https://advances.sciencemag.org/content/licensing-and-charges
Sur l'éditeur en chef (Holden Thorp) : https://www.documentcloud.org/documents/1344054-full-wainstein-report.html

[2] Par exemple, : http://www.cnrs.fr/fr/corneille-na-pas-ecrit-les-pieces-de-moliere ;
http://www.chartes.psl.eu/fr/actualite/jean-baptiste-camps-florian-cafiero-confirment-paternite-oeuvres-moliere ;
https://u-paris.fr/moliere-a-bien-signe-toutes-ses-pieces-corneille-ny-est-pour-rien/

https://twitter.com/UniNeuchatel/status/1199959956537139200 et
https://www.facebook.com/UniNeuchatel/posts/2813867858632000/

disent que les fichiers placés en ligne sur le site de la revue[3] comblent ces manques et assurent la "transparence" et la "reproductibilité" de leurs expériences.

Nous examinerons cette première question dans la section 2. Puis nous verrons que la sélection des textes a été faite de curieuse manière (section 3), que la qualité des données et des calculs ne sont pas à la hauteur de ce qui est attendu en pareil cas (section 4), et surtout que les quelques informations fournies démontrent un échec complet (section 5).

2. PAS DE TRANSPARENCE, CONTROLE IMPOSSIBLE

Examinons d'abord la question cruciale : Cafiero et Camps ont-ils fourni en ligne les fichiers attendus et, dans l'affirmative, ces informations suffisent-elles pour contrôler leurs conclusions ?

On trouve en ligne trois dossiers :

- aax5489_Data_file_S1.zip (que nous désignerons pas : "Data_S1")
- aax5489_Data_file_S2.7z (soit "Data_S2")
- aax5489_Data_file_S3.7z (soit "Data_S3")

Et un document à part : Adobe PDF - aax5489_SM.pdf (ci-dessous : "annexe en ligne")

Pour déterminer les données devant figurer dans ces dossiers en ligne, il est nécessaire de comprendre le cheminement des auteurs au long de la procédure classique en matière d'attribution d'auteur. Il faut également se souvenir que Cafiero et Camps affirment avoir répété trois fois la chaîne de traitements : une phase "exploratoire" pour démontrer l'efficacité de la méthode et préciser les critères de sélection des textes ; un test principal ("main") pour décider de l'attribution des pièces de Molière ; enfin un test de "contrôle" sur des textes comparables pour vérifier que tout fonctionne correctement. Pour chaque étape de l'attribution d'auteur, il faut donc trouver, dans les dossiers en ligne, trois jeux de données (exploratory, main, control).

Suivons les cinq étapes d'une attribution d'auteur, en vérifiant l'existence et la consistance de ce que Cafiero et Camps ont mis en ligne.

1. La sélection des textes et le calibrage de l'étiqueteur

Ce sont les premières opérations obligées de toute attribution d'auteur. La solidité des conclusions dépend de leur qualité.

---

[3] http://advances.sciencemag.org/cgi/content/full/5/11/eaax5489/DC1. Dernière consultation le dimanche 15 décembre à 9h00. A cette date aucun changement n'était intervenu dans les annexes mises en ligne avec l'article (27 novembre 2019).

Etrangement, l'article de Cafiero et Camps ne contient pas de liste des textes utilisés mais les trois "corpus" (exploratoire, principal, contrôle) seraient présentés dans les tableaux S1, S2 et S3 au début de l'annexe en ligne. En gros, le corpus "exploratoire" contiendrait des pièces par des auteurs contemporains de Molière ; celui de "contrôle", des pièces jouées après les morts de Molière et de P. Corneille. La principale expérience ("main") serait conduite sur un corpus qui contiendrait une trentaine de pièces présentées par Molière, les deux frères Corneille et deux autres auteurs. Mais, comme expliqué plus bas, un grand flou règne quant au nombre et aux titres des pièces.

Parallèlement à cette sélection, les auteurs ont procédé à "l'apprentissage" de l'étiqueteur. En fait, il s'agit d'un calibrage ou d'un étalonnage, car les algorithmes utilisés ne sont pas "intelligents". Ce programme découpe les mots dans les textes et attache à chacun d'eux une étiquette (comportant leurs entrées de dictionnaire et leurs catégories grammaticales). De la qualité de ces opérations dépend celle des quatre phases suivantes et des conclusions, exactement comme un télescope mal réglé ne donnera pas d'image fiable et ne permettra aucune conclusion concernant l'univers.

Pour l'étalonnage de leur étiqueteur, Cafiero et Camps indiquent avoir utilisé 41 pièces étiquetées remises par S. Gabay (Université de Neuchâtel) qui aurait collaboré avec eux à cette phase préliminaire. Dans l'article, ils affirment : "The data and models are available as supplementary material Data S1."

En fait, dans Data_S1, le dossier "Models" est vide.

Les pièces utilisées pour le calibrage de l'étiqueteur ne figurent dans aucun dossier mis en ligne[4].

Dès le début, il est donc impossible de répéter l'expérience. Toutefois, le lecteur intéressé par les détails de cette question, peut se reporter à notre annexe 5, il verra que Cafiero et Camps ont laissé en ligne un élément qui révèle un taux d'échec supérieur à 7% des mots. L'étiquetage est extrêmement médiocre. Normalement, l'expérience aurait dû s'arrêter là.

2. "Lemmatisation".

Sur chacun des textes sélectionnés, l'étiqueteur découpe les mots ("tokens") et attache à chacun d'eux une étiquette (son entrée de dictionnaire et sa catégorie grammaticale) que Cafiero et Camps nomment "lemmes". Du fait des trois expériences (exploratoire, principale, contrôle), il devrait figurer en ligne 2*3 groupes de pièces (textes seuls en entrée de l'étiqueteur et textes étiquetés en sortie).

Dans le Data_S2, le premier jeu de données (les textes seuls soumis à l'étiqueteur) manque totalement[5].

---

[4] Le 3 décembre 2019, nous les avons demandées à M. Gabay. Le 9 décembre 2019, il a répondu que "tout est en ligne" et qu'il n'a rien de plus à nous communiquer (refus réitéré à deux reprises). L'origine de ces pièces serait-elle inavouable ? En tous cas, les lecteurs sont privés d'un élément décisif d'appréciation.

[5] Autrement dit, il est impossible de refaire tourner l'étiqueteur ou tout autre programme de ce genre. De plus les auteurs ont effectué un nombre important d'opérations (par exemple, ils auraient retiré toutes les didascalies). Comment contrôler la qualité de ces opérations et l'intégrité des textes utilisés ? Là encore : ces textes n'étaient-ils pas montrables ?

Pour les textes étiquetés : dans le dossier "DataS2_labelled-corpus", on trouve trois dossiers mais l'un d'eux est vide. Les deux restants sont : "control" et "main". La phase "exploratoire" – censée vérifier l'efficacité de la méthode – est totalement manquante.

A cette étape également,

- des éléments essentiels sont à nouveau "oubliés",

- il est impossible de contrôler et de répéter les expériences des auteurs, à commencer par la plus importante aux yeux de tout usager sérieux (qui souhaite évidemment savoir si les essais sont concluants avant de s'embarquer à l'aventure).

3. Tri et indexation.

Rappel : les auteurs ont choisi d'étudier six variables : lemmes, mots ("word forms"), rimes, mots-outils ("function words"), affixes, POS-Ngrams (groupes de trois étiquettes consécutives par exemple, "nom+adjectif+verbe").

Ces 6 variables vont maintenant être mesurées. En effet, à l'issue de l'étiquetage, on connaît les mots employés dans chaque texte et leurs vocabulaires (les étiquettes). Sur chaque fichier étiqueté, un indexeur calcule le nombre de fois que les caractères recherchés apparaissent dans chaque texte. Les résultats sont consignés dans des tableaux (index) contenant en ligne les différentes attestations du caractère (par exemple, les mots), et en colonne le nombre de ses apparitions dans les textes du corpus considéré. Etant donné que Cafiero et Camps ont choisi d'étudier six variables et qu'il y a trois corpus, on devrait trouver en ligne : 6*3 tableaux d'indexation.

Dans le dossier Data_S3, on trouve 8 de ces tableaux (de "control-lemmas" à "explorat-words"). Il en manque donc 10. Lesquels ?

- les six tableaux du "main" ont été "oubliés". On ne dispose donc d'aucun élément pour contrôler les données ayant servi à l'expérience qui, d'après les auteurs, démontre que Molière est bien l'auteur de ces pièces. Le lecteur peut imaginer les raisons de cette étrange disparition et en tirer les conclusions.

- 2 caractères ont été totalement "oubliés" : les mots-outils et les affixes. On ne dispose donc d'aucun élément pour vérifier la véracité des dires des auteurs concernant ces deux variables. Or, d'après eux, il s'agirait des variables les plus discriminantes en matière d'attribution d'auteur (spécialement les mots-outils). Là encore, le lecteur peut imaginer les raisons de cette absence et en tirer les conclusions.

4. Le calcul des distances.

Comme indiqué dans leur article, à l'aide des index obtenus à l'étape précédente, Cafiero et Camps ont calculé des distances séparant chaque couple de pièces au sein des 3 corpus avec les six variables mesurées à l'étape précédente.

Les distances utilisées auraient été "MinMax" et "Burrows". Dans l'article, très peu d'informations sont fournies sur ces calculs et il n'y a aucun chiffre, contrairement à l'habitude. Usuellement, on indique au moins les couples les plus proches, les plus éloignés, les distances moyennes et les écarts-types. On recherchera en vain ces renseignements dans l'article et,

dans les documents en ligne, il n'y a aucune matrice de distances.

Pourquoi cette absence ?

Ces matrices ne sont pas volumineuses. Par exemple, le principal corpus contiendrait 37 pièces. Pour chaque caractère, cela donnerait deux petits tableaux de 37 lignes par 37 colonnes. Il n'y a donc aucun obstacle pratique pour les mettre en ligne.

Dans notre première réponse (mise en ligne le 27 novembre 2019) nous avons signalé cette étrange absence. Depuis lors, rien n'a changé, ce qui aurait certainement été le cas s'il s'agissait d'un "oubli" malencontreux.

Cette absence est donc délibérée.

Le lecteur se trouve privé de l'information décisive qui, seule, lui permettrait de contrôler l'exactitude de la dernière étape et notamment l'attribution à Molière des pièces de Molière…

5. La classification et l'attribution

C'est la dernière étape d'une attribution d'auteur. A partir des matrices de distances, un programme de classification, recherche le meilleur classement possible en groupant les couples les plus proches et en formant des groupes de textes les plus homogènes possibles et les plus contrastés entre eux. Ces opérations de groupement ("clustering") sont retracées dans des "dendrogrammes" (comme ceux reproduits dans les figures 1 et 2) : plus les traits horizontaux reliant les pièces entre elles sont situés bas, plus leur proximité est grande, ce qui suggèrerait un même auteur. En dessous de ces figures, les auteurs proclament que leurs classifications rejoignent "généralement" les auteurs supposés. Cependant, les noms des pièces et des auteurs sont si petits qu'il est impossible de vérifier l'information. De plus, ces noms ne figurent pas en légende des graphiques, contrairement aux usages en la matière.

Autrement dit, en l'absence de légendes sous les figures (et sans les matrices de distances ayant servi à tracer ces graphiques), il est impossible de contrôler quoi que ce soit. Il faudrait faire confiance et croire sur parole les conclusions des auteurs. Ce n'est pas l'attitude scientifique normale…

En fait, nous verrons plus bas que les auteurs ont laissé, dans l'annexe en ligne, les légendes de quelques figures. Ces légendes en disent plus long qu'ils ne le souhaiteraient sans doute (raison pour laquelle, elles ne figurent pas sous les graphiques comme cela est l'usage).

En tous cas, la conclusion de ce rapide tour d'horizon est sans appel.

Cafiero et Camps ont "oublié" de mettre en ligne les principaux éléments qui permettraient de contrôler leurs affirmations.

Sont-ils inexpérimentés, maladroits et désordonnés ? Cela soulève surtout de sérieux soupçons qui sont renforcés par beaucoup de choses intrigantes, notamment la sélection des pièces.

3. UNE ETRANGE SELECTION

Le bons sens suggère que, avant d'être appliquée aux cas douteux, une méthode d'attribution d'auteur doit être testée sur des textes d'origine sûre. Le bon sens indique aussi que les textes qui feront l'objet de l'expérience d'attribution doivent être exclus des tests préliminaires. Dans l'article de Cafiero et Camps, c'est le rôle des corpus "exploratoire" et de "contrôle". La composition de ces corpus réserve de nombreuses surprises.

Un étrange corpus "exploratoire"

Du fait des principes de bon sens énoncés ci-dessus, on s'attend à ce que le corpus destiné à démontrer l'efficacité de la méthode ne contienne aucun texte de Molière et de Corneille. D'après le tableau S2 de l'annexe en ligne, ce corpus contient 34 pièces (mais le titre du tableau indique 30) et l'on trouve dans cette liste :

- Molière : *Dom Garcie*, *Mélicerte*, *Sganarelle*. P. Corneille : *Don Sanche*, *Pulchérie*, *Tite and Bérénice*. Les auteurs sont donc sûrs de savoir qui a écrit ces pièces ? L'article ne donne aucune explication. On ne comprend pas (ou plutôt on comprend trop bien) pourquoi ces pièces – ayant déjà servi dans la phase de test - sont sorties de l'expérience principale destinée à démontrer que Molière est un "grand auteur" sans lien avec P. Corneille. Ces pièces gênaient la "démonstration".

- Autre étrangeté : *Ragotin* de La Fontaine. Cafiero et Camps ignorent que cette pièce a été jouée et publiée sous le nom de… Champmeslé. Cet acteur de la troupe de l'hôtel de Bourgogne était un 'comédien poète' tout comme l'était Molière (et sa vie est très similaire à celle de Molière). Après la mort de Champmeslé et de La Fontaine, un libraire hollandais a republié cette pièce sous le nom de La Fontaine sans expliquer les raisons de cette attribution posthume. En tous cas, cette pièce n'a rien à faire dans un corpus d'œuvres "certaines". Ou alors, puisque Champmeslé est un comédien poète exactement semblable à Molière, il faut mettre cette pièce sous son nom.

En fait, la composition de ce corpus exploratoire était peut-être différente comme le suggère le tableau S5 dans la même annexe en ligne. Ce document fournit la légende des figures 1 à 3 dans l'article. Examinons la première partie de ce tableau (légende de la figure 1) portant sur cette expérience

"exploratoire". Il apparaît que la calibration de la méthode d'attribution d'auteur a été faite en utilisant non pas seulement trois pièces présentées par Molière (comme indiqué plus haut en fonction du tableau S2 dans l'annexe en ligne) mais sur douze. Par ordre alphabétique : *Amphitryon*, *le Dépit amoureux*, *Dom Garcie*, *l'Ecole des femmes*, *l'Ecole des maris*, *l'Etourdi*, *les Fâcheux*, *les Femmes savantes*, *Mélicerte*, *le Misanthrope*, *Sganarelle*, *le Tartuffe*). De même, on trouve onze pièces de P. Corneille et non pas seulement les 3 listées dans le tableau S2.

Les 12 premiers groupes du tableau S5 (annexe en ligne) – l'expérience préliminaire - comptent en tout 72 pièces (et non pas 34) : Boursault a 7 pièces (au lieu de 6 listée dans S5) ; P. Corneille : 11 (au lieu de 3) ; T. Corneille : 11 (au lieu de 1) ; Molière : 12 (au lieu de 3) ; Ouville : 2 (au lieu de 3) ; Rotrou : 4 (au lieu de 0) et Scarron : 7 (au lieu de 1).

De plus, dans les 4 fichiers censés présenter les index utilisés au cours de cette phase préliminaire (Data_S3 : explorat-lemma, explorat-POS3gr, explorat-rhymes, explorat-words), on découvre que cette première phase a porté en réalité sur… 85 pièces[6].

Encore plus fort : dans le dossier "DataS2_labelled-corpus\ xml_main" censé contenir les pièces étiquetées ayant fait l'objet de l'expérience principale ("main"), on retrouve ces 85 pièces plus une !

Résumons : pour cette expérience préliminaire, le texte de l'article dit "a large sample", dans l'annexe en ligne, le titre du tableau S2 annonce 30 pièces ; mais ce même tableau en liste 34 ; la classification automatique aurait porté sur 72 pièces mais les données servant à cette classification sont issues de 85 ou 86 pièces qui, toutes sauf une, auraient servi également à l'expérience principale…

Les auteurs sont pour le moins légers.

Il est impossible de vérifier quelque chose dans un tel chaos.

En fait, tout s'éclaire en considérant que les trois expériences sont une fable. Le corpus "principal" (main) – celui qui sert à attribuer les pièces de Molière - était déjà présent en entier dans les 85 (ou 86) pièces de l'expérience préliminaire telle qu'elle a été réalisée réellement. Autrement dit, la calibration de la méthode ne visait pas à vérifier l'efficacité de celle-ci mais à repérer les pièces qui allaient dans le sens voulu - démontrer que Molière est un grand auteur - et à éliminer les autres, afin de ne présenter que des succès dans le prétendu "main".

Cela explique aussi pourquoi, lorsque les auteurs consentent à mettre en ligne quelques données – qui ont été présentées plus haut -, ces données vont toujours par paires et jamais par trois !

En définitive, Cafiero et Camps ont retiré beaucoup de pièces. C'est pourquoi l'expérience "principale" est entachée de nombreux oublis.

[6] Boursault 8, Chevalier, 8, P. Corneille 12, T. Corneille 11, Donneau de Visé 6, Dorimond 3, Gillet de Tessonerie 3, La Fontaine 3, Molière 12, Ouville 3, Quinault 3, Rotrou 4, Scarron 8.

Des "oublis" dans l'expérience cruciale

Parmi les 37 pièces prétendument soumises à cette épreuve cruciale, on trouve :
- 9 pièces présentées par Molière alors qu'il y en a 33.
- 8 par P. Corneille alors qu'il en a présenté 33 (34 avec *Psyché*).
- 10 par T. Corneille alors que 37 de ses pièces sont disponibles.
etc., etc.

Comme, selon Cafiero et Camps, leur expérience démontre que Molière a bien écrit *toutes* ses pièces, cela signifierait que les 24 sur lesquelles ne porte pas l'expérience sont incontestables ? Parmi celles-ci toutes les pièces en prose, notamment les plus célèbres et les plus jouées (*Les précieuses ridicules*, *Dom Juan*, *l'Avare*, *le Bourgeois gentilhomme*, *le Malade imaginaire*, etc.). Dans les pièces en vers, l'expérience principale "omet" : *Sganarelle* (1660), *Dom Garcie* (1661), *Mélicerte* (1666), *les Amants magnifiques* (1670), *Psyché* (1671).

Psyché a été le plus grand succès de Molière (plus de 70 représentations sans interruption, sous le seul nom de Molière et une recette supérieure au chiffre d'affaire annuel de la troupe). Six mois après ce triomphe, la pièce a été publiée avec le seul nom de Molière sur la couverture et la page de garde mais avec un avertissement du libraire (l'éditeur) en page suivante :

"Le libraire au lecteur

Cet ouvrage n'est pas tout d'une main. M. Quinault a fait les paroles qui s'y chantent en musique, à la réserve de la plainte italienne. M.de Molière a dressé le plan de la pièce, et réglé la disposition, où il s'est plus attaché aux beautés et à la pompe du spectacle qu'à l'exacte régularité. Quant à la versification, il n'a pas eu le loisir de la faire entière. Le carnaval approchait, et les ordres pressants du Roi, qui se voulait donner ce magnifique divertissement plusieurs fois avant le carême, l'ont mis dans la nécessité de souffrir un peu de secours. Ainsi, il n'y a que le prologue, le premier acte, la première scène du second et la première du troisième dont les vers soient de lui. M. Corneille a employé une quinzaine au reste ; et, par ce moyen, Sa Majesté s'est trouvée servie dans le temps qu'elle avait ordonné."

Psyché a donc été présentée au public sous le seul nom de Molière (comme toutes les autres). Cependant, dans ce cas, grâce à une indiscrétion (il y en a eu d'autres du vivant de Molière), la collaboration entre Corneille et Molière n'est pas discutable.

Les passages, que chacun est censé avoir écrits, sont clairement identifiés permettant de constituer trois textes (annexe 2 à ce document). Dès lors, ces passages fournissaient "l'épreuve décisive" qui aurait permis de juger de l'efficacité de la méthode de Cafiero et Camps : est-elle capable de distinguer

ces passages et de les attribuer correctement ? L'absence de *Psyché* dans leur corpus est un aveu ou une insigne maladresse.

De toute façon, l'absence de Psyché met par terre toute leur démonstration.

4. ENCORE DES ETRANGETES

D'autres éléments encore soulèvent de nombreux doutes sur la qualité du travail de Cafiero et Camps.

Etranges calculs

Premièrement, le tableau 2 en page 13 de l'article n'a que deux cadres : la phase "exploratoire" a totalement disparu. Elle était pourtant supposée démontrer l'efficacité de la méthode sur des cas sans problème. Cette absence n'est pas une inadvertance de plus. Puisque la phase préliminaire avait comme objectif de repérer les pièces qui ne "marchaient" pas, les taux de succès sont évidemment faibles et il fallait les faire disparaître.

Deuxièmement, les auteurs n'ont pas utilisé tout le matériel disponible. Par exemple, pour l'expérience décisive, ils ont utilisé 1789 "lemmes" sur les 8 781 qui constitueraient le vocabulaire du corpus (soit environ 1 "lemmes" sur 5). Est-il acceptable de prétendre avoir attribué un texte alors qu'on a pris en considération une proportion si petite de son vocabulaire ?

Plus important encore, le tableau 2 indique que, pour chaque variable, Cafiero et Camps ont sélectionné la proportion du caractère en fonction du taux de succès par rapport à… la conformité de la classification avec l'hypothèse : Molière n'est pas Corneille. Cette approche a été systématique. Par exemple, en raisonnant comme eux, mais en recherchant la conclusion inverse, leur tableau 2 indique qu'il suffisait de sélectionner l'analyse sur 75% des mots (word forms) pour démontrer que leur attribution échoue dans 43% des cas...

Troisièmement, sauf pour les "mots-outils" (function words), les "taux de succès" affichés –mais invérifiables – sont très bas (souvent inférieurs à 90%). Dès lors, leurs expériences sont entachés d'un taux d'erreur d'au moins 10%... ce qui est totalement inacceptable pour une attribution d'auteur décisive pour notre histoire littéraire.

Mais sur quoi portaient réellement ces calculs ?

Brouillard sur les mots-outils

Le lecteur attentif aura remarqué l'insistance de Cafiero et Camps à propos de l'efficacité remarquable des "mots-outils". D'où vient cette efficacité ? Selon eux, cette efficacité provient du fait que leur utilisation serait inconsciente et propre à chaque auteur.

Dans l'acception habituelle, les "mots-outils" sont les adverbes, les conjonctions, les déterminants, les prépositions et les pronoms. Puisque chaque mot est doté d'une étiquette, il était très facile de demander au programme de n'étudier que ces mots-là. Surprise : Cafiero et Camps ont choisi une autre façon de procéder. Ils ont sélectionné les 250 mots (word forms) les plus fréquents – pourquoi 250 ? impossible de le savoir -, puis ils ont sorti de ces 250 les mots supposées être des outils, avec le risque évident d'erreurs inhérent à toute intervention manuelle.

Une remarque préliminaire : tous les index pour cette phase principale sont absents des données mises en ligne. Il est donc impossible de vérifier ces opérations. Cependant, on peut le faire indirectement en se reportant à l'index des mots de la phase "exploratoire" (Data_S3 : explorat-words.cvs) qui, en fait, contient aussi tout le corpus "principal".

La table S4 (annexe en ligne) comporte 112 "mots" (et non 110 comme annoncé dans le texte). On y trouve certains adverbes comme *autant*, *aussi*, etc. mais les auteurs ont oublié, parmi ces adverbes présents dans les 250 : "aujourd'hui", "bien", "comment", "peut-être" et "toujours" (alors qu'il y a "jamais"). Parmi les déterminants, ils ont "omis" "l'" et "leurs".

Dans la présentation au-dessus de cette liste, il est indiqué que les "pronoms personnels" ont été enlevés (alors que tout le monde est d'accord pour les considérer comme des "outils"). Pourtant, on trouve dans la liste de Cafiero et Camps "s'" et "se" – qui sont des pronoms – beaucoup moins fréquents que "je", "tu", "il", "elle", "nous", "vous", "ils", "elles", "moi" qui sont effacés alors qu'ils sont présents dans les 250.

Une autre surprise survient quand on constate que, dans cette liste, figurent les verbes "être", "avoir" et "fait" que personne jusqu'à maintenant n'avait rangés dans les mots-outils. Dans ce cas, il aurait fallu trouver dans la liste Cafiero et Camps, d'autres conjugaisons comme "as", "faire", "fais", "fut", etc. qui sont présents dans les 250 mais qui ont été enlevés par les opérateurs pour des raisons mystérieuses.

Par-dessus tout, cette courte liste de 112 "mots-outils" contient une dizaine d'erreurs typographiques évidentes : "-ce", "-là", "-même", "qu", "c", "n", "l", "d", "s", "jusqu", au lieu de : ce, là, même, qu', c', n', l', d', s' jusqu' (qui sont aussi dans la liste de Cafiero et Camps). Ces étrangetés s'expliquent de trois manières.

Premièrement, la mauvaise qualité de certains textes qui n'ont pas été corrigés ni débarrassés de quelques didascalies (indications de scène) et des appels de notes (voir annexe 5).

Deuxième explication (valables pour –ce, -là et –même) : des erreurs de l'étiqueteur incapable de reconnaître les mots français (voir également notre annexe 5) : par exemple, "moi-même" est codé en deux mots ("moi +-même", avec le tiret devant le deuxième terme).

Troisièmement, ces signes étranges (qu, c, n, l, d, s, jusqu) ont été introduits en grand nombre pour forcer certains textes à se ranger là où ils doivent aller. En effet, dans l'index des mots du corpus "exploratoire"[7] (Data_S3 : explorat-words.cvs qui contient aussi le "main"), on ne trouve aucun "l", seulement trois "d" et deux "c". Dès lors, leur présence dans la liste des 250 est invraisemblable, sauf s'il y a eu manipulation des données. Ce serait pour cacher cette manipulation que les deux index des mots-clefs ont disparu des fichiers mis en ligne dans le dossier Data_S3. Il est donc impossible de savoir quels textes ont subi ces outrages mais on en devine le résultat.

Pourquoi toutes ces contradictions, erreurs et manipulations alors qu'il aurait été si simple d'utiliser les étiquettes pour repérer avec certitude les mots-outils ?

Bien sûr, on aura deviné que cela était nécessaire pour amener les pièces présentées par Molière à leur place, le plus loin possible de celles de P. Corneille. Le lecteur sceptique consultera le tableau 2 (p. 13). Les deux derniers cadres montrent que les auteurs ont utilisé tous les prétendus "mots-outils" (alors qu'ils n'ont retenu qu'une proportion plus ou moins faible des lemmes, mots, rimes, etc.). Autrement dit, la sélection a été faite selon l'habitude des auteurs : ajuster les paramètres de l'expérience jusqu'à ce que le but soit atteint.

En conséquence, le lecteur ne sera pas surpris par des déclarations comme celles-ci :

"The highest agglomerative coefficient is obtained for the analysis of function words. In this analysis, all plays signed by Molière are clustered together" (p. 3).

"The function words are deemed to reflect most accurately the less conscious variations in individual style" (sous la Figure 1, p 4).

Au passage, on notera le "all plays" que les auteurs écrivent à plusieurs reprises, notamment en conclusion, alors que, officiellement, ils n'attribuent que 9 pièces sur les 33 présentées par Molière.

En l'absence des matrices de distances, les mêmes soupçons pèsent sur l'ensemble des 5 autres variables qui sont mobilisées ou non, et toujours à proportion variable, selon qu'elles réussissent ou non à isoler ces 9 pièces prétendument de Molière, sans que jamais une explication claire soit donnée au lecteur.

Enfin, soulignons que l'orthographe de certains textes n'a pas été corrigée. Ils comptent de nombreux résidus (annexe 5), comme si l'on n'avait pas pris de le temps de les relire ni de consulter les index des mots et des lemmes dans lesquels ces coquilles sont pourtant bien visibles.

[7] aax5489_Data_file_S3.zip\DataS3_datasets-and-scripts\explorat-words.cvs.

Rappelons une règle de base de la statistique : la qualité des conclusions dépend de celle des observations du phénomène. Ici le peu d'information disponible est suffisant pour montrer que la qualité minimale requise dans ce genre de travail n'est pas au rendez-vous.

De même les résultats ne sont pas à la hauteur des prétentions des auteurs.

5. LE DERNIER TABLEAU (S5) REVELE UN ECHEC COMPLET

Quel devrait être les résultats obtenus par le classificateur automatique sur le corpus exploratoire (Fig 1) pour que la méthode puisse être considérée comme capable de reconnaître les textes de différents auteurs ?

On s'attend à ce que :

- chaque auteur sûr trouve sa place dans un groupe ;
- tous ces groupes sont homogènes (pas de groupe mélangeant plusieurs auteurs).

De plus, le même classement doit être obtenu pour les 6 variables.

En considérant ces standards, l'échec de Cafiero et Camps est complet. Il faut garder en mémoire que – comme montré dans le tableau 2 de l'article – ils ont choisi pour chaque variable la proportion la plus favorable à leurs hypothèses.

Deux fois plus d'échecs que de succès

Le tableau S5 (annexe en ligne) permet de mesurer l'efficacité de la méthode Cafiero et Camps – en l'absence de toute matrice des distances - et de corriger les omissions du tableau 2, notamment quant au taux de succès dans la première étape.

Le calcul se déroule ainsi : pour chacun des groupes énumérés dans ce tableau S5,

- le succès est atteint quand le groupe comprend un seul auteur. Il est accepté qu'un auteur figure dans plusieurs groupes à condition qu'il y soit seul à chaque fois. En cas de succès, on attribue au groupe une note égale au nombre de textes correctement classés.
- lorsque plusieurs auteurs sont mélangés dans un même groupe, la note de ce groupe est nulle.

Par exemple, c'est le cas du premier groupe dans le tableau S5 qui mélange trois auteurs : Boursault (*la Comédie sans titre*, *les Mots à la mode*, *le Portrait du peintre, la Satire des satires) ;* Donneau de Visé (*les Embarras de Godard*, *le Gentilhomme Guépin*, *les Intrigues de la loterie)* ; La Fontaine (*Climène).*

Le second groupe reçoit également une note nulle car il mélange 3 pièces présentées par Molière *(les Fâcheux, Mélicerte, Sganarelle)* avec une de T. Corneille *(Festin de Pierre),* une de La Fontaine *(Ragotin)* et une de Donneau de Visé *(la Cocue imaginaire).*

En revanche le troisième reçoit une note de 4 puisqu'il groupe 4 pièces d'un même auteur (Chevalier : *les amours de Calotin, l'Intrigue des carrosses, les Barbons amoureux, le Pédagogue amoureux)*, etc.

Le tableau ci-dessous présente les scores obtenus dans l'expérience préliminaire de Cafiero et Camps (qui était censée prouver l'efficacité de la mesure) pour les six variables (en colonne) et bpour chaque groupe, ou "cluster" (en ligne).

Tableau 1. Scores des groups lors de l'expérience préliminaire de Cafiero et Camps

Groupes	Lemmes	Rimes	Mots	Affixes	N-Grams	Mots-outis	Total
1	0	0	0	0	0	0	0
2	0	0	4	5	0	0	9
3	4	4	0	5	4	0	17
4	0	0	6	0	6	0	12
5	0	0	0	0	0	0	0
6	6	0	4	6	6	3	25
7	6	0	6	0	0	0	12
8	2	0	2	4	5	6	19
9	0	0	0	0	0	0	0
10	6	0	6	5	0	4	21
11	0	0	0	0	0	6	6
12	0	6	4	4	6	4	24
Total	24	10	32	29	27	23	145
Taux de succès %	33	14	44	40	38	32	34

Le corpus compte 72 pièces : c'est le score maximum pouvant figurer au bas de chaque colonne. Par exemple, avec les lemmes, 24 pièces ont été correctement classées, le taux de succès est de 24/72 = 33%.

Pas une variable ne franchit le seuil des 50%.

Puisque l'expérience considère successivement les six variables sur les 72 pièces, le score normal qu'une attribution d'auteur doit atteindre est de 72*6 = 432. En acceptant un taux d'erreur de 5% le score minimum est 410 (432*0,95). Cafiero et Camps en sont très loin : il y a deux fois plus d'échecs (287) que de succès (145).

Le numéro de la ligne est également à considérer car la plupart des classificateurs commencent par considérer les couples les plus proches. Dès lors, l'ordre des lignes donne une idée de la relative homogénéité des groupes classés avec succès. Cet ordre montre que la classification de Cafiero et Camps commence toujours par des échecs complets (première ligne égale à 0 pour toutes les variables) et récolte des succès dans les zones les moins sûres.

Enfin, on notera que le nombre des groupes est toujours de 12 quelle que soit la variable. Ce genre d'événement n'a aucune chance de se produire par hasard. Cafiero et Camps ont fixé une contrainte au classificateur : toujours trouver 12 groupes. De fait, cette expérience préliminaire couvrait 11 ou 12 auteurs (supposés). Si Corneille était la plume de l'ombre de Molière, alors il y avait 11 auteurs, et

sinon : 12. Cafiero et Camps ont décidé *a priori* : 12. C'est un exemple de leur manière de penser : ils savent le résultat des expériences avant même de les lancer.

Des auteurs introuvables

L'échec commence avec "Molière". Dans le tableau S5, *Amphitryon*, *Dom Garcie de Navarre*, *Le Dépit amoureux, les Fâcheux, Mélicerte, Sganarelle* ne sont jamais classés de manière stable et complète, sauf avec une variable (le lecteur a deviné laquelle : les mots-outils). Le cas le plus amusant est *Dom Garcie* (présentée par Molière). Avec 5 des six variables (lemmes, mots, rimes, n-grams, affixes), cette pièce est groupée 13 fois avec des pièces de P. Corneille, 12 fois avec celles de T. Corneille, 4 fois avec Rotrou, 1 fois avec Ouville et Quinault, et jamais avec une autre quelconque de Molière… Mais étrangement, lorsqu'on utilise les "mots-outils", *Dom Garcie* rentre au bercail dans un groupe contenant 7 autres pièces présentées par Molière. De fait, il ne s'agit pas de "mots outils" mais d'une liste élaborée spécialement pour obtenir cette heureuse issue.

D'autres pièces par le même auteur, comme celles de La Fontaine, Donneau de Visé ou Boursault, se trouvent un peu partout et rarement ensemble (jamais dans le cas de La Fontaine). Les pièces de T. Corneille ou de Quinault sont douées d'une extraordinaire mobilité. Elles se retrouvent avec pratiquement tous les autres auteurs (spécialement Molière) et rarement aux mêmes endroits selon la variable utilisée. En effet, non seulement les classements sont aberrants mais, de plus, ils varient en fonction de la variable.

Le tableau S5 démontre sans appel que la méthode de Cafiero et Camps est incapable de reconnaître : Boursault, Chevalier, Donneau de Visé, Gillet de la Tessonnerie, Pierre Corneille, Thomas Corneille, La Fontaine, Ouville, Quinault, Rotrou... et Molière ! Pas un seul n'échappe au naufrage.

Conclusion : l'expérience devait s'arrêter là.

Que font ces étranges expérimentateurs ?

Premièrement, ils effacent cette mauvaise nouvelle du tableau 2 (article page 13) : le cadre concernant l'expérience "exploratoire" a disparu et ils éliminent la légende sous les figures, spécialement la figure 1 (où le naufrage était impossible à masquer si cette légende avait été mise à la disposition du lecteur).

Deuxièmement, ils retirent toutes les pièces qui posent problème, soit 49 sur les 86 initialement retenues (près de 6 sur 10). Passent totalement à la trappe sans une explication : Boursault, Chevalier, Donneau de Visé, Gillet, La Fontaine, Ouville, Quinault.

Ne restent dans le tour final que Molière, P. et T. Corneille, Rotrou et Scarron.

Les œuvres des trois premiers sont massivement amputées (pour Molière, il n'en reste que 9 sur 33).

Rotrou (1609-1650) et Scarron (1610-1660) ne peuvent pas avoir écrit les pièces présentées par Molière (entre 1660 et 1673)… puisqu'ils étaient morts !

Miracle final

Avec ces amputations drastiques et force ajustements et nettoyages… ça marche. Ou du moins, les auteurs parviennent enfin là où ils voulaient arriver depuis le tout début : 9 des 33 pièces présentées par Molière seraient enfin groupées ensemble à l'écart de P. Corneille (mais comment vérifier puisque nous n'avons pas les matrices de distances ?).

Les autres auteurs continuent à se mélanger ? Aucune importance. Au contraire, voyez comme Molière est singulier dans ce XVIIe où tout le monde écrivait de la même manière.

L'article affirme que le "corpus de contrôle" ne pose pas de problème semblable… excepté pour Dancourt et Régnard. Dancourt (Florent Carton, un comédien poète comme Champmeslé et Molière) a présenté plus de 40 pièces et Régnard une douzaine. Ils ont dominé la scène théâtrale à la fin du XVIIe et au début du XVIIIe. La méthode de Cafiero et Camps échoue donc à nouveau sur ce cas emblématique. Mais ce n'est pas un problème pour eux : ils ont toujours une explication "ad hoc" sous la main pour expliquer l'inexplicable. Ici, Dancourt a joué un rôle dans une pièce de Régnard, la femme de Dancourt était une actrice, elle connaissait aussi Régnard, etc. Les explications des pages 3 et 6 sont de la même eau.

Cafiero et Camps ont un bagout extraordinaire, mais ils n'ont pas conscience que la multiplication de ces rustines détruit leur présupposé central selon lequel chaque auteur aurait des caractéristiques stylistiques particulières "inconscientes" que leur méthode miracle serait capable de détecter.

5. CONCLUSIONS

Une cascade d'anomalies

Récapitulons les principales anomalies :

- Disparition des textes ayant servi au calibrage de l'étiqueteur,
- Dissimulation des taux d'échec réels de l'étiqueteur,
- Disparition des textes en entrée de l'étiqueteur,
- Disparition des textes étiquetés ayant servi à l'expérience principale sur Molière,
- Disparition des 6 index concernant l'expérience principale,
- Disparition des 3 index portant sur les mots-outils,
- Disparition de toutes les matrices de distances ayant servi à la classification et à l'attribution d'auteur,
- Absence de légende sous les figures censées présenter la classification et justifier l'attribution d'auteur,
- Impossibilité de savoir précisément sur quels textes a porté la prétendue expérience exploratoire, censée démontrer l'efficacité de la méthode,
- Sélection arbitraire dans les œuvres et mise sous le boisseau d'un grand nombre d'entre elles, notamment : *Psyché* et *Dom Garcie*.
- Incohérences des calculs et cascades d'étrangetés, notamment sur les "mots-outils".
- Conclusions à rebours des résultats réels.

Tout cela éveille plus qu'un doute sur la bonne foi des auteurs et sur le sérieux de leur travail. En tous cas, ce papier n'avait aucune chance auprès d'une revue sérieuse spécialisée en statistiques appliquées ou en "humanités digitales". Voilà pourquoi Cafiero et Camps se sont adressés à un site en "open access" : ils étaient sûrs de ne pas être relus.

Une mauvaise méthode

Au-delà de cette anthologie de tout ce qu'il ne faut pas faire dans un travail de recherche, la principale erreur réside dans le fait d'avoir appliqué cette méthode aux seules pièces de théâtre de l'ancien régime. Les auteurs ne semblent pas avoir réalisé que le prérequis essentiel pour des outils d'attribution d'auteur est d'être capable de survivre à un grand nombre de tests sévères dans des conditions très variées. Au contraire, nous avons ici une méthode "ad hoc" imaginée seulement pour montrer que Molière est Molière. En définitive, toutes les manipulations n'auront servi à rien puisque l'échec final est bien visible :

La méthode Cafiero et Camps est incapable de reconnaître les auteurs des pièces du XVIIe siècle y compris celui des pièces présentées par Molière.

Corneille a bien écrit 19 des 33 pièces présentées par Molière

La conclusion de notre étude de 2001 n'est pas réfutée. Au contraire, le peu d'information donnée par Cafiero et Camps (comme l'étrange proximité de *Dom Garcie* avec les pièces contemporaines de P. Corneille), tout ce qu'ils ont maladroitement caché (comme *Psyché*) et toutes leurs manipulations renforcent nos conclusions d'il y a 18 ans.

Voici ces deux conclusions inchangées aujourd'hui.

Premièrement, *Dom Garcie* et *Psyché* (présentées par Molière respectivement en 1661 et 1671) sont les deux sœurs des 10 tragédies de P. Corneille entre 1659 et 1674.

Deuxièmement, les deux comédies de P. Corneille – *le Menteur* (1643) et *la Suite du menteur* (1643) – sont les sœurs aînées de 17 parmi les 31 autres comédies présentées par Molière : *l'Etourdi* (1659), *le Dépit amoureux* (1659), *Sganarelle* (1660), *l'Ecole des maris* (1661), *les Fâcheux* (1661), *l'Ecole des femmes* (1662), *la Princesse d'Elide* (1664), *le Tartuffe* (1664), *Dom Juan* (1665), *le Misanthrope* (1666), *Mélicerte* (1666), *Amphitryon* (1668), *l'Avare* (1668), *les Amants magnifiques* (1670), *le Bourgeois gentilhomme* (1670), *les Femmes savantes* (1672), *le Malade imaginaire* (1673).

Comme nous l'écrivions pour les mathématiciens du CNRS en 2011 :

«Il reste deux solutions. Molière aurait éprouvé un fort mimétisme envers P. Corneille pendant toute sa vie créatrice. Ou bien Corneille et Molière auraient collaboré selon la procédure usuelle à cette époque et consistant à faire endosser certaines comédies par un « comédien poète ». La convergence de plusieurs indices statistiques - distances, classifications, combinaisons des verbes usuels, sens des principaux mots, longueurs et structure des phrases -, avec de nombreux indices historiques, rend possible une conclusion en faveur de la seconde solution. »[8]

De lourdes questions

Au-delà des questions apparemment techniques posées au début de cette conclusion, Cafiero et Camps font face à trois lourdes interrogations.

- D'où viennent les données utilisées pour l' "apprentissage" de leur étiqueteur. Si elles sont l'œuvre de Cafiero, Camps et Gabay, pourquoi refusent-ils de les communiquer ? Si c'est le travail de quelqu'un d'autre, ce nom aurait dû figurer dans les remerciements ("aknowledgments") placés à la fin de l'article et cette personne aurait dû avoir copie de l'article avant sa parution. Si ces règles simples ne sont pas respectées, il n'y a plus de coopération scientifique possible car personne ne peut admettre que son travail soit ainsi pillé et dénaturé.

---

[8] CNRS, *Images des mathématiques. La recherche mathématique en mots et en images*. 28 mars 2011. (https://images.math.cnrs.fr/_Labbe-Dominique_.html). Pour le rôle des comédiens poètes dans la naissance de l'industrie moderne du spectacle ; https://www.researchgate.net/publication/333236909.

- Pourquoi avoir fait appel à une officine de bas-étage dirigée par une personne qui a été chassée de son université à la suite d'une enquête fédérale ? Pourquoi avoir payé si cher ? Ne savent-ils pas que ces pratiques sont condamnées par la communauté scientifique, que leur travail, publié dans de telles conditions, perd toute valeur aux yeux des gens sérieux ?

- Ont-ils clairement indiqué, aux responsables de leurs équipes de recherche respectives, la véritable nature du "journal" dans lequel leur papier a été publié moyennant finances ? Et sinon pourquoi ont-ils tu cette information capitale ?

Pourquoi avoir affirmé au *Journal du CNRS*[9], à *Pour la Science*[10], à la presse écrite et radio-télévisée que *Science Advances* est un "journal scientifique", voire une "revue prestigieuse" alors qu'ils ont déboursé une grosse somme pour être publiés sans relecture par les pairs, ce qui est contraire aux règles de la recherche.

Enfin, un certain nombre de gens ne se sont pas renseignés sur la nature de *Science Advances* et n'ont pas jeté un coup d'œil au papier de Cafiero et Camps avant d'engager dans cette galère le prestige de leurs organisations[11]. Il s'agit notamment de (liste non exhaustive) :

CNRS, Ecole des Chartes, Université de Paris, Université de Neuchâtel[12],

AFP, *Journal du CNRS*, *Pour la Science*, *Le Point*[13], *le Parisien*[14], *les Echos*[15], *le Temps*[16], *France Culture*[17], *France Inter*[18], *France Info*[19], *TV5 Monde*[20], etc.

Qui aura l'honnêteté de reconnaître publiquement s'être fait duper ?

---

[9] https://lejournal.cnrs.fr/articles/moliere-est-bien-lauteur-de-ses-pieces

[10] https://www.pourlascience.fr/sd/linguistique/moliere-est-bien-lauteur-de-ses-oeuvres-18445.php

[11] Signalons toutefois quelques articles nuancés – comme celui du *Figaro*. Un seul journaliste nous a contacté et a essayé de faire un travail équilibré (https://www.huffingtonpost.fr/entry/corneille-a-t-il-ecrit-les-pieces-de-moliere-deux-chercheurs-veulent-clore-le-debat_fr_5dde43d6e4b0d50f32999e3d).

[12] Pour ces quatre institutions, voir note 2 au début de ce dossier.

[13] https://www.lepoint.fr/culture/definitivement-corneille-n-a-pas-ecrit-les-pieces-de-moliere-27-11-2019-2350130_3.php

[14] http://www.leparisien.fr/culture-loisirs/corneille-n-a-pas-ecrit-les-pieces-de-moliere-assure-une-etude-28-11-2019-8203911.php

[15] https://www.lesechos.fr/idees-debats/sciences-prospective/moliere-etait-bien-moliere-nen-deplaise-aux-sceptiques-1154547

[16] https://www.letemps.ch/sciences/science-tranche-corneille-na-ecrit-oeuvres-moliere

[17] https://www.franceculture.fr/theatre/deux-chercheurs-prouvent-que-corneille-na-pas-ecrit-les-pieces-de-moliere

[18] https://www.franceinter.fr/corneille-a-t-il-ecrit-certaines-pieces-de-moliere-l-intelligence-artificielle-relance-le-debat

[19] https://www.francetvinfo.fr/culture/spectacles/theatre/c-est-bien-moliere-qui-a-ecrit-ses-pieces-pas-corneille-tranche-une-etude-linguistique_3722281.html

[20] https://information.tv5monde.com/culture/corneille-n-pas-ecrit-les-pieces-de-moliere-tranche-une-etude-334306

Remerciements

John L. Klause, Cyril Labbé, Thomas Merriam, Corinne Rossari et Jacques Savoy nous ont aidé à rédiger cette réponse.

Edward Arnold, Guy Bensimon, Jean-Guy Bergeron, Mathieu Brugidou, Pierre Hubert, Cyril Labbé, Nelly & Jean Leselbaum, Xuan Luong, Thomas Merriam, Denis Monière, Jacques Picard, André Pibarot, Mathieu Ruhlman et Jacques Savoy ont collaboré à la mise au point de la méthode d'attribution d'auteur.

Les sites Gallica et Théâtre classique nous ont permis de constituer une bonne partie du corpus des pièces du XVIIe et XVIIIe siècles.

Les logiciels de lemmatisation et d'analyse lexicométriques ont été réalisés par Cyril et Dominique Labbé.

Toutes nos recherches ont été réalisées sans aide publique ni mécénat.

Annexe 1

Les pièces présentées par Molière

		Création	Genre	Longueur (mots)
01	La jalousie du barbouillé*	Before 1659	Comédie prose	3 501
02	Médecin volant*	Before 1659	Comédie prose	3 876
03	L'étourdi	1659	Comédie vers	18 671
04	Dépit amoureux	1659	Comédie vers	16 242
05	Précieuses ridicules*	1660	Comédie prose	6 648
06	Sganarelle*	1660	Comédie verse	6 042
07	Dom Garcie*	1661	Comédie héroïque vers	17 049
08	L'école des maris	1661	Comédie vers	10 536
09	Les fâcheux	1661	Comédie vers	7 922
10	L'école des femmes	1662	Comédie vers	16 625
11	Critique de l'école des f.*	1663	Comédie prose	8 610
12	L'impromptu*-	1663	Comédie prose	7 168
13	Mariage forcé*	1664	Comédie prose	6 058
14	Princesse d'Elide*	1664	Comédie vers	11 333
15	Le Tartuffe	1664	Comédie vers	18 271
16	Dom Juan*	1665	Comédie prose	17 452
17	L'amour médecin*	1665	Comédie prose	6 147
18	Le Misanthrope	1666	Comédie vers	17 180
19	Médecin malgré lui*-	1666	Comédie prose	9 317
20	Mélicerte*	1666	Comédie vers	5 540
21	Coedy pastorale*	1667	Comédie vers libres	732
22	Le sicilien*	1667	Comédie prose	5 375
23	Amphytrion	1668	Comédie vers libres	15 117
24	Georges Dandin*	1668	Comédie prose	11 009
25	L'avare*	1668	Comédie prose	21 033
26	M. de Pourceaugnac*	1669	Comédie prose	11 803
27	Amants magnifiques*	1670	Comédie vers & prose	11 983
28	Bourgeois gentilhomme*	1670	Comédie prose	17 132
29	*Psyché* (see below appendix 2)*	*1671*	*Comédie héroïque vers*	*16 282*
30	Fourberies de Scapin*	1671	Comédie prose	14 245
31	Comtesse d'Escarbagnas*	1671	Comédie prose	5 564
32	Femmes savantes	1672	Comédie verse	16 863
33	Malade imaginaire*	1673	Comédie prose	19 919

* Cafiero et Camps ont enlevé de leurs expériences 24 plays des 33 pièces présentées par Molière (pièces marquées d'un astérisque).

Annexe 2

29. *Psyché* (1671) parts respectives de Corneille, Molière et Quinault, d'après l'éditeur

	Authors	Genre	Length (tokens)
29.1	Corneille	vers	10 067
29.2	Molière	vers	4 816
29.3	Quinault	vers	1 399

Annexe 3
Les pièces de P. Corneille (toutes les pièces sont en vers)

		Creation	Genre	Length (tokens)
01	Mélite	1630 ?	Comédie	16 690
02	Clitandre*	1631	Tragi-Comédie	14 402
03	La Veuve	1631	Comédie	17 661
04	La Galerie du Palais	1632	Comédie	16 140
05	La Suivante	1633	Comédie	15 160
06	Comédie des Tuileries*	1634	Comédie	3 627
07	Médée*	1635	Tragédie	14 269
08	La Place Royale	1634	Comédie	13 801
09	L'illusion comique	1636	Comédie	15 428
10	Le Cid*	1636	Tragi-Comédie	16 677
11	Cinna*	1639	Tragédie	16 126
12	Horace*	1640	Tragédie	16 482
13	Polyeucte*	1641	Tragédie	16 472
14	Pompée*	1642	Tragédie	16 492
15	Le Menteur	1642	Comédie	16 653
16	Le Menteur (suite)	1643	Comédie	17 675
17	Rodogune*	1644	Tragédie	16 842
18	Théodore*	1645	Tragédie	17 121
19	Héraclius*	1647	Tragédie	17 433
20	Andromède*	1650	Tragédie	15 514
21	Don Sanche*	1650	Comédie héroïque	16 947
22	Nicomède*	1651	Tragédie	16 923
23	Pertharite*	1651	Tragédie	17 121
24	Œdipe*	1659	Tragédie	18 618
25	Toison d'Or*	1661	Tragédie	20 343
26	Sertorius*	1662	Tragédie	17 675
27	Sophonisbe*	1663	Tragédie	16 858
28	Othon*	1664	Tragédie	16 971
29	Agésilas*	1666	Tragédie	18 227
30	Atilla*	1667	Tragédie	16 788
31	Tite et Bérénice*	1670	Comédie héroïque	16 697
32	Pulchérie*	1672	Tragédie	16 630
33	Suréna*	1674	Tragédie	16 545

* Cafiero et Camps ont retire de leurs experiences (25 des 33 pièces de P. Corneille (pieces marquees d'une astérisque.

Annexe 4
Notre méthode d'attribution d'auteur

Dès les premières lignes de Cafiero et Camps, une chose est sûre, ils ne nous ont pas lus.

Par exemple, dans la première colonne de la page 1, ils affirment que notre méthode donne un poids plus grand aux mots de fortes fréquences. S'ils avaient pris la peine de lire notre article (Labbé, Labbé 2001), ils auraient vu que la distance intertextuelle prend en compte tout le vocabulaire des textes et donne à chaque mot exactement son poids dans chaque texte sans aucune déformation (ce qui n'est pas le cas des distances utilisées par Cafiero et Camps). Un autre exemple : dans leur article (première colonne de la page 2), ils nous font dire que les *Précieuses ridicules* (présentées par Molière) sont de P. Corneille (c'est d'ailleurs la seule pièce qu'ils citent à notre propos). Pas de chance ! Dans notre article de 2001, cette pièce n'est pas attribuée à P. Corneille et nous n'avons pas varié depuis.

Cafiero et Camps veulent toujours trop prouver…

S'ils nous avaient lus– ou s'ils avaient accepté de discuter avec nous comme il est d'usage dans la communauté scientifique – ils auraient vu que, depuis 1999, notre méthode a été couronnée de succès et qu'elle a passé les tests les plus rigoureux. Par exemple, un test en aveugle avec deux chercheurs britanniques (Labbé 2007) ou l'identification de la plume de l'ombre qui a servi deux Premiers ministres québécois (Monière, Labbé 2006).

Après un test et un examen approfondi de nos formules, les mathématiciens du CNRS ont publié notre méthode dans leur journal en ligne (*Images de mathématiques :* Labbé, Labbé 2011).

Le dernier succès en date est l'identification de la plume qui se cache dans l'ombre d'E. Ferrante (Savoy 2018). Dans ce dernier cas, la conclusion a été validée par les résultats de sept autres équipes de recherche à travers le monde (Tuzzi 2018; Tuzzi, Cortelazo 2018).

Plus important, cette méthode est utilisée pour repérer certaines fraudes dans les publications scientifiques (Labbé, Labbé 2012) avec des succès qui ont été salués à trois reprises par la revue *Nature* (Van-Norden 2014 ; Phillips 2017 ; Byrne 2019). Ces outils ont été intégrés dans le processus de décision du deuxième groupe mondial d'édition, Springer-Mac-Millan (Minh Tien 2018). Nous avons placés ces algorithmes dans le domaine public et ils sont utilisés par les principaux éditeurs scientifiques mondiaux. Récemment, ils ont permis de détecter une fraude massive dans la recherche sur le cancer (Byrne, Labbé 2016), ce qui a valu à l'une des membres de notre réseau de recherche (J. Byrne) le titre de "Scientist of the Year 2017" délivré par le journal *Nature.*

En d'autres mots, nos outils d'attribution d'auteurs ont résisté à tous les tests depuis plus de 20 ans, ils sont efficaces à travers le monde… pourquoi pas sur Molière ?

Références

J. Byrne, We need to talk about systematic fraud, *Nature*, 06 February 2019.
https://www.nature.com/articles/d41586-019-00439-9

J. Byrne, C. Labbé, Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines, *Scientometrics*, 28 December 2016.
http://membres-lig.imag.fr/labbe/Publi/ByrneLabbe2016.pdf

C. Labbé, D. Labbé, Inter-Textual Distance and Authorship Attribution Corneille and Molière, *Journal of Quantitative Linguistics*, 8-3, p. 213-231, December 2001 /
https://www.researchgate.net/publication/32222053

C. Labbé, D. Labbé, La classification des textes. Comment trouver le meilleur classement possible au sein d'une collection de textes ? *Images des mathématiques. La recherche mathématique en mots et en image,* 28 March 2011.
http://images.math.cnrs.fr/La-classification-des-textes.html

C. Labbé, D. Labbé, Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*, 22 June 2012.
https://www.researchgate.net/publication/257663143

D. Labbé, Experiments on Authorship Attribution by Intertextual Distance in English, *Journal of Quantitative Linguistics*, 14(1), p. 33-80, April 2007
https://www.researchgate.net/publication/32222131

N. Minh Tien. *Detection of automatically generated texts*. Ph-D Thesis. Grenoble-Alpes University. 3-04-2018.
https://tel.archives-ouvertes.fr/tel-01919207

D. Monière, D. Labbé. L'influence des plumes de l'ombre sur les discours des politiciens. In Condé Claude et Viprey Jean-Marie. Actes des 8e Journées internationales d'Analyse des données textuelles. Besançon : 19-21 avril 2006, II, p. 687-696.
https://www.researchgate.net/publication/237648469

N. Philipps. Online software spots genetic errors in cancer papers. *Nature*. 20 November 2017.
https://www.nature.com/news/online-software-spots-genetic-errors-in-cancer-papers-1.23003

J. Savoy, *Elena Ferrante Unmasked*. Neuchatel University, September 2017.
https://www.researchgate.net/publication/320131096_Elena_Ferrante_Unmasked

A Tuzzi, *It Takes Many Hands to Draw Elena Ferrante's Profile*. Padova University Press, 2018.
https://www.researchgate.net/publication/326723646_It_Takes_Many_Hands_to_Draw_Elena_Ferrante's_Profile

A. Tuzzi, M. Cortelazo, What is Elena Ferrante? A comparative analysis of a secretive bestselling Italian writer, *Digital Scholarship in the Humanities*, Volume 33, 3, p 685–702, September 2018

R. Van Noorden, Publishers withdraw more than 120 gibberish papers, *Nature*, 24 February 2014.
https://www.nature.com/news/publishers-withdraw-more-than-120-gibberish-papers-1.14763

Annexe 5
Qualité du découpage des mots et de l'étiquetage Cafiero-Camps-Gabay

Avant toute analyse de contenu – ou d'attribution d'auteur – se situe le traitement automatique des textes dont la portée est cruciale (puisque tous les calculs seront opérés sur les résultats de cette première phase). Il est donc nécessaire d'apprécier le sérieux de ce travail pour lequel Cafiero et Camps auraient bénéficié de l'aide de S. Gabay (Université de Neuchâtel).

Le dossier Data_S1 ne contient pas les fichiers étiquetés utilisés pour l'apprentissage de l'étiqueteur. Sans doute, leur origine était-elle inavouable ?

On trouve dans ce dossier un seul fichier consistant : "train" (ce qui pourrait vouloir dire "entrainement", "apprentissage"). Il s'agit d'un fichier de sortie de l'étiqueteur qui comporte 82 063 lignes et 70 852 mots étiquetés. Ce sont des extraits d'une vingtaine de pièces du XVIIe collés bout à bout, commençant par *La Veuve* (Pierre Corneille) et se terminant par la *Comtesse d'Orgueil* de Thomas Corneille.

Avant d'analyser ce document, rappelons ce qu'est la lemmatisation et ses enjeux.

1. Le traitement automatique des textes et ses enjeux

Ce traitement comporte deux opérations principales : le découpage des mots et leur "étiquetage". Mais auparavant, l'orthographe doit être soigneusement corrigée et standardisée.

Correction orthographique

La liste complète des "mots" présentées dans "explorat-words", comporte de nombreuses coquilles et des mots étranges comme : 232 guillemets comptés comme des "mots", "[" (49 fois) ; "]" (56 fois), "*" (5 fois)… et même "$", "2coutez", "5ᵉ", "6meu", etc.

Nous n'insistons pas sur point. Dans leur article, les auteurs reconnaissent ne pas avoir procédé à cette correction orthographique. On notera simplement que ces textes auraient mérité un traitement plus respectueux et que, du fait de l'absence d'une relecture soignée - les résultats finaux sont entachés d'une marge d'erreur non négligeable.

Découpage des mots.

Le mot (word form) est l'unité insécable du lexique de la langue française. Voici à l'aide de trois exemples simples, les principales difficultés auxquelles se heurte le découpage d'un texte français en "tokens" :

- "d'abord" est un seul mot (adverbe) malgré la présence de l'apostrophe qui, habituellement, sépare les mots comme dans "l'abord" (l' + abord). Autres exemples : aujourd'hui, c'est-à-dire, etc. ;
- "moi-même", "lui-même", etc. sont des mots uniques malgré la présence du tiret qui, habituellement, sépare les mots (comme dans "dois-je ?"). Dans le premier cas, le programme conserve le mot unique avec son tiret ; dans le second (dois-je), le tiret est éliminé (dois + je) alors que le tiret de ponctuation est conservé (comme celui au début de ce paragraphe). Le français comporte une multitude de mots composés (avec des tirets) qu'il ne faut pas découper ;
- l'espace n'est pas toujours un séparateur de mots. Par exemple, "parce que" (ou parce qu') est un seul mot.

Etiquetage des mots

En même temps qu'il découpe le texte en mots (word tokens), le programme rattache chacun de ces mots à une entrée du dictionnaire, c'est-à-dire un "mot vedette" – l'infinitif du verbe, le masculin singulier de l'adjectif, etc. – et à une catégorie grammaticale. Le vocable ("word type") est la

combinaison de cette entrée de dictionnaire et de la catégorie grammaticale (on dit aussi "lemme"). Cette opération est indispensable, car dans tout texte en français, au minimum un mot sur trois peut être rattaché à plus d'une entrée de dictionnaire (homographies). Par exemple, "garde" : verbe garder ou substantif masculin ou substantif féminin. ; "tout" : nom, pronom, adjectif indéfini, pronom ? etc.

Le programme qui découpe les mots et attache les étiquettes est appelé lemmatiseur ou étiqueteur ("tagger").

Ces opérations préliminaires sont cruciales. Sans elles, Cafiero et Camps n'auraient pas pu analyser les pièces de Corneille et de Molière ni réaliser leur prétendue attribution d'auteur. On comprend également que de la qualité de cette opération préliminaire dépend celle des conclusions.

Evaluation de l'étiqueteur et corrections

Comment évaluer le travail de l'étiqueteur ?

Normalement, nous devrions disposer des textes qui ont été utilisés pour l'étalonner. En effet, à chaque fois que celui-ci fait une erreur, deux explications sont possibles :

- les données d'étalonnage ne comportaient pas ce problème (ou contenaient l'erreur). Dans ce cas, il faut reprendre le processus en intégrant ce cas dans le jeu d'apprentissage.
- la règle est présente dans le jeu d'apprentissage et, si on ne peut intervenir sur le code source du programme, il faut prévoir de poser des "rustines" après le passage de l'étiqueteur. S'il y a plusieurs cas de ce genre, l'étiqueteur sera mis au rencart.

Faute des données d'apprentissage, dans les dossiers mis en ligne, il reste un fichier "Train" qui permet d'apprécier l'étiquetage réalisé par l'étiqueteur sous la supervision de Cafiero, Camps et Gabay.

2. Qualité de l'étiquetage Cafiero-Camps-Gabay

Le fichier "Train" comporte 82 063 lignes et 70 852 mots. Faute de disposer de l'index, on a procédé par sondages, ce qui a permis de détecter un nombre considérable d'erreurs d'étiquetage (quelques exemples significatifs dans le tableau ci-dessous)

Exemples d'erreurs relevées dans le fichier "Train" (les mots, entrées de dictionnaire et catégories grammaticales sont celles qui figurent dans le fichier) avec le nombre de fois qu'elles apparaissent).

Mots	Entrée diction.	Code grammatical	Effectifs	Observations
-elle	il	PROper	24	Le tiret aurait dû être enlevé
-elles	il	PROper	2	id + erreur sur le lemme (il au lieu de ils)
-il	il	PROper	106	id
-ils	il	PROper	9	id + erreur sur le lemme
-je	je	PROper	79	Id.
-le , les	il	PROper	10 + 5	id + erreurs sur le lemme ("il" au lieu de "le") et le code grammatical (pronom relatif et non pas personnel)
-lui	il	PROper	10	id
-même	même	ADJind	60	id
-mêmes	même	ADJind	2	id les deux fois "eux-mêmes" coupés en deux mots
-moi , -on , -nous , -vous, -toi , -tu	je , on, nous , vous , toi , tu	PROper	61 + 30 + 15 + 116 + 20 + 86 + 21 + 62	id

a-t	avoir	VERcjg	21	Erreur de segmentation (a t)
abord	abord	NOMcom	62	l'adverbe d'abord a été tronçonné en deux mots (de abord)
au, aux	à_le	PRE.DETdef	291 + 146	Ces deux entrées et le code n'existent pas en français
des	un	DETndf	179	Erreur (on aurait dû avoir « de_le », voir au dessous 'du')
du	de_le	PREDETdef	418	l'entrée et le code n'existent pas en français
(…)				
elles	il	PROper	12	Errreur sur le lemme (ils)
eux	il	PROper	425	Erreur sur le lemme (ils)
est-à-dire	c'est-à-dire	CONcoo	12	La forme a perdu son "c' " initial
(…)				
l', le, la, les	il	PROper	302 + 599 + 117 + 86	Le lemme est "le", pas "il". C'est un pronom relatif par personnel
(…)				
mâchoire	mâchoire	VERppe	2	nom et pas verbe au participe passé
parce	parce	ADVgen	22	au lieu de « parce que » conjonction

Classons les erreurs par catégories

I. Erreurs sur le découpage des mots

a. *Problèmes avec les formes contractes "au, aux, du, des"*

- selon Cafiero-Camps-Gabay, "au" et "aux" seraient des flexions de "à_le" ; "du" et "des" des flexions de "de_le". Si on admet cette étrangeté, il faudrait avoir les mots : "à_la" et "de_la" qu'on ne trouve pas.

Aucune nomenclature de dictionnaire ne comporte de pareilles entrées, aucune grammaire ne contient la partie du discours "préposition-déterminant-défini". C'est une solution illogique, ridicule, contraire à la convention généralement adoptée (décomposer en deux mots du et des). Est-ce une erreur de l'étiqueteur "oubliée" par les trois superviseurs ou une de leurs étranges inventions ?

- le lemme de "des" serait "un, déterminant" (au lieu de "de_le"). Du coup, il y a 179 "un" de trop ! Les superviseurs n'ont rien vu ?

b. Mots composés et tirets agglutinants

- De très nombreux "mots composés" sont tronçonnés à commencer par "moi-même", "toi-même", "par-là" et, systématiquement, les très nombreux noms composés du type "loup-garou", etc

- Des "mots" commençant par un tiret : -elle, -elles, -il, -ils, -je, -le , -les, -lui, -même, -mêmes, -moi, -on , -nous , -vous, -toi , -tu. Ces "mots" figurent aussi dans le texte sans tiret initial. Ce sont des mots différents ?

Dans le fichier "explorat-words" qui liste les mots utilisés dans les 85 pièces utilisées par Cafiero-Camps-Gabay, on trouve 25 mots commençant par un tiret initial (dont ceux cités ci-dessus) pour un total de 13 912 occurrences soit 1,4% de la surface du corpus.

- Des coupures arbitraires et des tirets oubliés

Par exemple : "a-t-il" est tronçonné 34 fois en :

a-t avoir verbe

-il il pronom
Soit deux erreurs à chaque fois. Les superviseurs ont laissé passer !
Autre mot inconnu provoqué par cette même défaillance :
est-à-dire c'est-à-dire CONcoo
le mot a perdu son c' initial alors que l'entrée de dictionnaire l'a conservée,
etc.

II. Erreurs sur les lemmes et codes grammaticaux

a. Erreurs sur les lemmes

Le pronom relatif "le" est codé pronom personnel "il" à la place de "le" pronom relatif. Ex "il les garde". Il y en a 1104 " = taux d'erreur 1,5%

- participes présents étranges. Par exemple : le mot "médisants" (ils sont médisants) se voit affecter le lemme "médire participe présent", De même pour "il est obligeant". Plusieurs centaines d'adjectifs sont ainsi codés verbes…

b. Erreurs sur les modes et les temps des verbes

- 136 verbes conjugués ont un temps inconnu. Par exemple dis et dit sont codés temps inconnu la moitié du temps alors qu'il s'agit clairement du présent dans plus de neuf cas sur dix.
- au moins 110 présents sont codés passés simples : "moi qui chéris", "je chéris sa présence", etc.
- des participes passés erronés. Confusions avec l'indicatif – "moi qui t'en promis" (promis est codé participe passé). Confusions avec l'adjectif. Par exemple, sont codés verbes au participe passé les adjectifs : "son frère ébloui", "un secret gardé", "son cœur tombé", "l'œil radouci", "la porte entrouverte", etc. L'étiqueteur code systématiquement ces adjectifs "participe passé" même quand il n'y aucun auxiliaire à l'horizon. Un sondage sur 2000 mots indique une proportion de 43% d'erreurs sur les participes passés, soit sur l'ensemble de « train » environ 660 erreurs.

c. Les substantifs sans genre.

Dans la nomenclature du français, tout substantif est doté d'un genre (masculin ou féminin).

Cafiero-Camps-Gabay ne donnent pas de genre à 421 substantifs comme "garde" (20 fois), "mode" (11), "poste" (5), "tour" (25)… C'est d'autant plus curieux que, très souvent, le déterminant devant ou un adjectif associé permettent facilement la reconnaissance de ce genre (ex. "la mode", "la poste", "ce tour"), etc.

Beaucoup d'autres choses étonnent sans pouvoir être décomptées aisément (comme "mâchoire" codé verbe avec comme entrée de dictionnaire "mâchoire"). Autre exemple, les majuscules initiales au début de vers. La codification de certains noms communs en majuscule semble univoque comme si l'étiqueteur ne reconnaissait plus les homographies. Par exemple, "Soit" est systématiquement codé conjonction même quand c'est le verbe être : "l'outrage… Soit vu par les autres", "Soit dit sans vous déplaire", etc. De même, certains adjectifs sont systématiquement codés verbes quand ils ont une majuscule initiale. Ex : "le considérer, Interdit et confus" (interdit est codé verbe au participe passé).

Récapitulons. Un simple sondage non-exhaustif sur ces 70 852 mots étiquetés donne :

- 1 692 découpages erronés,
- 1 720 erreurs sur les mots vedettes (des, il, les…),
- 906 erreurs sur la codification des verbes,
- 421 substantifs sans genre.

Total : 4 739 erreurs soit 7%.

C'est donc avec beaucoup d'étonnement que l'on voit Cafiero et Camps :

- annoncer des taux de réussite de 97% et 98% pour les deux étiqueteurs utilisés (article, p. 7). Soit les auteurs ne connaissent pas les règles de base de l'étiquetage du français, soit ils mentent.

-déclarer que : "POS 3-grams (i.e., sequences of tags, such as "NOUN ADJECTIVE VERB"), proved to be effective criteria to discover the author of a text" (article, p. 2 2e colonne). Avec une pareille quantité d'erreurs sur les étiquettes, cette prétention est évidemment dérisoire.

Conclusion : l'étiquetage est absolument pas fiable ; aucune étude sérieuse ne peut être réalisée sur de telles bases.

Pourtant, Cafiero et Camps ont poursuivi leur chemin… vers une catastrophe de plus grande ampleur encore.

3. Catastrophe finale.

Après avoir étiqueté les textes, on établit leurs index sur lesquels sont calculées les distances qui servent à tracer les graphes et à réaliser l'attribution d'auteur. Chacune de ces étapes doit être réalisée avec soin et en se gardant des erreurs.

Dans Data_3, huit fichiers permettent de juger de la qualité de cette indexation (de control-lemma.csv à explorat-words.csv).

- Dans les deux fichiers "-word", on retrouve toutes les erreurs de segmentation qui conduisent à un taux d'erreur certainement supérieur à 7%. Ces données étaient donc inutilisables. Les classifications opérées par Cafiero et Camps et leur attribution d'auteur, sur les "mots", reposent sur du sable.

- La surprise principale provient des "lemmes" et des "rimes" (qui sont étudiées de la même manière). Rappelons que les lemmes (word types) sont formés par l'association d'une entrée de dictionnaire et d'un code grammatical. Or dans les 4 fichiers "-lemmas" et "-rhymes", ne figurent que les entrées de dictionnaire sans les catégories grammaticales associées. Du coup, dans l'index, il n'y a plus qu'une seule ligne pour "le" mélangeant les pronoms et les articles, un seul "être" amalgamant le substantif ("l'être humain") dans l'océan de toutes les conjugaisons de "être" ; de même pour "avoir", "pouvoir", "devoir", "savoir"… ; une seule ligne pour "tout" alors qu'il en faudrait quatre, etc.

Dans tout texte en français ces homographies – disparues dans les index de Cafiero-Camps-Gabay - concernent au minimum un mot sur trois. C'est donc également l'incertitude qui a pesé sur les prétendus calculs effectués sur les lemmes et sur les rimes. Autrement dit, l'attribution d'auteur s'est réalisée dans un brouillard total.

Annexe 6
Première réponse à MM. Cafiero et Camps.

Le 26 novembre 2019 un journaliste nous a communiqué l'article de Cafiero et Camps. Après une étude rapide, nous avons mis en ligne, sur researchgate, la réponse ci-dessous le 27 novembre à 20h00 (à la sortie de l'article de *Science Advances*).

Nous avons averti de notre réponse : le CNRS, l'AFP, l'Ecole des Chartes et les principaux médias. Tous ont fait une large place aux affirmations de Cafiero et Camps. A part deux journaux étrangers, personne n'a fait état de ma réponse.

HAL (les archives ouvertes du CNRS) a bloqué ce texte pendant six jours avant que nos protestations auprès de la direction générale finissent par le faire mettre en ligne.

Nous le reproduisons ci-dessous afin de fournir un dossier complet au lecteur.

A ce jour (18 décembre) aucune des graves accusations contenues dans ce texte n'a reçu le moindre commencement d'une réponse.

Dominique Labbé

Dominique.labbe@umrpacte.fr

https://www.pacte-grenoble.fr/membres/dominique-labbe

Réponse à

Florian Cafiero et Jean-Baptiste Camps. Why Molière most likely did write his plays. *Science Advances*. 5: eaax5489. 27 November 2019.

Grenoble le 27 novembre 2019

Résumé

Dans cet article, MM. Cafiero et Camps prétendent apporter la preuve que P. Corneille n'a écrit aucune des pièces présentées par Molière. Ils utilisent pour cela 6 "caractérisques" (lemmes, formes, mots outils, rimes, affixes, n-grams) couplées avec des classifications automatiques. En fait, les auteurs fournissent peu d'informations précises sur ces méthodes et aucune donnée chiffrée. Les quelques informations, notamment dans les annexes en ligne, suffisent pour soulever beaucoup de doutes. Par exemple, la liste des "mots outils" comporte de nombreuses étrangetés qui ne peuvent s'expliquer simplement par des maladresses. De même, ils ont opéré un tri dans les pièces de Molière, retirant de l'expérience 24 des 33 pièces. Parmi ces pièces écartées : *Psyché* qu'il ne fallait surtout pas retirer ! Enfin, le détail des classifications (publié dans une annexe en ligne séparée de l'article) montre un échec total. Leurs méthodes se révèlent incapables de reconnaître : Boursault, Chevalier, Dancourt, Donneau de Visé, Gillet de la Tessonnerie, Pierre Corneille, Thomas Corneille, La Fontaine, Ouville, Quinault, Régnard, Rotrou… et Molière.

Deux remarques préalables.

Premièrement, une série de données est disponible sur le site :

"https://advances.sciencemag.org/content/suppl/2019/11/21/5.11.eaax5489.DC1

Ces données, intéressantes pour les chercheurs, prétendent répondre à l'exigence de transparence indispensable étant donnée la portée du débat. Nous conseillons au lecteur d'y jeter un coup d'œil et de se faire sa propre opinion. En particulier : trouvera-t-il les données chiffrées qui ont servi à tracer les graphiques et qui conduisent aux conclusions si fermes des auteurs ?

Deuxièmement, dès le premier paragraphe de cet article, il est clair que nos travaux sont la principale cible. Alors pourquoi n'avoir pas pris contact avec nous ? Pourquoi ne pas nous avoir donné la possibilité de travailler ensemble et d'arriver à une position commune ou à une liste raisonnée de divergences ? C'est ainsi que la science progresse. L'échange des données et la possibilité de répondre pour les chercheurs mis en cause sont des règles basiques dans la communauté scientifique. MM. Cafiero et Camps ignorent ces règles ?

Voici quelques observations que MM. Cafiero et Camps nous contraignent à rendre publiques, alors que nous les leur aurions volontiers communiquées de manière privée s'ils avaient respecté les usages de la communauté scientifique.

Dès les premières lignes, une certitude : ils ne nous ont pas lus. Par exemple,

- dans la première colonne de la page 1, ils affirment que notre méthode donne plus de poids aux mots de fortes fréquences. C'est faux ! S'ils s'étaient donnés la peine de lire nos articles, ils auraient vu que la distance intertextuelle donne à chaque mot exactement son poids dans les textes sans aucune déformation ;

- dans la première colonne de la page 2, ils nous font dire que les *Précieuses ridicules* seraient de P. Corneille (c'est la seule pièce qu'ils citent à notre propos). Pas de chance ! Dans notre article de 2001, cette pièce n'est pas attribuée à P. Corneille et nous n'avons pas varié depuis.

A vouloir trop "prouver"…

Sur le fond, le 27 novembre au soir, à sa parution, l'article et les annexes en ligne ne comportent aucun tableau de données, seulement des graphiques difficiles à comprendre. Dans l'état actuel de ce qui a été mis en ligne, il est impossible de contrôler sans refaire les expériences, ce qui demandera de longues semaines.

En attendant, il faut donc faire confiance ?

Malheureusement dans ce travail, beaucoup de choses soulèvent la défiance.

1. Qualité des données et des analyses.

Nous nous appuyons sur la table 1 (page 10) et la table S4, dans l'annexe mise en ligne avec l'article à l'adresse mentionnée ci-dessus (*Supplementary Materials for Why Molière most likely did write his plays*). Ces deux tables montrent que :

- MM. Cafiero et Camps ne travaillent pas sur les lemmes – contrairement à ce qu'ils affirment. Par exemple, la table 10 contient "ta", "ton", "tes" qui sont des flexions ("word forms") d'un même article possessif. Ou encore, la table des "mots fonctionnels" (S4) comporte : c', jusqu', l', etc. qui sont des flexions des lemmes : 'ce', 'jusque', 'le', etc.

- dans cette liste des mots "fonctionnels", on trouve les conjonctions de coordinations sauf 'or' ; on trouve 'après' mais pas 'avant' ; 'jamais' mais pas 'toujours' ;'ici' et 'là' mais pas 'ailleurs' ; 'voilà' mais pas 'voici' ; 'se' mais pas 'il', 'elle' ou 'me', 'te' ; on trouve 'tous' et 'toute' mais pas 'toutes', etc. Toutes les prépositions, conjonctions, articles, pronoms, adverbes qui sont absents de cette liste ne seraient donc pas des "outils" ? Certaines flexions d'un même mot le seraient et d'autres pas ?

Autre surprise : figurent dans cette liste des outils : les verbes "être" et "avoir". Ce seraient donc des "outils" et pas des "mots lexicaux" ? Il faudrait alors trouver dans la table S4 toutes les flexions de ces deux verbes, ce qui est très loin d'être le cas. Dès lors, certaines de ces flexions seraient des mots "lexicaux" – comme "suis" ou "sommes" qui ne sont pas dans la liste alors qu'ils sont très utilisés dans les comédies - et d'autres des "outils", comme "est" ou "êtes", également très utilisés, qui sont, eux, dans la liste de la table S4 ?

Les auteurs ne connaissent pas le français ?

Ces curiosités sont nécessaires pour amener les pièces de Molière à la "bonne place". Effectivement, sans surprise, on lit p. 3 : "The highest agglomerative coefficient is obtained for the analysis of function words. In this analysis, all plays signed by Molière are clustered together" (p. 3).

Nous verrons aussi que "all" n'est pas exact puisque cette affirmation ne concerne que 9 pièces attribuées à Molière sur les 33 qu'il a présentées.

La même question peut être posée à propos de toutes les "features" (lemmes, formes, rimes, affixes, n-grams, mots outils) sélectionnées ou enlevées – selon qu'elles réussissent ou non à isoler certaines pièces attribuées à Molière - sans qu'aucune explication claire, ni chiffres précis soient donnés au lecteur.

- A la fin de cette même annexe, on trouve les "mots outils" suivants : qu / c / n / l / d / s / jusqu. Naturellement, ces mots auraient dû avoir une apostrophe terminale. Autrement dit, les textes utilisés par Caferio et Camps comportent de nombreuses fautes de frappe qui n'ont pas été corrigées.

Rappelons la règle de base en statistique : la qualité des conclusions dépend de celle des observations du phénomène. Ici, le peu d'informations disponibles suffit largement pour démontrer que la qualité n'est pas au rendez-vous.

A cela s'ajoute une grande désinvolture dans le choix des pièces.

2. Une curieuse sélection

Avant d'être appliquée aux cas douteux, une méthode d'attribution d'auteur doit être mise à l'épreuve sur des cas qui ne le sont pas. Dans l'article de Cafiero et Camps, c'est le rôle des corpus "exploratoire" et de "contrôle" qu'on peut découvrir en détail dans les tableaux S2 et S3 (du document annexe). En gros, le corpus exploratoire contient, outre quelques pièces de Molière, celles d'auteurs contemporains ou antérieurs. Le corpus de "contrôle" est composé d'une trentaine de pièces postérieures.

Ces listes réservent de nombreuses surprises. Par exemple :

- La Fontaine *Ragotin*. Les auteurs ignorent que cette pièce a été jouée et publiée sous le nom de... Champmeslé. Cet acteur de la troupe de l'Hôtel de Bourgogne était comédien poète exactement comme Molière et son histoire est très semblable à celle de Molière. Après la mort de Champmeslé (et de La Fontaine), un libraire d'Amsterdam a republié cette pièce sous le nom de La Fontaine sans expliquer les raisons de cette attribution posthume. Cette pièce n'avait donc rien à faire dans un corpus de textes de paternité indiscutable. Ou alors, puisque, au début de leur papier, MM. Cafiero et Camps mettent en doute l'existence même du système du comédien poète, il fallait l'enregistrer sous le nom de Champmeslé…

- Molière : *Dom Garcie*, *Mélicerte*, *Sganarelle* figurent dans le corpus des pièces utilisées pour étalonner la méthode de manière "exploratoire". Ces pièces seraient donc "sûres" ? L'article n'apporte aucune explication concernant cette étrange décision. On ne comprend pas (ou plutôt on comprend trop bien) pourquoi ces pièces ne figurent pas dans l'expérience "finale" censée démonter l'existence de Molière "grand auteur".

3. Des "oublis" fâcheux

La table S1 de l'annexe liste les 30 pièces sur lesquelles a porté l' "expérience" finale d'attribution d'auteur :

- 9 de Molière alors que celui-ci en a présenté 33 : la paternité des autres est donc certaine ?

- 8 de Pierre Corneille alors qu'il en a présenté 33 (ou 34 avec *Psyché*). Les autres ne seraient pas de lui ?

- 10 de Thomas Corneille alors que 37 pièces de lui sont disponibles.

etc., etc.

Pour Molière, Cafiero et Camps ont écarté toutes les pièces en prose (*Les précieuses ridicules*, *Dom Juan*, *l'Avare, le Bourgeois gentilhomme*, *le Malade imaginaire*, etc.). Elles ne posent donc aucun problème d'attribution ? En effet, leur conclusion affirme bien imprudemment que Molière est l'auteur de *toutes* ses pièces !

Dans les pièces en vers, ils ont "oublié" :

Sganarelle (1660)

Dom Garcie (1661)

Mélicerte (1666)

Les amants magnifiques (1670)

Psyché (1671)

Cette dernière pièce est le plus grand succès de Molière (plus de 70 représentations d'affilée, sous le seul nom de Molière, et une recette supérieure au chiffre d'affaire annuel de la troupe). Six mois après ce triomphe, la pièce est publiée – avec en couverture et en page de garde, le seul nom de Molière – mais avec un avertissement du "libraire" (l'éditeur) :

"Le libraire au lecteur

Cet ouvrage n'est pas tout d'une main. M. Quinault a fait les paroles qui s'y chantent en musique, à la réserve de la plainte italienne. M.de Molière a dressé le plan de la pièce, et réglé la disposition, où il s'est plus attaché aux beautés et à la pompe du spectacle qu'à l'exacte régularité. Quant à la versification, il n'a pas eu le loisir de la faire entière. Le carnaval approchait, et les ordres pressants du Roi, qui se voulait donner ce magnifique divertissement plusieurs fois avant le carême, l'ont mis dans la nécessité de souffrir un peu de secours. Ainsi, il n'y a que le prologue, le premier acte, la première scène du second et la première du troisième dont les vers soient de lui. M. Corneille a employé une quinzaine au reste ; et, par ce moyen, Sa Majesté s'est trouvée servie dans le temps qu'elle avait ordonné."

Psyché a donc été présentée au public sous le seul nom de Molière (comme toutes les autres). Cependant, dans ce cas, grâce à une indiscrétion (il y en a eu d'autres du vivant de Molière), la collaboration entre Corneille et Molière n'est pas discutable. Les passages, que chacun d'eux est censé avoir écrits, sont clairement identifiés. Dès lors, ces passages fournissaient "l'épreuve décisive" qui aurait permis de juger de l'efficacité de la méthode de MM. Cafiero et Camps : est-elle capable de distinguer ces passages et de les attribuer correctement ?

Dans leur article, page 6, au bas de la deuxième colonne, les auteurs admettent implicitement que la réponse est négative et que, pour cette raison, ils ont retiré *Psyché* de toutes leurs expériences.

Faute d'avoir réalisé cette expérience décisive sur *Psyché*, leurs prétentions s'écroulent.

4. La dernière table (S5) révèle un échec complet

Cette table donne – pour trois seulement des 6 séries de graphiques de l'article (graphiques difficilement compréhensibles et incontrôlables en l'absence des données chiffrées) - la composition des "clusters" découpés dans les corpus (mais sans les chiffres). On s'attend à ce que, dans l'expérience préliminaire et cruciale (décrite par les six diagrammes de la figure 1), chaque auteur "sûr" trouve sa place dans l'un des groupes. Il faut aussi que ces groupes soient homogènes et que le même classement se retrouve quelle que soit la "feature" utilisée. Sinon, l'expérience doit s'arrêter puisque le contrôle préalable a échoué.

Le ratage commence avec "Molière" dont *Amphitryon*, *Dom Garcie de Navarre*, *le Dépit amoureux*, les *Fâcheux*, *Mélicerte*, *Sganarelle* ne sont jamais groupées de manière stable et complète… Le titre de l'article de Cafiero et Camps est donc trompeur sinon mensonger.

D'autres pièces d'un même auteur, comme celles de La Fontaine, Donneau de Visé ou Boursault, se trouvent dans beaucoup de groupes différents et jamais toutes ensemble ! Les pièces de Thomas Corneille ou de Quinault se dispersent presque partout, groupées avec pratiquement tous les autres auteurs mais pas toujours les mêmes selon les "features", etc., etc.

En effet, non seulement les classements sont aberrants mais ils varient en fonction des "features" utilisées. C'est le cas notamment pour La Fontaine dont les pièces se trouvent un peu partout, rarement à la même place et jamais ensemble !

Cette table prouve que **la méthode de MM. Cafiero et Camps est incapable de reconnaître : Boursault, Chevalier, Donneau de Visé, Gillet de la Tessonnerie, Pierre Corneille, Thomas Corneille, La Fontaine, Ouville, Quinault, Rotrou… et Molière !** Pas un n'échappe au naufrage.

Et que font ces curieux expérimentateurs ?

Ils retirent toutes les pièces qui posent problème ! A la trappe : Boursault, Chevalier, Donneau de Visé, Gillet, La Fontaine, Ouville, Quinault. Ce ne sont donc pas de vrais auteurs ? Ceux qui restent, au tour final, se trouvent amputés de la plupart de leurs œuvres. Molière en garde 9 (moins d'une sur trois).

Et… miracle ! Cela marche. Ou, du moins, on arrive enfin là où on voulait en venir depuis le début : 9 des 33 pièces assignées à Molière seraient isolées (mais comment contrôler puisque nous n'avons aucun chiffre). Et peu importe que les autres auteurs continuent à se mélanger.

On lit dans l'article que le "corpus de contrôle" ne poserait pas de problème de ce genre… sauf Dancourt et Régnard. Dancourt (Florent Carton, comédien poète lui aussi) a présenté plus de 40 pièces et Régnard plus d'une dizaine. Ils ont dominé la comédie à la fin du XVIIe et au début du XVIIIe. Et la méthode échoue sur ce cas emblématique ! Cela ne pose aucun problème à MM. Cafiero et Camps. Ils ont toujours une hypothèse "ad hoc" sous la main pour expliquer l'inexplicable : ici, la femme de l'un était actrice, devait certainement connaître l'autre et influencer les deux.

Les explications embrouillées de la page 3 et de la deuxième colonne de la page 6 sont de la même eau.

Les méthodes de MM. Cafiero et Camps sont incapables de reconnaître les auteurs des pièces du XVIIe, y compris celles attribuées à Molière qui aurait mérité de meilleurs avocats !

Nos travaux sont en ligne (Archives ouvertes du CNRS et ResearchGate).

Une présentation de la méthode d'attribution d'auteur :

https://images.math.cnrs.fr/La-classification-des-textes.html

Le système du comédien poète :

https://www.researchgate.net/publication/293334890

https://www.researchgate.net/publication/333236909

Nos corpus sont en ligne sur le site du Centre de Linguistique de Corpus (Université de Neuchâtel : http://www.unine.ch/clc/home/Corpus.html