Qui a écrit Aétius, Juba et Tachmas ?

Dominique Labbé

Laboratoire PACTE
(CNRS - Université de Grenoble-Alpes)

Dominique Labbé

dominique.labbe@umrpacte.fr
http://www.pacte-grenoble.fr/blog/membres/labbe-dominique/

Qui a écrit *Aétius*, *Juba* et *Tachmas* ?

Une attribution d'auteur par ordinateur.

Rapport technique

Décembre 2014

Résumé

Qui a composé *Aétius*, *Juba* et *Tachmas*, pièces inédites du fonds Maniban-Campistron aux Archives départementales de la Haute-Garonne à Toulouse ? La réponse est apportée grâce à une procédure d'attribution d'auteur par ordinateur. Un calcul de distance mesure précisément la plus ou moins grande ressemblance de ces textes par rapport à un vaste corpus de référence comportant 235 pièces contemporaines. Les textes séparés par les distances les plus faibles sont écrits par un même auteur. Des procédures de classification repèrent les groupements optimaux au sein de cette population et mesurent le degré d'appartenance de chaque texte à un groupe donné. Ces procédures permettent d'identifier l'auteur de ces trois manuscrits qui est aussi celui de toutes les tragédies représentées sous le nom de Jean-Galbert Campistron et de Jean de La Chapelle.

De nombreux indices lexicaux et stylistiques viennent confirmer cette attribution, notamment les longueurs, structures et fonctions des phrases qui caractérisent le style de cet auteur par rapport à ses contemporains. Enfin, une mesure de l'influence du temps permet d'offrir une datation de ces œuvres.

La lecture de ce texte ne demande aucune connaissance en mathématique et statistique.

Mots clefs : lexicometrie ; attribution d'auteur ; théâtre français ; XVIIe siècle ; Corneille ; Racine ; Campistron ; La Chapelle

Abstract

Who wrote *Aetius*, *Juba* and *Tachmas*, three manuscript tragedies stored in the Maniban-Campistron fund in the Departmental Archives of the Haute-Garonne (Toulouse)? The answer is provided by a computer-assisted authorship attribution. A distance calculation accurately measures the similarities between these texts and a vast reference corpus of 235 contemporary plays. The texts separated by the smallest distances are written by the same author. Classification procedures spot the best groupings within this population and measure the degree of membership of each text in a given group. These procedures leads to the identification of the author of these three manuscripts as well as all the tragedies represented under the name of Jean-Galbert Campistron and Jean de La Chapelle.

A large number of lexical and stylistic evidences confirm these attributions, including the lengths of sentences, their structures and functions that characterize the style of one author compared to his contemporaries. Finally, a measure of the influence of the chronology makes it possible to date approximatively the composition of these texts.

This text requires no prior knowledge in mathematics and statistics.

Key words : lexicometry ; authorship attribution ; French theatre ; seventeenth century ; Corneille ; Racine ; Campistron ; La Chapelle

Remerciements

Jean-Charles Basson a assuré les négociations avec les Archives départementales de Haute-Garonne à Toulouse. Il a cliché les manuscrits présents dans le fonds Maniban-Campistron et en a assuré la transcription en français contemporain en collaboration avec Dominique Labbé.
Il a participé à toutes les étapes de la recherche qui a permis l'attribution présentée dans ce rapport.

La méthode d'attribution d'auteur a été mise au point avec Cyril Labbé (Laboratoire d'Informatique de Grenoble – Université Joseph Fourrier). Il a collaboré à l'écriture des programmes informatiques et à la mise au point des procédures statistiques utilisées dans cette recherche.

Nos travaux d'attribution d'auteur ont bénéficié de l'aide d'un grand nombre de chercheurs, notamment : Dominique Andolfatto, Edward Arnold, Jean-Guy Bergeron, Mathieu Brugidou, Pierre Hubert, Nelly & Jean Leselbaum, Xuan Luong, Thomas Merriam, Denis Monière, Mathieu Ruhlman, Jacques Savoy.

SOMMAIRE

INTRODUCTION 9

PREMIERE PARTIE
MESURE DES PROXIMITES ENTRE TEXTES 13

CHAPITRE I.
LA DISTANCE INTERTEXTUELLE 15
I. PREPARATION ET TRAITEMENT DES TEXTES 15
Transcription des textes en français contemporain 15
Balisage et standardisation graphique 17
Etiquetage 18
II. CALCUL DE LA DISTANCE 20
Principes du calcul 20
Calcul sur des textes de longueurs différentes 22
Limites du calcul 25
II. QUELS SONT LES FACTEURS QUI DETERMINENT LA DISTANCE ? 26

CHAPITRE II.
UN EXEMPLE : LES FRERES CORNEILLE ET J. RACINE 27
I. DISTANCES INTRA-CORPUS 28
Pierre Corneille 28
Thomas Corneille 32
Jean Racine 33
II. DISTANCES INTER-CORPUS 37
Jean Racine et Pierre Corneille 38
Jean Racine et Thomas Corneille 40
Conclusions du chapitre 42

CHAPITRE III.
IDENTIFICATION DE L'ECRIVAIN 43
I. TESTS STATISTIQUES STANDARDS 43
Comparaison d'une œuvre à une autre 43
Un cas limite 46
II. UN MÊME MOULE ? 48
Une population unique : la tragédie classique ? 48
Trois écrivains et trois œuvres distinctes 50
III. LE POIDS DE L'ECRIVAIN 51
Calcul du poids de l'écrivain 51
Examen direct des intervalles 52
Examen graphique 54
Conclusions du chapitre 56

CHAPITRE IV.
J. RACINE – J.-G. CAMPISTRON ET J. DE LA CHAPELLE 59
I.DISTANCES INTRA 59
J. de La Chapelle 59
J.-G. Campistron 60
II. DISTANCES INTER 62
J. de La Chapelle et les frères Corneille 62
J.-G. Campistron et les frères Corneille 65
J.-G. Campistron et J. Racine 68
Tachmas 71
J. de la Chapelle, J. Racine et J.-G. Campistron 72

Conclusions de la première partie 76

DEUXIEME PARTIE
CLASSIFICATIONS 79

CHAPITRE V.
CLASSIFICATIONS HIERARCHIQUES 81
I. UN EXEMPLE DE CLASSIFICATION : P. CORNEILLE 82
La classification 82
Le dendrogramme 83
II. J. RACINE, LES FRERES CORNEILLE… 86
Classification des œuvres présentées par J. Racine 86
Vérification sur les trois écrivains 87
III. … J. DE LA CHAPELLE ET J.-G. CAMPISTRON 88
Les résultats attendus… 88
… loin de la réalité
Conclusions du chapitre 91

CHAPITRE VI.
CLASSIFICATIONS NON-HIERARCHIQUES 93
I. METHODES 94
Voisinage et indice d'appartenance 94
Limites 94
II. J. RACINE ET LES FRERES CORNEILLE… 96
Constitution de groupes homogènes 96
Les textes les plus lointains 100
II. …LA CHAPELLE ET J.-G. CAMPISTRON 102
Trois auteurs 102
J. de La Chapelle à la charnière 104
Conclusions de la deuxième partie 106

TROISIEME PARTIE
COMPLEMENTS POUR UNE ATTRIBUTION D'AUTEUR 107

CHAPITRE VII.
LE TEMPS 109
I. L'EFFET DU TEMPS DANS LES ŒUVRES D'UN ECRIVAIN 110
Chronologie et attribution d'auteur 110
Le temps dans l'œuvre présentée par J. Racine 111
Le temps dans l'œuvre des frères Corneille 115
II. L'EFFET DU TEMPS DANS LA COMPARAISON ENTRE PLUSIEURS ECRIVAINS
Le temps dans la comparaison entre les pièces des frères Corneille et de J. Racine 116
Le temps dans la comparaison entre J. Racine, J. de la Chapelle et J.-G. Campistron 118

CHAPITRE VIII.
PHRASE ET CHOIX STYLISTIQUES 121
I. PHRASE ET STYLISTIQUE 121
Délimitation de la phrase et décompte 121
La phrase dans les pièces présentées par J. Racine 123
II. COMPARAISON DES PHRASES 125
Tests statistiques 125
Plusieurs types de phrases 129
III. STYLES DANS LES PIECES PUBLIEES PAR J. RACINE, J. DE LA CHAPELLE ET J.-G. CAMPISTRON 130
Tests statistiques 130
Contrôles graphiques 132
IV. QUATRE TYPES DE PHRASES 134
Conclusions du chapitre 136

CHAPITRE IX.
UNE ECHELLE DE LA DISTANCE 137
I. L'ECHELLE 137
II. CONDITIONS DE VALIDITE ET PRECAUTION D'UTILISATION 138
La qualité de la mesure 138
La précision de la mesure 139
III. ATTRIBUTION D'AUTEUR 139
Zones d'acceptation et de rejet 139
Influence, collaboration ou plume de l'ombre ? 140

Conclusion générale 143
Paternité des œuvres parues sous les noms de J. Racine, J. de La Chapelle et J.-G. Campistron 143
Sur l'attribution d'auteur 144
Lexicométrie et sciences humaines 145

Annexes
- 1. L'attribution d'auteur assistée par ordinateur 147
Références 149
- 2. Les corpus 152
- 3. Les comédiens poètes du XVIIe et leurs "œuvres" 154
- 4. Corpus électronique des pièces de théâtre du XVIIe siècle 158

Le sujet de cette tragédie est assez du goût de l'ancienne Athènes ; il est grand, simple, propre à inspirer la terreur et la compassion, et par dessus cela, il est tout neuf.

L'incident qui fait le fonds de cette pièce, est peut-être la plus grande époque de l'histoire profane : c'est la ruine entière de la république romaine, et l'établissement de la monarchie universelle de Jules César.

Tout le monde sait qu'après la défaite et la mort de Pompée, Juba, roi de Mauritanie, le plus fidèle et le plus courageux de ses amis, recueillit dans l'Afrique les débris du parti de ce grand homme, et qu'ayant joint ses forces à celles de Scipion, de Varus et de Caton, il renouvella une guerre où César faillit succomber. Mais il fallut céder enfin, et la bataille de Thapse, qui fut pour le moins aussi sanglante que celle de Pharsale, ayant donné le dernier coup à la liberté de Rome, Juba, Scipion et Caton qui se virent sans ressource, finirent leur vie par leurs propres mains.

Cette aventure tragique a paru si propre pour le théâtre à l'illustre Monsieur Racine, qu'un de ses amis [mention manuscrite en marge : M. Campistron], qui depuis plusieurs années remplit si dignement sa place, lui a ouï dire qu'il était résolu de la traiter avant que de renoncer à la tragédie.

On ne trouvera pas mauvais, j'en suis sûr, que nous prenions les devants, et que nous commencions par mettre les ombres du tableau, en attendant que ce grand maître veuille y mettre les couleurs.

(Dominique de Colonia. Préface de *Juba*. Lyon : Guerrier, 1695)

Introduction

Selon le Père Colonia, en 1695, J. Racine (1639-1699) n'aurait pas abandonné le théâtre[1] et écrirait une tragédie sur Juba. Le Père Colonia affirme tenir cette information de l'un des amis de J. Racine - Jean-Galbert Campistron (1656-1723) - qui tient dignement sa place depuis plusieurs années. Et Colonia se vante d'avoir devancé J. Racine.

Dans les années 1950, les descendants de J.-G. Campistron remettent aux Archives départementales de la Haute-Garonne à Toulouse les papiers de la famille contenant plusieurs manuscrits appartenant à leur ancêtre, parmi lesquels une tragédie en alexandrins et en cinq actes intitulée… *Juba*[2]*,* mais aussi les quatre premiers actes d'une autre tragédie (*Aétius)* - représentée à la comédie française en 1693, sous le nom de J.-G. Campistron, et restée inédite jusqu'à ce jour - ainsi que des fragments du début d'une troisième à l'état d'ébauche : *Tachmas*.

Deux questions se posent.

- Le manuscrit de *Juba*, présent dans les archives de J.-G. Campistron, est-il la pièce à laquelle J. Racine travaillait en 1695 ?

- Que signifie la formule du Père Colonia selon laquelle, en 1695, J.-G. Campistron remplit dignement la place de J. Racine depuis plusieurs années ? J.-G. Campistron était-il le prête-nom de J. Racine ? Ou simplement son digne successeur ?

La statistique appliquée au langage[3] apporte des réponses à ces deux questions. Ce rapport les présente de manière détaillée. Il a été rédigé à l'intention des non-statisticiens et spécialement des spécialistes du XVIIe siècle. Il est mis en ligne lors de la parution de la transcription des trois manuscrits accompagnés d'un dossier historique[4].

---

[1] Les biographies de J. Racine indiquent qu'il abandonne le théâtre en 1677, lorsqu'il devient historiographe du roi.

[2] Sur l'histoire de ce dépôt et une première analyse de *Juba* : Gérard Pierre. Une tragédie inédite de Campistron. Juba, roy de Mauritanie. *Actes du 11e congrès d'études de la fédération des sociétés académiques et savantes Languedoc-Pyrénées – Gascogne* (Albi 11-13 juin 1955). Albi : Imprimerie des orphelins apprentis, p. 36-42.

[3] Voir la note en annexe 1 et la bibliographie placée à la fin de cette annexe. Les travaux publiés par notre réseau de recherche sont consultables en ligne (notamment sur le site hal.archives-ouvertes.fr. Une liste peut être consultée sur la page personnelle de D. Labbé.

[4] Racine Jean. *Aétius, Juba, Tachmas. Tragédies inédites transcrites et présentées par Jean-Charles Basson & Dominique Labbé*. Montréal : Monière-Wollank Editeurs, 2015.

Au sein de notre bibliothèque électronique de plus de 35 millions de mots, figurent 235 pièces de théâtre du XVIIe présentées par 32 "auteurs" différents[1]. Cette section de la bibliothèque est présentée dans la dernière annexe placée à la fin du rapport. *Juba*, *Aétius* et *Tachmas* ont été comparées à ces pièces afin d'identifier la plume qui les a composées.

Au-delà de l'énigme posée par le Père Colonia, une remarque préalable fera comprendre l'enjeu de cette discussion. Il y a lieu de distinguer l'écrivain, le copiste (scripteur) et l'"auteur" pour la galerie.

L'**écrivain** a écrit, c'est-à-dire imaginé le sujet, créé et composé la pièce. C'est lui que la statistique permet d'identifier.

Le **copiste** recopie le manuscrit. Par exemple, les restes de la correspondance de J.-G. Campistron indiquent qu'il a recopié des textes pour les envoyer à son père à Toulouse. Cela n'en fait pas l'écrivain. Au XVIIe, il n'existe pas de machines à écrire ni de photocopieuses. Quand une pièce est considérée comme achevée, les copistes en produisent autant de copies que nécessaires pour la troupe de théâtre et l'imprimeur. Ces copies sont détruites quand le livre est imprimé. C'est pourquoi, il n'existe aucun manuscrit autographe des pièces de J. Racine ou de P. Corneille. De plus, la plume de l'ombre peut souhaiter ne pas voir ses manuscrits circuler. Ainsi, R. Gary faisait recopier ses manuscrits "Ajar" par son neveu prête-nom[2].

L'**intermédiaire** présente la pièce aux comédiens, puis, en cas de succès, il la vend à un "libraire" (les éditeurs du temps). Parfois, il est aussi **producteur** en finançant la création[3]. Il passe pour l' "auteur" vis-à-vis du public, des gazettes, des autorités… et de la postérité parce que son nom figure sur la couverture du livre, sur les affiches et dans les gazettes alors que cela indique simplement qu'il a **présenté** la pièce.

Il peut arriver qu'une même personne joue les trois rôles mais, au XVIIe siècle c'est l'exception. La majorité des pièces de théâtre ne sont pas présentées sous le nom des écrivains qui

---

[1] Boisrobert, Boursault, Boyer, Campistron, Champmeslé, Corneille P., Corneille T., Dancourt, Desfontaines, Desmarets, Donneau de Visé, Dufresny, de l'Estoile, Genest, Hauteroche, La Calprenède, La Chapelle, La Fontaine, La Fosse, Mairet, Molière, Montleury (père et fils), Poisson, Pradon, Quinault, Racine, Regnard, Rotrou, Ruyer, Scarron, Scudery, Villiers. En fait, le nombre des écrivains est nettement plus réduit ! Pour le détail des pièces, voir annexe 3.

[2] Bellos David. *Romain Gary : A Tall Story*. London : Harvil Secker, 2010.

[3] C'était le cas de Molière (Labbé Dominique. *Si deux et deux sont quatre*. Paris : Max Milo, 2009).

les ont composées mais sous celui des intermédiaires qui les achètent à ces écrivains et les négocient avec les troupes. Beaucoup de comédiens du temps jouent ce rôle d'intermédiaires. Molière est le plus connu mais ils sont très nombreux (annexe 4).

Dès lors la question devient : J.-G. Campistron est-il l'un de ces intermédiaires – comme le suggère le Père Colonia – ou bien a-t-il composé les pièces parues sous son nom ainsi que *Juba* et *Tachmas* ?

La recherche de l'auteur d'un texte est un cas particulier d'une question plus générale : comment trouver le meilleur classement possible au sein d'une vaste collection de textes écrits dans une même langue ? Pour répondre à cette question, deux outils sont nécessaires. D'une part, un calcul de distance entre les textes afin de mesurer précisément la plus ou moins grande ressemblance (similarité) de chacun des textes par rapport aux autres (la première partie présente cette mesure). D'autre part, des procédures de classification qui, à l'aide de ces distances, repèrent les "meilleurs groupements possibles" au sein de cette population (deuxième partie).

A cette procédure d'attribution d'auteur, s'ajoute l'examen d'une série d'indices lexicaux ou stylistiques qui pourront aussi répondre à la question : « puisque l'on est capable d'identifier l'écrivain, qu'apprend-on sur son style et son oeuvre ? » (troisième partie).

Enfin, grâce aux multiples expériences comme celles présentées dans cette note, une échelle de la distance intertextuelle permet de prendre une décision rapide concernant un texte, ou un petit groupes de textes, anonymes ou douteux – à condition de disposer de textes non douteux produits à la même époque par l'écrivain plume de l'ombre - sans avoir besoin de parcourir toutes les étapes présentées dans cette note.

Ce texte a été spécialement rédigé à l'intention des littéraires et ne demande aucune connaissance préalable en mathématique et statistique.

PREMIERE PARTIE
MESURE DES PROXIMITES ENTRE TEXTES

Un lecteur est désarmé quand il cherche à identifier celui qui a écrit un texte anonyme ou douteux car le cerveau humain n'est pas outillé pour cela. Certes, les érudits peuvent avoir des intuitions mais ils ne peuvent les fonder sur des critères intrinsèques aux textes susceptibles de caractériser un écrivain, par son vocabulaire et son style, et de le distinguer des autres auteurs contemporains. Pour le théâtre du XVIIe siècle, la question n'est pas mineure puisque plus de la moitié des pièces de théâtre n'ont pas été présentées sous le nom des écrivains qui les avaient composées mais sous le nom d'intermédiaires, souvent de riches comédiens qui se faisaient passer pour les "auteurs".

Les lunettes pallient à la myopie, le télescope et le microscope permettent à l'œil de voir plus loin ou plus près. De même, l'ordinateur peut reconnaître l'écrivain, là où notre cerveau est désarmé. En effet, celui-ci ne peut garder simultanément en mémoire des centaines de milliers de mots, pour comparer un grand nombre de textes, ce que l'ordinateur peut faire aisément.

Le premier chapitre récapitule les principales étapes du traitement des textes, les formules du calcul de distance, les propriétés de la mesure. Puis cette méthode est appliquée à trois corpus de tragédies du XVIIe (Chapitre II). Les résultats montrent que la méthode est capable d'attribuer chaque pièce sans erreur et de discriminer des œuvres différentes quand elles ont été composées par des écrivains différents (chapitre III). Elle permet de reconnaître une même plume dans les œuvres présentées par J. Racine, J. de La Chapelle et J.-G. Campistron (chapitre IV).

Cette méthode a été mise au point selon les procédures usuelles en sciences de l'ingénieur. Les étapes de cette mise au point sont récapitulées dans l'annexe 1.

CHAPITRE 1
LA DISTANCE INTERTEXTUELLE

La distance entre deux textes ("intertextuelle") est mesurée comme la distance séparant deux objets dans l'espace. L'unité de mesure n'est pas le mètre mais le "mot". Comme pour un calcul de distance dans l'espace, la qualité du résultat dépend de celle des procédures d'observation et des instruments de mesure. Si ceux-ci ne sont pas précisément définis – voire non connus comme c'est le cas pour de nombreuses recherches en "analyse des données textuelles" -, si elles fluctuent selon les lieux et les observateurs, les résultats sont sans utilité.

Ces conventions portent d'abord sur la préparation des textes et leur traitement (première section). Elles concernent ensuite la mesure de la distance et le respect de certaines contraintes dues aux propriétés de cet indice (deuxième section). Grâce à ces procédures, les principaux facteurs qui déterminent cette distance ont pu être mis au jour et pondérés (troisième section).

I. PREPARATION ET TRAITEMENT DES TEXTES

Pour le théâtre du XVIIe, la première difficulté réside dans la transcription. En effet, seules un petit nombre de pièces ont été transcrites en français contemporain, notamment celles de J. Racine par P. Mesnard et celles de P. Corneille par C. Marty-Laveaux (qui nous ont servi de modèles).

Transcription des textes en français contemporain

La langue du XVIIe est bien le français moderne dans sa structure syntaxique et dans les principaux composants de son lexique, mais elle diffère du français contemporain par la graphie et la ponctuation. Le document ci-dessous illustre cette difficulté. Il s'agit des premiers vers d'*Adrien*, l'une des tragédies parues sous le nom de J.-G. Campistron, copiés dans les *Œuvres de M. Campistron* en 1734.

ACTE I.

SCENE PREMIERE.

VALERIE, JULIE.

JULIE.

VOus vous cachez, Madame, & vous fuyez mes ſoins ;
Mes yeux ſont ils ici de profanes témoins ?
Troublent-ils la douceur de votre ſolitude ?
Parlez ; c'eſt à Julie un ſupplice trop rude
D'adorer Valerie, & de voir chaque jour
Que fuyant les plaiſirs d'une ſuperbe cour,
Elle vient en ces lieux enfevelir ſes charmes ;
Payer à ſes chagrins un tribut de ſes larmes :
Chagrins d'autant plus vifs ; que toujours renfermez. . . .

VALERIE.

Hélas !

JULIE.

Quoi, mes reſpects tant de fois confirmez,
Quoi, mon attachement & ſi pur & ſi tendre,
N'obtiendront point de vous ce que j'oſe prétendre ?

VALERIE.

Laiſſe, laiſſe, Julie ; & ne demande plus
L'aveu de ces chagrins dans mon cœur retenus ;
Qu'il les dévore ſeul.

JULIE.

Quels malheurs les font naître ?
Et pourquoi craignez-vous de les faire paroître ?
Plus j'en cherche la cauſe, & moins je l'entrevoi,
Des deſtins, votre rang ſemble braver la loi.
Fille d'un Empereur que l'univers revere,
Seul objet de l'amour de cet auguſte pere ;
Digne prix des lauriers que le fier Adrien
Moiſſonne à pleines mains pour Dioçletien,
Sûre que dès long-tems ce vainqueur vous adore ;
Aux douleurs, votre ſein peut-il s'ouvrir encore ?

VALERIE.

Hé, quel eſt le mortel parfaitement heureux ?

JULIE.

J'entens. Un tendre amour tyranniſe vos vœux.
L'abſence d'Adrien faiſoit couler vos larmes,
Mais ce jour vous promet la fin de vos allarmes :

Début de : Campistron Jean-Galbert. Adrien. *Œuvres.* Paris : Compagnie des Libraires, 1734, tome 2, p. 3-4. (Déchargée sur Google-livres en 2011).

Non seulement les 's' et les 'f' sont indifférenciés, les 'aî', les 'oî' et les 'ais' sont confondus (*faifoit* pour *faisait* ; *naître*, rime avec *paroître*, etc.), les accents sur le 'e' sont souvent omis (*témoin* ou *père* mais *Valerie* ou *revere*), le participe passé pluriel est tantôt 'es', 'és', 'ez' (*renfermez, confirmez*), etc. On trouve aussi des mots composés étranges (*long-temps*) et des orthographes singulières (*allarmes*)… La ponctuation est parfois difficile à déchiffrer, à cause de taches sur les vieux livres, et assez aberrante, notamment l'usage immodéré du point virgule dans des endroits inattendus (*Laisse, laisse, Julie ; et ne demande plus*).

Aucun scanner ne peut reconnaître un tel texte. Il est donc nécessaire de le saisir à nouveau et, plus précisément, de le transcrire en français contemporain, en suivant les conventions typographiques usuelles (ce qui permet d'offrir au lecteur les textes d'*Aétius*, de *Juba* et de *Tachmas*).

Balisage et standardisation graphique

Ceci fait, on isole par des balises tout ce qui n'est pas le texte mais qui doit être conservé. Par exemple pour le théâtre, en tête du texte figurent les indications bibliographiques : nom de l'auteur, titre de l'ouvrage, date et lieu de première présentation – quand ces renseignements sont connus – références de l'édition utilisée, identité du transcripteur et date de l'opération. Dans le corps du texte, les noms des personnages, les indications scéniques, les numéros des vers sont entourés de balises qui indiquent qu'on ne doit pas les compter comme du texte mais qu'ils sont conservés pour être restitués à l'utilisateur.

Ces opérations préalables effectuées, le texte est soumis à un logiciel qui standardise les graphies et attache à chaque mot du texte une étiquette. Voici, un vers d'*Adrien* :

Que fuyant les plaisirs d'une superbe cour,

"Que" (avec majuscule initiale) est le "mot" ou la "forme graphique" qui reste sans changement et à sa place dans le texte. Mais l'automate doit y ajouter une information indiquant qu'il s'agit d'une "majuscule initiale d'usage en début de vers" et non d'un nom propre, contrairement à Julie, Valérie, Adrien, Dioclétien qui conservent cette majuscule initiale. Il faut également identifier les majuscules données aux "faux noms propres" (Madame, Seigneur).

Cette opération ne modifie pas le texte original, elle ajoute de l'information. Par exemple, *Madame* ou *Seigneur* gardent leur majuscule — puisque l'écrivain (ou le typographe) l'ont voulu ainsi et cette graphie sera retournée à l'écran ou sur papier — mais, l'ordinateur traite ces mots comme des noms communs.

La standardisation consiste également à reconnaître les mots composés. *A priori*, n'entrent dans cette catégorie que les mots composés disposant d'une entrée dans les dictionnaires de langue. Par exemple, "a priori", "aujourd'hui" ou "bric à brac" sont considérés comme un seul mot mais pas "pomme de terre" ou "chemin de fer". Ces formes composées "libres" ne sont pas perdues pour autant : d'autres programmes permettent de les retrouver dans les vastes collections de textes comme la bibliothèque électronique du français contemporain[1]. Elles sont souvent utiles pour caractériser le style d'un écrivain.

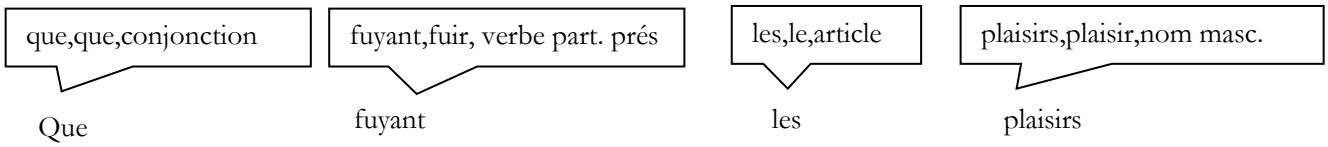
Ces graphies standardisées servent ensuite à étiqueter le texte.

Etiquetage

La seconde étape est l'étiquetage de chaque mot du texte (lemmatisation), comportant notamment la résolution des homographies (une seule graphie mais plusieurs entrées de dictionnaire). Par exemple, sur les huit mots du vers ci-dessus, cinq peuvent être rattachés à plusieurs entrées de dictionnaire :

- 'que' : conjonction (que) ou pronom (que)
- 'fuyant' : verbe (fuir) au participe présent ou adjectif (fuyant)
- 'les' : article (le) ou pronom (le)
- 'une' : article (un) ou pronom (un)
- 'superbe' : nom féminin (superbe) ou adjectif (superbe) ?

Ci-dessous l'étiquetage de ces mots par le logiciel mis au point dans les années 1980 et perfectionné depuis lors[2].



"Que" est le "mot" ou la "forme graphique brute" qui reste sans changement et à sa place dans le texte. Dans l'étiquette, le premier "que" est la "forme graphique normalisée". Dans un texte comme celui présenté ci-dessus, plus d'un mot sur huit a une graphie standard différente de celle présente dans le texte. Que dirait-on de mesures réalisées sur des observations dont une sur

[1] Pibarot André, Picard Jacques & Labbé Dominique (1998). Les syntagmes répétés dans l'analyse des commentaires libres. In Mellet Sylvie (ed). *4e Journées d'analyse des données textuelles*. Nice, 1998, p. 507-516.

[2] Labbé Dominique. *Normes de saisie et de dépouillement des textes politiques*. Grenoble, 1990, Cahiers du CERAT.

huit seraient fausses ? C'est pourtant ce que font les logiciels "d'analyse des données textuelles" puisqu'aucun ne réduit les majuscules initiales des mots communs comme 'que' (début de vers).

La seconde partie de l'étiquette ("que, conjonction") est son "entrée de dictionnaire" ou "lemme" (mot vedette et catégorie grammaticale). Par exemple le mot vedette d'un verbe est son infinitif, celui des déterminants ou des adjectifs, leur masculin singulier, le singulier du nom, etc.

Par exemple :

— sous l'entrée "le, article", ou "le, pronom", on trouve : L', La, Le, Les, l', la, le, les… Mais pas la note de musique "la" ni "La Chapelle (Monsieur de…)"

— sous l'entrée "fuir, verbe", on trouve : fuis, fuirai, fuyais… mais pas : "fuyant, adjectif" ;

La lemmatisation est complète et univoque (tout mot reçoit une étiquette et une seule) stable (les mêmes conventions sont strictement appliquées du début à la fin de l'opération).

Cette **nomenclature** fait consensus parmi les usagers de la langue. Elle est en usage dans tous les dictionnaires de langue. Par exemple, en français, toutes les flexions d'un verbe (modes, temps, personnes) sont regroupées sous l'infinitif de ce verbe. Mais lorsqu'elle est implantée dans un programme informatique, elle est systématique, c'est-à-dire qu'elle explicite et "durcit" ces conventions. Par exemple,

— si l'on distingue certains substantifs par leur genre (garde, livre, mode, page, tour…) alors tous les substantifs doivent avoir un genre et tous les substantifs de même graphie et de même genre doivent être regroupés sous une seule entrée ;

— toutes les formes verbales ayant même infinitif doivent être groupées ensemble (voler, intransitif et transitif : une seule entrée). Sinon, il faudrait distinguer les emplois transitifs et intransitifs de tous les verbes…

La lemmatisation doit être aussi automatique que possible, ce qui oblige à renoncer à fournir certaines informations (comme le partage entre l'indicatif et le subjonctif présent des verbes du premier groupe). A condition de s'en tenir à une nomenclature synthétique (type dictionnaire de langue), en moyenne 99% des mots peuvent être étiquetés automatiquement. Le "résidu" n'est pas négligeable : pour l'étiquetage des œuvres de Corneille, Molière et Racine, cela représente plus de 10 000 cas à résoudre manuellement. Les principales difficultés résident dans les mots les plus usuels ("suis", "tout", "même"…) et dans le caractère idiomatique de la langue : toute "règle" a des exceptions.

Toute intervention manuelle étant source d'erreurs ou de fluctuation dans les décisions, l'algorithme offre des solutions pertinentes, pour les cas non-résolus, et limite au maximum les possibilités d'erreur.

En effet, c'est le principe le plus important : l'étiquetage est sans erreur, par rapport aux conventions retenues qui, elles-mêmes, sont entièrement explicites. Les étiquetages fantaisistes et lacunaires sont contre-productifs car ils engendrent le scepticisme sur l'ensemble des informations contenues dans la base et des conclusions tirées de celles-ci.

II. CALCUL DE LA DISTANCE

Un exposé détaillé - en français et destiné aux non-mathématiciens - est disponible en ligne dans *Images des mathématiques*, revue des mathématiciens du CNRS destinée à un large public[1]. En voici un résumé.

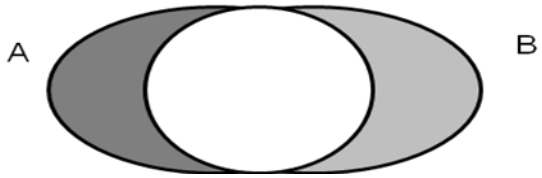
Principes du calcul

Soit deux textes A et B.

A Que fuyant les plaisirs de la superbe cour.

B De la superbe cour, il fuit les plaisirs.

Ces deux textes sont superposés et on compte le nombre de mots différents (zones grisées dans le schéma ci-dessous).



On note :

- N_A et N_B : nombre de **mots** ("tokens" en anglais) dans A et respectivement B, ou **longueurs** de A et de B, ici 8 mots dans les deux cas ;

- V_A et V_B : nombre de "**vocables**" ("types" en anglais) dans A et respectivement B. C'est l'étendue de leurs vocabulaires respectifs : il y a 7 vocables différents dans A et autant dans B. $V_{(A,B)}$ est le vocabulaire total de A et B ;

- F_{iA} et F_{iB} : nombre de fois qu'un vocable i est utilisé dans A et respectivement B. Ce sont les **effectifs** ou les "fréquences absolues" de ce vocable. Dans l'exemple, les effectifs sont tous de 1 sauf pour l'article "le" (employé 2 fois dans A et autant dans B) ;

---

[1] Labbé Cyril & Labbé Dominique. 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 images*. (http://images.math.cnrs.fr/La-classification-des-textes.html). 28 mars 2011.

- $|F_{iA} - F_{iB}|$ la différence absolue des effectifs du vocable *i* dans *A* et dans *B*. L'adjectif "absolue" signifie que l'on ne tient pas compte du signe dans le résultat. Dans l'exemple ci-dessus, cette différence absolue est de 1 pour "il" et "que".

- $D_{(A,B)}$: la **distance** entre *A* et *B*.

Cette distance est le **nombre de mots différents** dans *A* par rapport à *B* (ou réciproquement). Pour calculer cette distance, les 8 vocables constituant le vocabulaire de *A* et de *B* – ensemble noté $V_{(A,B)}$ - sont rangés par ordre alphabétique dans le tableau 1 (semblable au tableau de calcul créé en mémoire de l'ordinateur) :

Tableau 1. Tableau de calcul de la distance entre les textes *A* et *B*

i =	Vocabulaire de A et B	F_{iA}	F_{iB}	$\|F_{iA} - F_{iB}\|$
1	cour (nom féminin)	1	1	0
2	de (préposition)	1	1	0
3	fuir (verbe)	1	1	0
4	il (pronom)	0	1	1
5	le (article)	2	2	0
6	plaisir (nom masculin)	1	1	0
7	que (conjonction)	1	0	1
8	superbe (adjectif)	1	1	0
	Total	8	8	2

Dans la troisième colonne : l'effectif du vocable d'indice *i* dans *A* ; dans la quatrième colonne, son effectif dans *B*, et, en dernière colonne, la différence absolue de ces 2 effectifs. La dernière ligne donne les résultats. La longueur de *A* (N_A), comme de *B* (N_B) est de 8 mots ($N_A = N_B$). La distance absolue entre *A* et *B* ($D_{(AB)}$) est égale à 2 mots.

Ces opérations sont résumées par la formule suivante :

$$(1)\ D_{(A,B)} = \sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - F_{iB} \right| \quad avec\ N_A = N_B$$

Dans cette formule, la lettre majuscule sigma signifie "somme" et les notations en dessous et en dessus de ce symbole signifient "effectuer le calcul, figurant à droite de ce symbole, pour les vocables de rang *i* appartenant à *A* et/ou *B*, avec *i* variant de 1 à $V_{(A,B)}$".

La distance absolue entre A et B est donc de 2 mots.

Pour pouvoir comparer les résultats obtenus sur des populations importantes de textes, la distance relative est calculée :

$$(2)\ D_{rel(A,B)} = \frac{\sum_{i\in(A,B)}^{V_{(A,B)}} \left| F_{iA} - F_{iB} \right|}{N_A + N_B}$$

Dans l'exemple ci-dessus, la distance relative entre A et B est 2/16 ou encore 0.125.

Cet indice varie entre 0 (mêmes vocables avec les mêmes effectifs dans les deux textes) et 1 aucun mot en commun. Cette variation est uniforme (ni seuil ni saut).

$D_{(A,B)}$ est une **distance euclidienne** (longueur du segment de droite unissant deux points). L'adjectif "euclidien" signifie "conforme à la géométrie d'Euclide" (par un point il ne passe qu'une parallèle à une droite située hors de ce point). Les propriétés d'une distance euclidienne sont :

- l'**identité** (la distance d'un point à lui-même est nulle),

- la **symétrie** (le résultat est le même que l'on mesure *AB* ou *BA*), ce qui dispense d'utiliser les vecteurs,

- l'**inégalité triangulaire** (le chemin direct entre *A* et *B* est toujours plus court qu'en passant par un point *C* n'appartenant pas au segment *AB*).

Ces propriétés ont d'importantes conséquences. Par exemple, on peut réaliser une représentation graphique de toutes les distances au sein d'une vaste population de textes, comme on dresse la carte d'une ville ou d'un quartier. Cette "cartographie" des bases de textes sera évoquée dans la seconde partie de ce rapport.

Trois remarques :

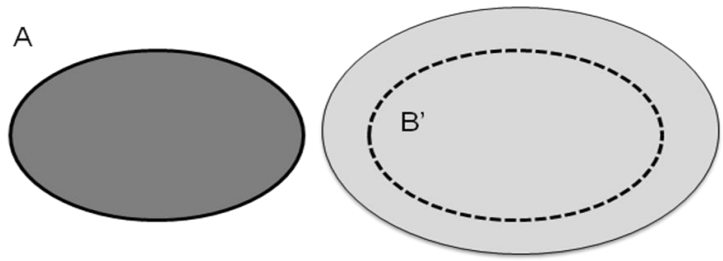
- le calcul porte sur la totalité des vocables et leurs effectifs, autrement dit sur l'ensemble du texte ("intertextuelle"),

- l'ordre des mots dans le texte importe peu, seule compte l'idée véhiculée (comme le maître de rhétorique l'explique à M. Jourdain),

- la longueur des textes est la même. Généralement ce n'est pas le cas. La formule a donc été adaptée pour s'appliquer à des textes de longueurs différentes.

Calcul sur des textes de longueurs différentes

Dans le cas de deux textes de longueurs inégales ($N_A < N_B$), la distance est estimée en "réduisant" B à la longueur de A (schéma ci-dessous) puis en superposant A et B' (comme dans le schéma précédent) et en comptant le nombre de mots différents entre A et B'.



Soit :

- *U* : le **rapport des longueurs** de *A* et *B*, c'est-à-dire la proportion dont il faut réduire *B* pour obtenir *B'* (ou "coefficient de proportionnalité") :

$$U = \frac{N_A}{N_B}$$

- $E_{iA(u)}$: **l'effectif théorique** dans un texte *B'* de la longueur de *A* d'un vocable *i* appartenant au vocabulaire de *B*. Cet effectif théorique est obtenu en pondérant l'effectif de *i* dans *B* par *U* (formule 3) :

$$(3)\ E_{iA(u)} = F_{iB} * U \quad avec U = \frac{N_A}{N_B}$$

Pour chacun des vocables de *B*, la formule (3) permet de calculer le nombre de fois que ce vocable apparaîtrait si *B* avait la longueur de *A*. En remplaçant, dans la formule (1), l'effectif de chacun des vocables de *B* par cet effectif théorique, on obtient une **estimation** de la distance intertextuelle (formule 4) :

$$(4)\ D_{(A,B')} = \sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - E_{iA(u)} \right|$$

Pour le calcul de la distance relative, on remplace, dans la formule (2) N_B par la somme des effectifs théoriques, c'est-à-dire la longueur théorique de *B'* ($N_{B'}$) :

$$N_{B'} = \sum_{i \in B}^{V_B} E_{iA(u)}$$

Aux arrondis près, $N_{B'}$ est égale à N_A. La formule (2) devient :

$$(5)\ D_{rel(A,B')} = \frac{\sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - E_{iA(u)} \right|}{N_A + N_{B'}}$$

Il s'agit d'une **estimation** et ceci pour au moins deux raisons.

La première raison a déjà été aperçue à propos du calcul de la distance relative. Les effectifs dans *A* sont des entiers naturels et les effectifs théoriques dans *B'* des rationnels *approchant* des entiers naturels (inconnus). Autrement dit, le résultat de la soustraction - au numérateur de (4) et (5) - comportera des décimales sans signification mais qui seront pourtant additionnées pour obtenir la distance… Ces décimales pèseront d'autant plus lourd que le vocable considéré aura des effectifs faibles - observés dans *A* et théoriques dans *B'*. Or, dans tout texte en langue naturelle, les vocables qui n'apparaissent qu'une fois sont toujours plus nombreux que ceux survenant deux fois, eux-mêmes plus nombreux que les effectifs trois, etc. Le fait que dans les formules (4) et (5), on cumule des différences absolues ne permet pas à ces "erreurs" de s'annuler. Au contraire, elles se cumuleront. Au passage, on remarquera que cette caractéristique est considérablement aggravée quand on élève au carré le résultat de la soustraction au numérateur de (4). C'est ce que font les "analyses en composantes principales" ou l'"analyse factorielle des correspondances", sans que les usagers aient toujours conscience des déformations massives auxquelles conduit cette élévation au carré.

Pour limiter cet inconvénient, on élimine du calcul :

- Les vocables absents de *A* et pour lesquels l'effectif théorique dans *B'* est inférieur à 1. La formule (3) devient :

$$\text{(3 bis)}\quad E_{iA(u)} = \begin{cases} 0 & si\ \ F_{iA} = 0\ \ et\ \ F_{iB} * U < 1 \\ F_{iB} * U & si\ \ F_{iA} > 0\ ou\ F_{iB} * U \geq 1 \end{cases}$$

- La différence des effectifs observés en *A* et des effectifs théoriques en *B* lorsque celle-ci est inférieure à 0.5. En effet, puisqu'il s'agit d'estimer un entier, ce résultat équivaut à zéro. La formule (4) devient :

$$\text{(4 bis)}\quad D_{(A,B')} = \sum_{i \in (A,B)}^{V_{(A,B)}} \left| F_{iA} - E_{iA(u)} \right| \ \text{ avec } \left| F_{iA} - E_{iA(u)} \right| = 0 \ \ si\ \ \left| F_{iA} - E_{iA(u)} \right| < 0.5$$

La formule (5) est complétée pour intégrer ces deux éléments.

La seconde raison tient aux postulats qui fondent le calcul de l'effectif théorique d'un vocable dans *B'* (formule 3 bis). Cette formule repose sur deux postulats[1].

[1] Labbé Cyril, Labbé Dominique et Hubert Pierre. Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics*, 11-3, 2004, p. 193-213.

- premier postulat : l'effectif d'un vocable augmente proportionnellement à l'allongement du texte. Ce premier postulat n'est valable que pour les mots les plus fréquents et non-spécialisés (ces derniers surviennent par paquets dans les passages où sont traités les thèmes auxquels ils appartiennent),

- deuxième postulat : l'apparition des vocables nouveaux se fait toujours au même rythme. En réalité, ce rythme est très rapide au début du texte – donc la formule (3 bis) ne peut pas s'appliquer à des textes trop courts – puis il décline ensuite lentement sans jamais se stabiliser. Dès lors la formule (5) n'est pleinement valable que lorsque les deux textes comparés ne sont pas de longueurs trop différentes et lorsque la longueur du plus court excède le point à partir duquel le rythme d'apparition des mots nouveaux devient sensiblement linéaire.

Limites du calcul

Une série d'expériences – dont certaines sont évoquées en annexe 1 - indiquent que :

- les deux textes doivent avoir plus de 1 000 mots, et que, en-dessous de 5 000 mots, le résultat de (5) peut être instable,

- le rapport (U) doit être inférieur à 1 : 10. En fait, plus ce rapport s'élève, plus le résultat doit être examiné avec prudence.

- dans ces limites, l'incertitude qui pèse sur la distance estimée est comprise entre ±1% (90% des valeurs sont comprises dans cet intervalle), avec des textes de longueurs supérieures à 5 000 mots et avec $U \leq 0.5$) et ±5%, avec la longueur du petit texte au moins égal à 5 000 mots et lorsque $U = 0.2$ (rapport de 1:5).

Les textes utilisés dans cette recherche comptent pour la plupart au moins 10 000 mots et leurs longueurs sont assez proches (voir les corpus en annexe 2). Les résultats sont donc présentés avec trois décimales. Du fait de l'incertitude introduite par l'estimation de la distance, la dernière décimale indique dans quel sens arrondir la seconde décimale qui est la dernière significative.

III. QUELS SONT LES FACTEURS QUI DETERMINENT LA DISTANCE ?

Les expériences effectuées pour mettre au point la méthode ont permis d'identifier et de mesurer l'importance des principaux facteurs qui déterminent la distance entre textes. Par importance décroissante, il s'agit de :

- le genre : oral et écrit, prose, vers, comédie et tragédie, etc.
- l'écrivain,
- l'époque où a été rédigé le texte car chaque époque possède un vocabulaire particulier et le lexique de chaque écrivain évolue avec le temps,
- le thème (vocabulaire propre à ce thème, nom des personnages, lieux, principaux motifs).

Le modèle ci-dessous systématise ces constats :

$$D_{(A,B)}=f\{Dmin_{(A,B)},\ (Genre_A,Genre_B),\ (Auteur_A,Auteur_B),\ (Epoque_A,Epoque_B),\ (Thème_A,Thème_B)\}$$

Le symbole *f()* signifie que la distance est **fonction** des termes indiqués entre crochets. Ces termes sont nommés "variables", car elles peuvent être chiffrées. Pour donner un **poids** à chacune de ces variables, on recherche des cas où toutes les autres sont nulles ou négligeables (raisonnement "toutes choses égales par ailleurs").

Par exemple, en utilisant des textes appartenant au même genre (théâtre, poésie, roman, correspondance, etc.), écrits à la même époque, on peut isoler l'importance relative de l'écrivain et du thème. Parfois, on a la chance que deux écrivains contemporains traitent le même thème, en aveugles, dans le même genre. Cette situation est assez courante au XVIIe ; par exemple : les deux *Mère coquette* – T. Corneille (sous le nom de P. Quinault) et J. Donneau de Visé en 1665 - les deux *Bérénice* de P. Corneille et J. Racine (en 1670) ; les deux *Phèdre* (J. Racine et J. Pradon en 1677) ; les deux *Comte d'Essex* (T. Corneille et C. Boyer en 1678) : alors il ne reste plus que le facteur auteur, ce qui révèle parfois des situations intéressantes.

Le chapitre VII de la présente note revient sur le poids spécifique du temps sur la distance entre textes écrits dans un même genre par les mêmes écrivains.

La conclusion essentielle est la suivante : dans un genre et à une époque donnée, la variable "auteur" l'emporte sur le thème et sur l'influence du temps, du moins quand celui-ci ne se compte pas en plusieurs dizaines d'années. Dès lors, pour déterminer le véritable auteur d'un texte d'origine douteuse ou inconnue, il suffit de le confronter à d'autres – *dont l'origine n'est pas douteuse* - écrits dans un même genre et à la même époque. NB : le théâtre doit être comparé au théâtre et les tragédies entre elles, les comédies entre elles, etc.

CHAPITRE II.
UN EXEMPLE : LES FRERES CORNEILLE ET J. RACINE

Pour exposer en détail la méthode, il a paru judicieux de commencer par une mise en oeuvre sur des textes qui ne posent apparemment pas de problème et dans des situations telles que, selon le modèle présenté dans le chapitre précédent, la variable auteur soit prépondérante (textes contemporains écrits dans le même genre).

Pierre Corneille (1606-1684) et Thomas Corneille (1625-1709) sont ceux qui ont produit – sous leur nom - le plus grand nombre de tragédies contemporaines de celles de Jean Racine (1639-1699). Ces pièces appartiennent au même genre, la tragédie classique en alexandrins. Elles suivent les mêmes conventions et traitent des thèmes proches – voire identiques dans le cas des deux *Bérénice*. Ces pièces sont destinées au public parisien, et parfois aux mêmes acteurs (la troupe de l'Hôtel de Bourgogne crée la plupart d'entre elles).

Enfin, les frères Corneille font les mêmes études, épousent deux sœurs, les deux ménages vivent sous le même toit et font bourse commune. Ils fournissent ainsi un étalon de la proximité maximale pouvant exister entre deux auteurs différents.

Ces trois corpus offrent une situation idéale où tous les facteurs autres que l'écrivain sont neutralisés (genre) ou le plus faible possible (temps et thème).

Voici les tragédies des frères Corneille – contemporaines des pièces présentées par J. Racine - qui ont été utilisées dans cette expérience (voir également la description des corpus en annexe 2) :

- Pierre, les tragédies présentées après son retour au théâtre en 1659 : *Œdipe* (1659), *la Toison d'or* (1661), *Sertorius* (1662), *Sophonisbe* (1663), *Othon* (1664), *Agésilas* (1666), *Attila* (1667), *Tite et Bérénice* (1670), *Pulchérie* (1672), *Suréna* (1674).

- Thomas : les tragédies présentées durant la même période *Stilicon* (1660), *Camma* (1661), *Persée et Démétrius* (1661), *Maximian* (1662), *Pyrrhus* (1663), *la Mort d'Annibal* (1669), *Ariane* (1672), *la Mort d'Achille* (1673), *le Comte d'Essex* (1678)[1].

- J. Racine : *la Thébaïde* (1664), *Alexandre* (1665), *Andromaque* (1667), *Britannicus* (1669), *Bérénice* (1670), *Bajazet* (1672), *Mithridate* (1672), *Iphigénie* (1674), *Phèdre* (1677).

Ces pièces vont être utilisées pour illustrer la méthode d'attribution d'auteur et pour la mettre à l'épreuve.

[1] Ces pièces ont été transcrites en français contemporains – en suivant les mêmes conventions que celles appliquées aux œuvres de J. Racine par P. Mesnard (annexe 1) et aux œuvres de P. Corneille par C. Marty-Laveaux. Leur analyse sera présentée dans un ouvrage ultérieur.

La confrontation entre les dix-neuf œuvres des frères Corneille et les neuf tragédies contemporaines présentées par J. Racine, depuis la *Thébaïde* (1663) jusqu'à *Phèdre* (1677), donne au total (28*27)/2 = 378 couples différents et autant de distances. Le tableau des distances (28 colonnes x 28 lignes) est trop grand pour être reproduit. L'analyse va se concentrer sur les principales zones critiques de ce grand tableau : les trois œuvres séparément puis la comparaison de J. Racine d'abord à P. Corneille puis à T. Corneille.

Cela donne cinq tableaux, comportant chacun deux zones : les distances séparant les textes d'un même écrivain (ou "distances intra") et les distances séparant les textes de cet écrivain à un des deux autres (ou "distances inter").

I. DISTANCES INTRA-CORPUS

Examinons d'abord les distances entre pièces d'un même écrivain (distances "intra") : 45 non nulles pour P. Corneille ; 36 pour T. Corneille et pour J. Racine.

Pierre Corneille

Le tableau 1 ci-dessous donne les résultats du calcul des distances sur les dix tragédies de la dernière période de création de P. Corneille.

Tableau 1. Les distances internes au corpus "Tragédies de P. Corneille (1659-1674)" (Classement chronologique).

	Oedipe	Toison d'	Sertorius	Sophonisbe	Othon	Agésilas	Attila	Tite et B.	Pulchérie	Suréna
Œdipe	0,000	0,194	0,196	0,190	0,194	0,211	0,196	0,208	0,206	0,194
Toison d'or	0,194	0,000	0,187	0,194	0,199	0,201	0,191	0,197	0,203	0,201
Sertorius	0,196	0,187	0,000	0,159	0,177	0,173	0,177	0,171	0,173	0,173
Sophonisbe	0,190	0,194	0,159	0,000	0,171	0,175	0,188	0,177	0,179	0,180
Othon	0,194	0,199	0,177	0,171	0,000	0,179	0,169	0,163	0,158	0,174
Agésilas	0,211	0,201	0,173	0,175	0,179	0,000	0,186	0,159	0,165	0,162
Attila	0,196	0,191	0,177	0,188	0,169	0,186	0,000	0,180	0,176	0,178
Tite et B.	0,208	0,197	0,171	0,177	0,163	0,159	0,180	0,000	0,153	0,156
Pulchérie	0,206	0,203	0,173	0,179	0,158	0,165	0,176	0,153	0,000	0,158
Suréna	0,194	0,201	0,173	0,180	0,174	0,162	0,178	0,156	0,158	0,000
Moyenne	0,199	0,196	0,176	0,179	0,176	0,179	0,182	0,174	0,175	0,175

La diagonale est nulle (première propriété d'une distance euclidienne : la distance d'un

individu à lui-même est nulle). Cette diagonale sépare le tableau en deux parties symétriques : le triangle supérieur droit contient les mêmes valeurs que le triangle inférieur gauche. C'est la deuxième propriété d'une distance euclidienne (symétrie) : la distance entre deux points A et B est identique que l'on mesure de A vers B ou de B vers A (ce qui permet de ne pas avoir recours aux vecteurs).

Ces distances sont rangées par ordre croissant et groupées dans des classes d'intervalles égaux (correspondant au seuil de précision de la mesure ici : 0.01). Trois valeurs centrales caractérisent cette série statistique :

- la distance médiane (Me) : 0,178 (la moitié des distances sont inférieures à 0.18 et l'autre moitié lui sont supérieures) ;

- la distance modale (Mo) ou valeur la plus fréquente (c'est-à-dire la classe de distances dont les effectifs sont les plus nombreux, ici 0.18) ;

- la distance moyenne ($\overline{D}$) : 0,181, soit 0,18.

Contrairement à une idée reçue, la moyenne n'est pas "au milieu" de la série, sauf dans un cas particulier : une distribution symétriquement et harmonieusement répartie autour de la moyenne – généralement en forme de "cloche", ce qui indique une population homogène dont la plupart des individus peuvent être considérés comme obéissant à une même loi de distribution. Dans ce cas, la moyenne est égale à la médiane et toutes deux sont comprises dans la classe modale. Aux arrondis près, c'est le cas ici. Sous réserve d'une vérification graphique, on peut considérer que les distances du tableau 1 sont distribuées de manière "normale" (c'est-à-dire qu'elles sont régies par une "loi normale" dite aussi de "Laplace-Gauss") ;

L'écart-type est la mesure standard de la dispersion des distances autour de la moyenne (moyenne quadratique des écarts entre les valeurs observées et la moyenne arithmétique de celles-ci). Les distances étant classées de 1 à *n*, soit D_i la *ième* distance, l'écart type (σ) est égal à :

$$\sigma = \sqrt{\frac{\sum_{1}^{n}\left(D_i - \overline{D}\right)^2}{n}} = 0{,}0154$$

La variation relative (des valeurs observées) autour de la moyenne est donnée par le coefficient :

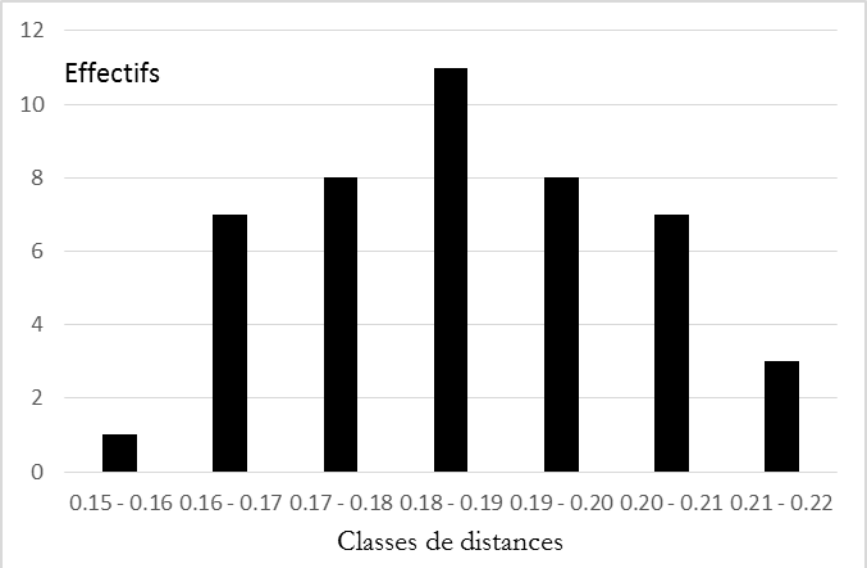
$$V = \frac{\sigma}{\overline{D}} * 100 = 8{,}4\%$$

Ce coefficient de variation relative signifie que, si les valeurs observées sont "normalement" distribuées autour de la moyenne, les deux tiers des valeurs sont comprises dans un intervalle égal à ± 8,4% de celle-ci et 95% autour de ± 16,7%. Plus ce coefficient est faible, plus les valeurs observées s'écartent peu de leur moyenne et plus celle-ci peut être considérée comme

représentative de la population entière.

Pour que le raisonnement soit pleinement valable, il faut s'assurer que la distribution est "normale". Cela se caractérise de trois manières. Premièrement, les paramètres centraux (Me, Mo et $\overline{D}$) sont égaux ou très proches. C'est le cas ici (0,18). Deuxièmement, les valeurs observées se distribuent harmonieusement autour de la moyenne. Le moyen le plus simple de le vérifier est de tracer l'histogramme des valeurs observées (tableau 2). La hauteur des barres indique les effectifs absolus de chacune des classes de distances (ces effectifs sont indiqués sur l'axe vertical à gauche du graphique).

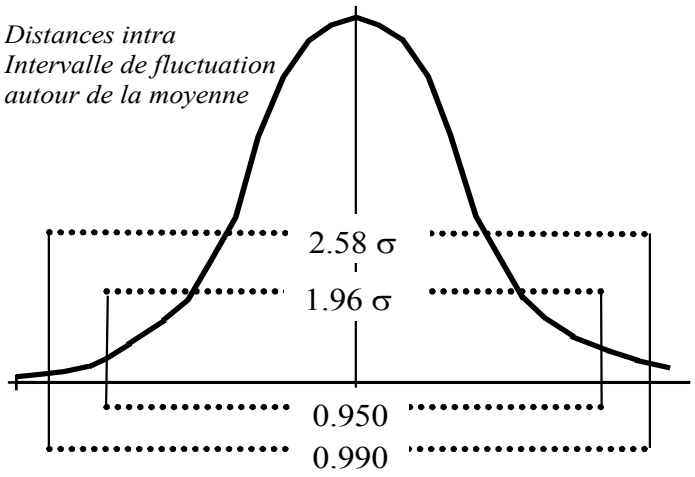
Tableau 2. Histogramme des distances entre les tragédies de P. Corneille (1659-1674)



Cette figure présente un profil "en cloche", dit "gaussien" qui confirme une distribution "normale" – c'est-à-dire une population obéissant à une même loi de distribution de paramètres $\overline{D}$ et σ. Cela signifie que les variations de la variable autour de la moyenne peuvent être considérées comme le fait du hasard (ou de plusieurs causes qui peuvent être négligées).

Troisième caractéristiques d'une distribution "normale" : 95% des valeurs sont comprises dans un intervalle égal à plus ou moins deux écarts-types (exactement 1,96) autour de la moyenne arithmétique et 99% dans un intervalle égal à plus ou moins 2,58 écarts types (schéma de principe ci-dessous).

Tableau 3. Schéma de principe des intervalles de fluctuation normale en cas de distribution normale des valeurs observées.



Si effectivement les distances intra-P. Corneille suivent une loi normale, on s'attend à ce que 95% des distances soient comprises entre 0,15 et 0,21 et 99% entre 0,14 et 0.22. Dans le tableau 1, on constate effectivement que :

- 94% des distances sont inférieures ou égales à 0.20

- toutes les distances sont supérieures à 0.14 et aucune n'atteint 0.210. Autrement dit, l'intervalle à 99% contient toutes les distances.

Une distribution normale des valeurs signifie qu'une même loi de distribution est en œuvre dans l'ensemble de la population étudiée et que les écarts existant entre les individus qui la composent sont le fait du hasard (ou de facteurs perturbateurs réguliers et suffisamment faibles pour être négligés).

L'un de ces facteurs perturbateurs est facile à identifier. En effet, les distances proches ou égales à 0.21 concernent toutes la première pièce *Œdipe* (1659) comparée à *Agésilas* (1666), *Tite et Bérénice* (1670), *Pulchérie* (1672) ce qui suggère une dimension chronologique (les pièces les plus éloignées dans le temps sont un peu plus distantes entre elles). A contrario, la plus petite distance sépare *Tite et Bérénice* (1670) de *Pulchérie* (1672).

Autrement dit, les tragédies présentées par P. Corneille durant la dernière partie de sa vie (sur une quinzaine d'années) sont sœurs mais leur père aurait légèrement changé ses habitudes entre 1659 et 1674 (ce qui sera vérifié dans le chapitre VII). Toutefois, ce léger changement ne remet pas en cause la distribution normale des distances.

Thomas Corneille

Les distances intra semblent semblables à celles observées chez son frère (tableau 4).

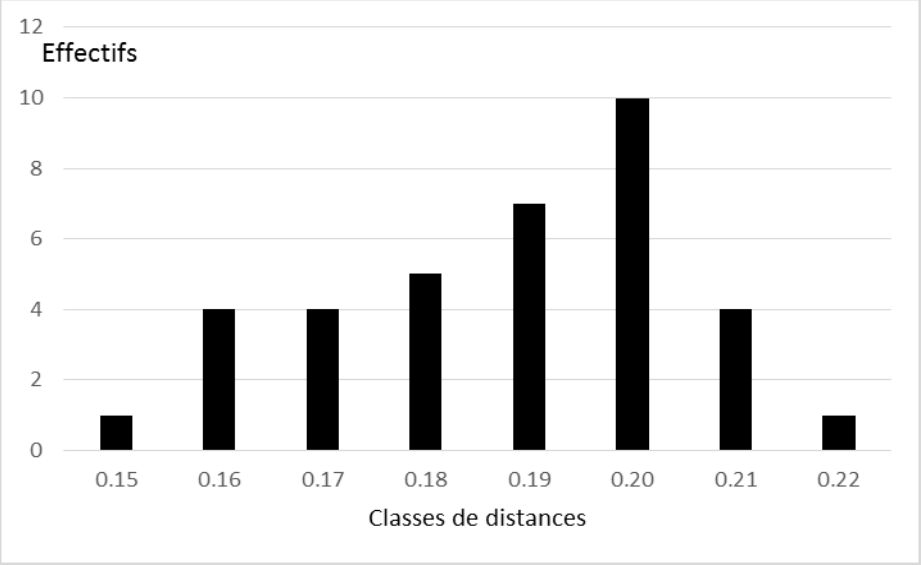
Tableau 4. Les distances internes au corpus "Tragédies de T. Corneille (1659-1678)" (Classement chronologique).

	Stilicon	Camma	Persée	Maximian	Pyrrhus	Annibal	Ariane	Achille	Essex
Stilicon	0,000	0,166	0,165	0,153	0,184	0,208	0,224	0,212	0,201
Camma	0,166	0,000	0,176	0,166	0,176	0,202	0,200	0,201	0,199
Persée	0,165	0,176	0,000	0,162	0,161	0,180	0,212	0,191	0,205
Maximian	0,153	0,166	0,162	0,000	0,171	0,203	0,206	0,197	0,194
Pyrrhus	0,184	0,176	0,161	0,171	0,000	0,190	0,202	0,186	0,200
Annibal	0,208	0,200	0,180	0,203	0,190	0,000	0,219	0,206	0,213
Ariane	0,224	0,200	0,212	0,206	0,202	0,219	0,000	0,190	0,194
Achille	0,212	0,201	0,191	0,197	0,186	0,206	0,190	0,000	0,196
Essex	0,201	0,199	0,205	0,194	0,200	0,213	0,194	0,196	0,000
Moyenne	0,189	0,186	0,181	0,181	0,184	0,203	0,206	0,197	0,200

Médiane : 0.196 ; moyenne : 0.192 ; mode : 0.20 ; écart-type : 0.0177.

La moyenne est légèrement plus élevée que chez Pierre mais la dispersion est du même ordre (9%). En s'en tenant au modèle présenté dans le chapitre précédent, il y aurait donc chez T. Corneille une diversité de thèmes et un renouvellement légèrement plus importants que chez son frère aîné. Le léger décalage existant entre les trois paramètres centraux (Me, Mo et $\overline{D}$) suggère une série moins homogène que chez P. Corneille. L'histogramme des distances (tableau 5) illustre cette caractéristique.

Tableau 5. Histogramme des distances entre les tragédies de T. Corneille (1659-1678)



Le mode (0,20) se trouve décalé par rapport à la moyenne (0,19) indiquant une asymétrie de la série sur la gauche et l'existence d'un phénomène perturbateur, en l'occurrence le temps (dont l'effet était déjà visible chez Pierre). En effet, les couples séparés par les plus faibles distances le sont également dans la chronologie. La plus petite distance sépare *Stilicon* (1660) de *Maximian* (1662), puis l'on trouve *Camma* (1661) - *Maximian* (1662), *Persée-Pyrrhus*, etc. Plus les deux pièces sont séparées par un nombre d'années importants, plus leur distance est grande. La plus forte distance sépare *Stilicon* (1660) d'*Ariane* (1672) puis *Camma* à *Arianne* et *Camma* à la *Mort d'Achille*. Autrement dit, les pièces de T. Corneille évoluent avec le temps de manière plus nette que chez son frère. Le prochain chapitre présentera un autre facteur perturbateur : une influence, voire une collaboration possible de Pierre à quelques pièces présentées par Thomas.

Ces deux facteurs perturbateurs ne remettent pas en cause la distribution normale des distances. Non seulement les valeurs centrales sont très proches mais 95% des distances internes au corpus T. Corneille sont comprises entre 0,157 et 0.225 (intervalle de fluctuation normale) ; et la totalité entre 0.148 et 0.236 (intervalle à 99%). Ce qui indique une population proche de la distribution gaussienne, donc un écrivain unique.

En est-il de même chez J. Racine ?

Jean Racine

La totalité de l'œuvre tragique présentée par J. Racine entre 1663 et 1691 est examinée en suivant la même méthode. Le résultat n'est pas aussi simple et clair que chez les deux frères Corneille (tableaux 6 et 7). En gras dans le tableau 6, les valeurs supérieures à 0.25 qui, à l'aune des frères Corneille, peuvent être considérées comme *anormalement* élevées.

Tableau 6. Les distances intra (tragédies présentées par J. Racine entre 1664 et 1691) (Classement chronologique)

	Thébaïde	Alexandre	Androma	Britannicu	Bérénice	Bajazet	Mithridat	Iphigénie	Phèdre	Esther	Athalie
Thébaïde	0,000	0,242	0,245	**0,260**	**0,276**	**0,258**	0,242	**0,255**	**0,275**	**0,317**	**0,295**
Alexandre	0,242	0,000	0,231	0,233	**0,260**	0,251	0,238	0,241	**0,266**	**0,315**	**0,295**
Andromaque	0,245	0,231	0,000	0,214	0,227	0,202	0,208	0,222	0,245	**0,331**	**0,306**
Britannicus	**0,260**	0,233	0,214	0,000	0,209	0,206	0,218	0,222	0,246	**0,314**	**0,302**
Bérénice	**0,276**	**0,260**	0,227	0,209	0,000	0,220	0,206	0,226	0,250	**0,346**	**0,329**
Bajazet	**0,258**	0,251	0,202	0,206	0,220	0,000	0,204	0,230	0,244	**0,325**	**0,304**
Mithridate	0,242	0,238	0,208	0,218	0,206	0,204	0,000	0,193	0,224	**0,313**	**0,288**
Iphigénie	**0,255**	0,241	0,222	0,222	0,226	0,230	0,193	0,000	0,216	**0,293**	**0,280**
Phèdre	**0,275**	**0,266**	0,245	0,246	0,250	0,244	0,224	0,216	0,000	**0,286**	**0,275**
Esther	**0,317**	**0,315**	**0,331**	**0,314**	**0,346**	**0,325**	**0,313**	**0,293**	**0,286**	0,000	0,215
Athalie	**0,295**	**0,295**	**0,306**	**0,302**	**0,329**	**0,304**	**0,288**	**0,280**	**0,275**	0,215	0,000
Moyenne	0,267	0,257	0,243	0,242	0,254	0,244	0,233	0,238	0,253	0,306	0,289

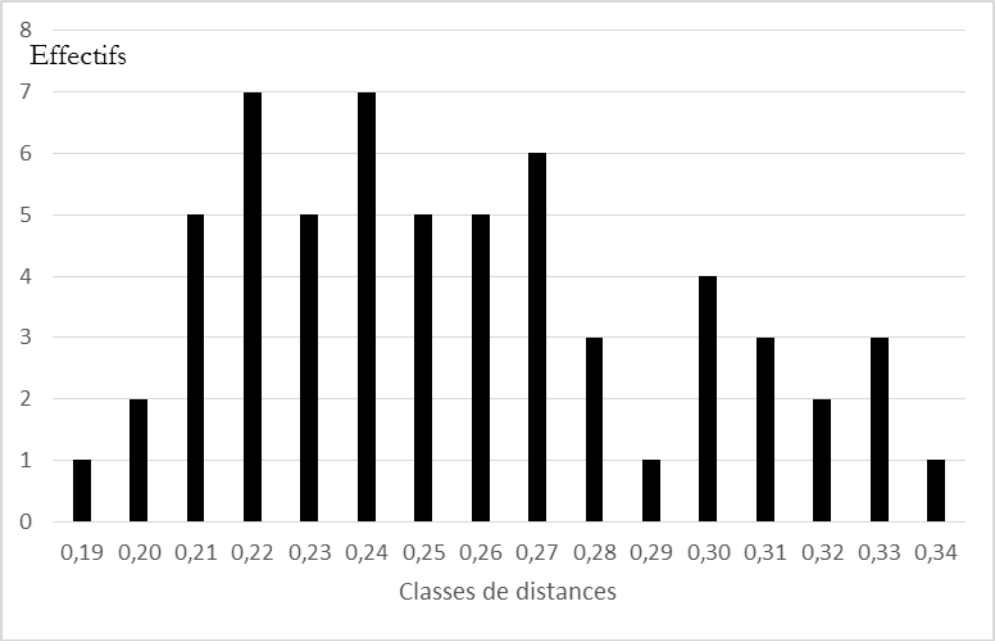
Les trois valeurs centrales, caractérisant cette série statistique, sont nettement plus élevées que chez les frères Corneille : Médiane : 0,246 ; moyenne : 0,257. L'écart-type : 0,040 confirme cette dispersion : le coefficient de variation relative : 16%, est deux fois plus élevé que chez les frères Corneille (mais aussi tous les corpus à auteur et genre uniques).

Ces valeurs sont de 30 à 100% plus fortes que chez P. Corneille avec une dispersion très importante. Logiquement, la distribution des distances est éloignée de la distribution normale (tableau 7).

La série est multimodale : deux modes principaux (0.22 et 0.24) et trois modes secondaires (0.27, 0.30 et 0.33).

Les résultats divergent grandement par rapport aux frères Corneille et par rapport à ce que laisse attendre des textes écrits par un écrivain unique dans un même genre : étalement plus grand, valeurs plus élevées, courbe multimodale. Tout cela indique une population hétérogène, voire le mélange de plusieurs populations différentes.

Tableau 7. Histogramme des distances entre les tragédies présentées par J. Racine (1664-1691)



On constate notamment que, sur les 55 distances :

- trois seulement sont inférieures ou égales à 0.20, il s'agit des couples *Mithridate-Iphigénie*, *Andromaque-Bajazet* et *Bajazet-Mithridate*, alors qu'il y en a 93% chez P. Corneille

- 27 distances sont comprises dans l'intervalle 0.21-0.25 ;

- enfin, 25 distances sont supérieures à 0.25 (en gras sur le tableau 5). Elles sont toutes concentrées sur quatre pièces. Dix-huit concernent les deux dernières pièces (*Esther* et *Athalie*), très proches entre elles mais remarquablement éloignées des autres tragédies présentées par J. Racine avant 1677. Sept concernent les deux premières pièces : *la Thébaïde* avec, par ordre de distances croissantes, *Iphigénie*, *Bajazet*, *Britannicus*, *Phèdre*, *Bérénice* - et *Alexandre* avec *Bérénice* et *Phèdre*.

Quelles explications donner à ces résultats qui divergent par rapport à ce qui est constaté chez un écrivain unique dans un seul genre ?

- Les deux dernières pièces présentées en 1689 (*Esther*) et 1691 (*Athalie*), proches l'une de l'autre, marquent une coupure nette par rapport aux œuvres présentées par J. Racine durant les années 1664-1677[1]. L'étrangeté *d'Esther* et *Athalie* par rapport aux autres pièces était déjà connue et généralement attribuée à leur caractère "sacré" et à l'étrangeté de leur commande[2].

[1] Labbé Cyril & Labbé Dominique. A Tool for Literary Studies: Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*. 21-3, 2006, p. 311-326.

[2] Composées à la demande de Mme de Maintenon, seconde épouse du roi Louis XIV, pour être représentées uniquement à Saint-Cyr l'Ecole (institution fondée par Mme de Maintenon pour l'éducation des jeunes nobles orphelines).

- Pour la période 1664-1677 : le temps et une certaine versatilité. En effet, plus l'intervalle de temps séparant deux pièces est important, plus la distance est élevée. De fait, la plus faible distance sépare *Mithridate* (1672) d'*Iphigénie* (1674) : 0,19 ou *Bajazet* (1672) - *Mithridate* (1672) : 0.20 ; la plus forte, la *Thébaïde* (1664) - *Phèdre* (1677) : 0,28.

Le décalage de chaque pièce par rapport à toutes les autres est mesuré par la dernière ligne du tableau 5 récapitulée dans le tableau 8 ci-dessous.

Tableau 8. Eloignement de chaque texte par rapport au centre de gravité de l'œuvre présentée par J. Racine

1	Mithridate (1672)	0.233
2	Iphigénie (1674)	0.238
3	Britannicus (1669)	0.242
4	Andromaque (1667)	0.243
5	Bajazet (1672)	0.244
6	Phèdre (1677)	0.252
7	Bérénice (1670)	0.254
8	Alexandre (1665)	0.257
9	Thébaïde (1664)	0.267
10	Athalie (1691)	0.289
11	Esther (1689)	0.306

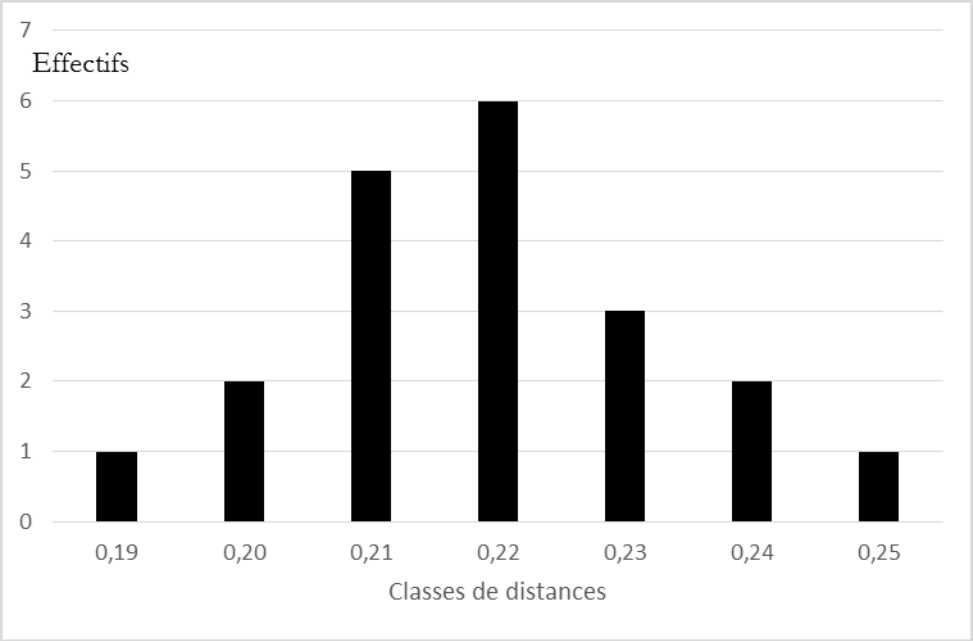
Les deux premières et les deux dernières pièces – n° 8 à 11 dans le tableau 8 - sont les principales "anomalies" de la série. Dans un cas de ce genre, il est recommandé de retirer les anomalies et de traiter sans elles le reste de la série statistique[1].

L'étude se concentre donc sur le cadre au centre du tableau 5. Cette sous-population est constituée des sept tragédies formant le "noyau central" de l'œuvre de J. Racine[2]. Toutes les distances, comprises dans ce cadre, sont inférieures ou égales à 0.25. Les valeurs centrales sont les suivantes : médiane : 0,220 ; moyenne : 0,220 ; mode : 0,220. L'égalité de ces trois valeurs centrales indique une distribution normale. L'homogénéité de la série est vérifiée par l'histogramme des distances classées de manière ascendante (tableau 9 ci-dessous).

[1] Les deux premières pièces feront l'objet d'une étude à paraître (*Les débuts de J. Racine*).

[2] Ce sont aussi les pièces qui ont connu le succès (voir Basson Jean-Charles & Labbé Dominique. *Op. Cit.)*

Tableau 9. Histogramme des distances intra (tragédies présentées par J. Racine entre 1667-1677).



La distribution des distances autour de la moyenne est normale (courbe en cloche), ce qui est la caractéristique d'œuvres écrites par un seul écrivain. Les valeurs centrales plus élevées que chez P. et T. Corneille suggèrent que cet écrivain est plus versatile à la fois dans ses thèmes et dans le temps que ne le sont les frères Corneille.

L'écart-type (0.0156), la dispersion autour de la moyenne (7%) indiquent l'homogénéité de cette population. 95% des valeurs sont comprises entre 0,190 et 0.250. 99% des valeurs sont comprises entre 0.185 et 0.255.

La suite de l'expérience portera donc sur ces 7 pièces - formant le noyau central de l'œuvre tragique présentée par J. Racine - comparées aux pièces contemporaines de P. et T. Corneille.

II. DISTANCES INTER-CORPUS

Les distances entre pièces d'écrivains différents ("inter") sont au nombre de 70 pour P. Corneille comparé à J. Racine et de 63 pour T. Corneille vs J. Racine. Ces distances "inter" se différencient-elles significativement des distances "intra" ? La variable "auteur" a-t-elle le poids que laisse attendre le modèle présenté dans le chapitre précédent ?

Jean Racine et Pierre Corneille

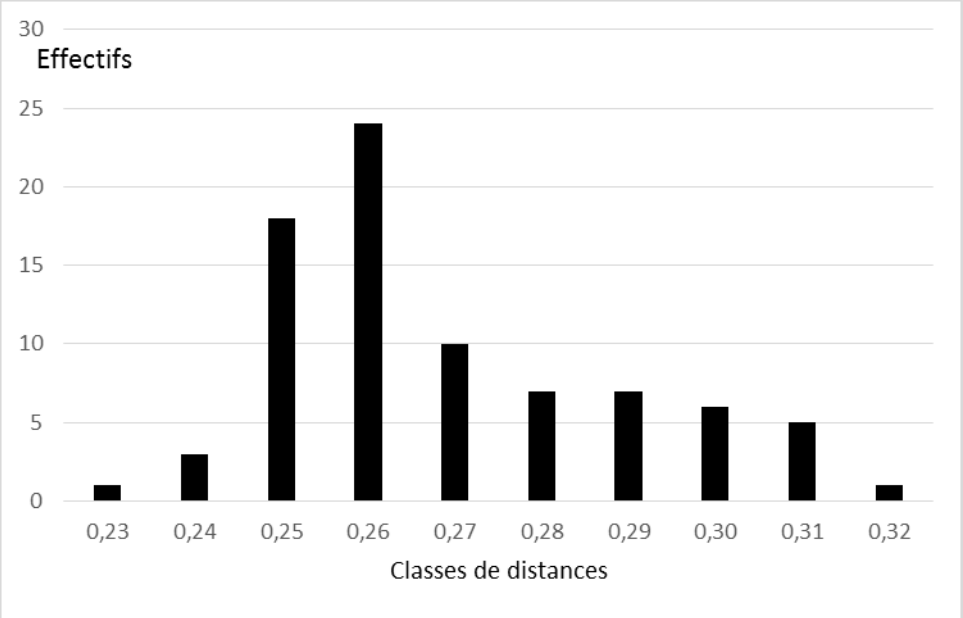
Les résultats sont récapitulés dans le tableau 10 ci-dessous. Ici toutes les cases sont différentes. Les marges du tableau (lignes et colonnes) indiquent les moyennes des distances de chaque pièce d'un écrivain par rapport à toutes les pièces de l'autre. Toutes les valeurs dans les marges du tableau sont supérieures à 0.25. La dernière case en bas à droite contient la moyenne générale : 0,264.

Tableau 10. Les distances entre les tragédies présentées par J. Racine et P. Corneille (distances "inter", classement chronologique)

Racine P. Corneille	Andromaque	Britannicus	Bérénice	Bajazet	Mithridate	Iphigénie	Phèdre	Moyenne
Œdipe	0,255	0,254	0,276	0,257	**0,239**	0,254	0,264	0,257
Toison d'or	0,246	0,262	0,272	0,250	**0,230**	0,252	0,271	0,255
Sertorius	0,253	0,246	0,258	0,257	**0,238**	0,260	0,297	0.257
Sophonisbe	0,253	0,247	0,262	0,250	**0,236**	0,264	0,296	0.258
Othon	0,260	0,250	0,274	0,257	0,248	0,275	0,303	0.267
Agésilas	0,263	0,267	0,278	0,271	0,257	0,285	0,318	0.277
Attila	0,263	0,266	0,289	0,271	0,249	0,275	0,306	0.274
Tite	0,259	0,251	0,256	0,262	0,246	0,281	0,302	0.265
Pulchérie	0,263	0,255	0,271	0,261	0,254	0,278	0,306	0.270
Suréna	0,257	0,252	0,264	0,255	**0,244**	0,275	0,298	0.264
Moyenne	0,257	0,255	0,270	0,259	0,245	0,270	0,296	**0.264**

La médiane 0,260, le mode 0.26 sont très proches de la moyenne (0,264), ce qui signale une population "normale" ou proche de la normale. La variation autour de cette moyenne (écart type : 0,018) indique une assez faible variation autour de ces valeurs centrales (7%). Les intervalles à 95% (0.229 - 0.300) et 99% (0.220 et 0.309) sont également vérifiés. La distribution de ces distances épouse à nouveau un profil en cloche mais avec une "queue" sur la droite (tableau 11).

Tableau 11. Histogramme des distances entre J. Racine et P. Corneille



La plus petite distance (0.230) sépare *Mithridate* (1672) de *la Toison d'or* (1661). La plus longue (0,318, soit 0.32) sépare *Phèdre* (1677) d'*Agésilas* (1666). D'autres facteurs que le temps semblent donc jouer ici.

La série est groupée autour du mode (0.26) avec deux "anomalies".

Premièrement, un petit nombre de valeurs sont décalées vers le haut ("queue de distribution"). Elles signalent la présence d'un facteur perturbateur provenant exclusivement de la dernière pièce de J. Racine - *Phèdre* (1677) – qui semble nettement plus éloignée de P. Corneille que les pièces antérieures présentées par J. Racine. Les moyennes en dernière ligne du tableau 9 indiquent en effet un changement après *Mithridate*. Cependant, comme on le verra dans le dernier chapitre de ce rapport, ces distances n'ont rien d'anormal pour deux écrivains différents travaillant dans le même genre sur des thèmes différents avec un décalage dans le temps non négligeable.

Deuxièmement, sur la partie gauche du graphe, cinq distances sont inférieures à 0,25. Elles concernent toutes *Mithridate* (1672) avec *Œdipe (*1659), *Sertorius* (1662), la *Toison d'or* (1661) *Sophonisbe* (1663) et *Suréna* (1674). Certes, dans une série de ce genre, on attend 5% des valeurs en dehors de la plage de variation normale. Mais ces valeurs devraient être réparties aléatoirement sur l'ensemble alors qu'elles sont concentrées sur une seule pièce (*Mithridate*).

La dernière ligne indique en effet que *Mithridate* (1672) est la plus "cornélienne" du noyau central des pièces présentées par J. Racine. La dernière colonne indique que *la Toison d'or* (1661)

est la plus "racinienne" des pièces de P. Corneille. Cette pièce est – avec *Sertorius* et *Psyché* – un très grand succès théâtral de l'époque.

Etant donné que les dates de création des pièces de P. Corneille les plus proches de *Mithridate* sont antérieures à la création de celle-ci, le sens de l'influence ne fait guère de doute. Au passage, on voit que la distance intertextuelle ne se limite pas à l'attribution d'auteur mais qu'elle détecte aussi des proximités et des influences intéressantes pour l'étude littéraire.

Ces principales conclusions sont confirmées par la comparaison entre J. Racine et T. Corneille

Jean Racine et Thomas Corneille

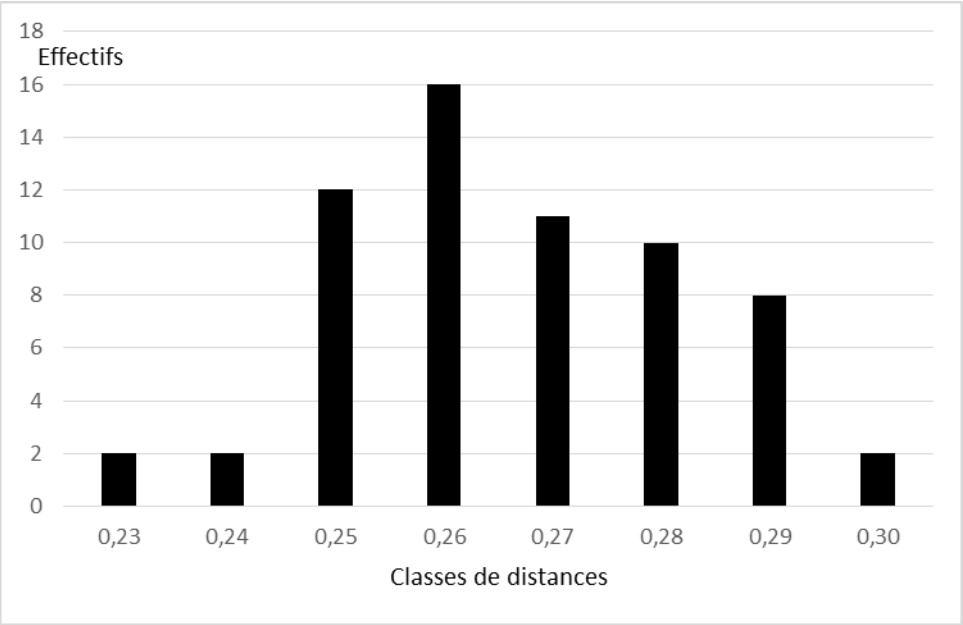
Le tableau 12 reproduit la partie de la matrice des distances inter concernant J. Racine et T. Corneille.

Tableau 12. Distances entre les tragédies contemporaines présentées par J. Racine et T. Corneille (distances "inter", classement chronologique)

Racine T. Corneille	Andromaque	Britannicus	Bérénice	Bajazet	Mithridate	Iphigénie	Phèdre	Moyenne
Stilicon	0,262	0,270	0,298	0,262	0,265	0,283	0,273	0,273
Camma	0,250	0,278	0,285	0,254	0,253	0,277	0,279	0,268
Persee	0,252	0,256	0,282	0,258	0,249	0,270	0,277	0,264
Maximian	0,264	0,266	0,295	0,261	0,266	0,282	0,285	0,274
Pyrrhus	**0,241**	0,260	0,282	0,257	0,254	0,275	0,286	0,265
Annibal	0,250	0,247	0,263	0,255	**0,233**	0,264	0,291	0,257
Ariane	**0,242**	0,272	0,259	0,249	0,258	0,292	0,292	0,266
Achille	**0,233**	0,267	0,269	0,257	0,251	0,257	0,274	0,258
Essex	0,249	0,266	0,284	0,252	0,267	0,289	0,291	0,271
Moyenne	0,250	0,265	0,280	0,256	0,255	0,277	0,283	0,266

La moyenne (0,266), la médiane (0,265) et le mode (0.26) sont très proches suggérant une distribution normale, ce que confirme l'histogramme des distances (tableau 13).

Tableau 13. Histogramme des distances "inter" (tragédies contemporaines présentées par J. Racine et T. Corneille)



La série présente un profil proche de la distribution normale mais avec une asymétrie et avec une "queue" à droite comparable à celle observée pour J. Racine/P. Corneille (Tableau 11). Le facteur perturbateur semble à nouveau ne pas être simplement la chronologie.

L'écart-type (0.019) indique une variation assez faible autour de la moyenne. Les intervalles de variation normale correspondent aux valeurs observées.

95% des distances sont comprises dans l'intervalle 0.237-0.296

Toutes les distances dans l'intervalle à 99% (0,230-0.304)

Au seuil de 95%, on note :

- distances remarquablement faibles : *Andromaque* (1667) – *Mort d'Achille* (1673) et *Mithridate* (1672)- *Mort d'Annibal* (1669) ;

- distance remarquablement forte : *Bérénice* (1670) – *Stilicon* (1660).

De nouveau, ces distances faibles ne sont pas exceptionnelles pour des écrivains contemporains travaillant dans le même genre, pour les mêmes acteurs, le même public et sur des thèmes proches. Ici la date de création de *la Mort d'Achille* (T. Corneille) est postérieure à celle d'*Andromaque* (J. Racine), mais celle de la *Mort d'Annibal* (T. Corneille) est antérieure à celle de *Mithidate* (J. Racine), on peut conclure à une influence mutuelle entre les écrivains auteurs de ces pièces.

Conclusions du chapitre

Les populations sur lesquelles travaille la lexicométrie ne sont pas tout à fait comparables au cas d'école de la statistique standard - comme les lancers de dés, les tirages dans une urne, etc. - qui, à partir d'un certain effectif, donnent des distributions en forme de cloches parfaites. Ici, les profils empiriques s'en écartent toujours un peu à cause de facteurs perturbateurs. Cependant, les distributions des distances intertextuelles observées chez les trois auteurs approchent ces profils théoriques et peuvent être considérées comme statistiquement normales. Autrement dit, dans les trois corpus qui viennent d'être présentés, les distances qui séparent chacun des textes de tous les autres – ou distances "intra" - forment des populations "gaussiennes" caractérisées par une moyenne et une certaine variabilité autour de cette moyenne, ce qui permet de définir des plages standards. Ces trois corpus sont séparés des deux autres par des distances "inter" qui sont également distribuées de manière normale et peuvent donc également être dotées de plages standards.

Dans ces conditions, il est possible d'utiliser les procédures habituelles en mathématiques appliquées pour répondre à deux questions principales :

- les mesures permettent-elles d'évaluer la contribution du facteur "auteur" à la distance entre textes ?

- dans le cas présent peut-on conclure qu'il y a trois écrivains distincts ? Et si oui avec quel degré de certitude ?

CHAPITRE III.
IDENTIFICATION DE L'ÉCRIVAIN

Dans l'expérience en cours, des quatre facteurs qui déterminent la distance, trois sont neutralisés ou minimisés. Le genre est neutralisé : tous ces textes sont des tragédies en alexandrins en cinq actes de même format (un spectacle entier). Leur création se situe durant la même période d'une quinzaine d'années. Seul le poids du thème reste relativement indéterminé mais il doit être relativement constant puisque ses variations ne sont pas suffisantes pour perturber la distribution normale des distances. La situation est donc optimale pour mesurer l'influence du facteur "auteur" sur la distance entre textes.

Ce poids suffit-il pour identifier celui qui a composé un texte ? En particulier, les propriétés mises en lumière dans le précédent chapitre permettent-elle d'utiliser les tests statistiques standards et de connaître le degré de fiabilité de cette attribution ?

I. TESTS STATISTIQUES STANDARDS

Puisque toutes les distances inter et intra semblent distribuées de manière normale, il est logique de recourir à un des tests paramétriques standards[1] utilisés pour comparer les moyennes et les dispersions observées dans des populations différentes afin de savoir si les écarts constatés permettent de conclure à des différences significatives entre ces populations : ici des auteurs différents puisque ce facteur est isolé.

Deux approches sont possibles. Les oeuvres sont comparées deux à deux ou la comparaison est généralisée.

Comparaison d'une œuvre à une autre

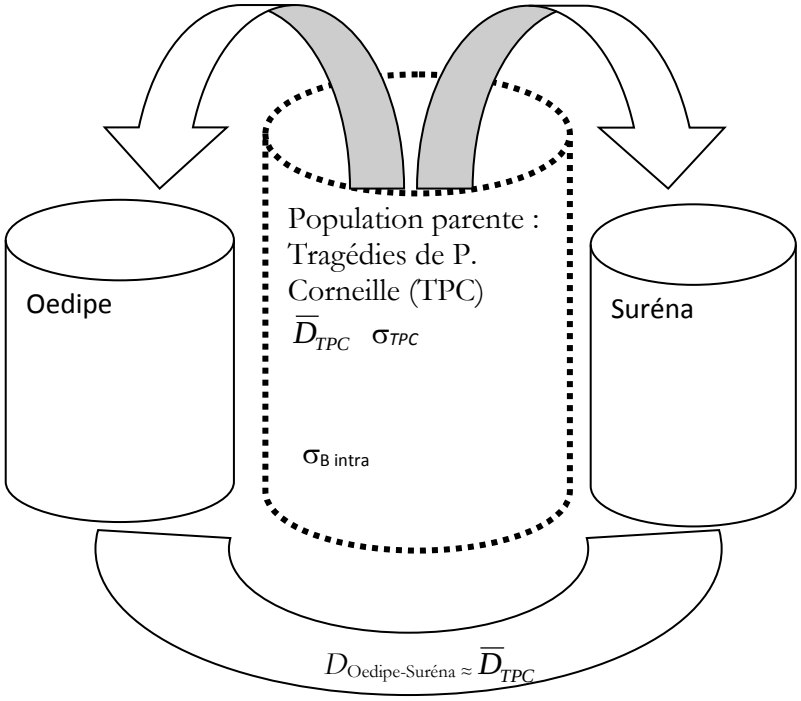
Le phénomène sous-jacent est le suivant : dans un laps de temps limité (ici une quinzaine d'années), les habitudes d'écriture peuvent être considérées comme "stables" et la diversité des thèmes imprime des différences limitées entre les textes (la réalité de ce phénomène a été

[1] Voir CISIA-CERESTA. *Aide-mémoire statistique*. Paris : CISIA-CERESTA, 1995 et Harris John W. & Stocker Horst. *Handbook of Mathematical and Computational Science*. New-York – Heidelberg : 1998.

préalablement vérifiée grâce à la proximité des valeurs centrales et à la distribution normales des observations autour de leur moyenne).

Considérons le premier corpus (P. Corneille). Chacun des textes peut être représenté comme un ensemble de mots appartenant à une population parente homogène : "Tragédies de la dernière partie de la vie créatrice de P. Corneille" (dans le schéma ci-dessous : tragédies de P. Corneille ou TPC).

Tableau 1. Schéma de principe "tragédies de la dernière partie de la vie créatrice de P. Corneille"

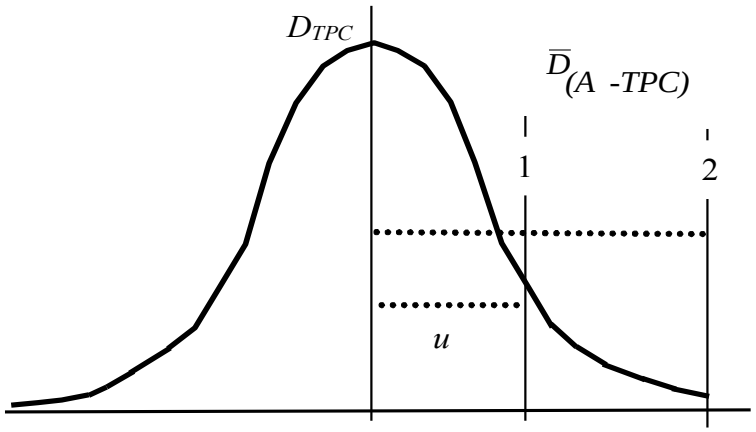


La population parente est composée des 10 dernières tragédies et comporte N_{TPC} mots différents. Son vocabulaire est composé de V_{TPC} vocables différents, chacun d'entre eux ayant dans N_{TPC} un effectif (F). Dans TPC, la distribution du caractère D, distance entre les pièces appartenant à TPC, suit une loi *normale* de paramètres ($\overline{D}_{TPC}$, σ_{TPC}). L'écart type donne la variabilité *normale* du caractère entre les tragédies. Plus ces valeurs sont faibles plus la population est homogène (faibles distances et faible variabilité de celles-ci) et, dans le cas présent, plus l'on sera sûr que l'écrivain est le même puisque ce facteur a été isolé en minimisant ou en annulant les autres facteurs agissant sur la distance.

Considérons une tragédie contemporaine A, n'appartenant pas à TPC et calculons $\overline{D}_{(A\text{-}TPC)}$, moyenne des distances entre A et les 10 tragédies formant TPC.

Deux situations sont examinées selon le schéma de principe suivant.

Tableau 2 Schéma de principe du test statistique standard de comparaison des moyennes



Premièrement $\overline{D}_{(A\ -TPC)}$ se situe dans la plage de variation normale autour de $\overline{D}_{TPC}$. On retient l'hypothèse d'un écrivain unique (tragédie ignorée de P. Corneille). Cette hypothèse est notée H_o (hypothèse nulle). Elle signifie ici que, au seuil choisi, la différence entre $\overline{D}_{(A\ -TPC)}$ et $\overline{D}_{TPC}$ est due au hasard (variabilité normale du caractère au sein de la population de référence). Naturellement, dire qu'une hypothèse est acceptée ne signifie pas qu'elle est "vraie" mais seulement que les observations disponibles ne sont pas incompatibles avec elle et que l'on n'a pas de raison de lui préférer l'hypothèse contraire[1]. Si l'auteur de A est inconnu ou douteux, on accepte l'hypothèse P. Corneille (nous discutons plus loin la question de sa "véracité").

Deuxièmement, $\overline{D}_{(A\ -TPC)}$ se situe en dehors de l'intervalle de variation normale. H_0 est rejetée et l'hypothèse inverse, notée H_1 est acceptée : le texte n'appartient pas à TPC. Si l'auteur de A est inconnu, peut-on affirmer qu'il ne s'agit pas de P. Corneille ? Il faut d'abord être certain qu'il s'agit bien d'une tragédie contemporaine, sur un thème pas trop éloigné de ceux traités par P. Corneille, dépouillée selon les normes indiquées au premier chapitre. Si ces conditions sont réunies, on rejette H_0 (pièce de P. Corneille) et l'on accepte H_1 (deux écrivains différents) avec un risque d'erreur (α) d'autant plus faible que la différence entre les deux moyennes est grande.

Pour choisir entre ces deux hypothèses, il faut calculer $u = \overline{D}_{(A\ -TPC)} - \overline{D}_{TPC}$ et rapporter cette valeur à l'écart type σ_{TPC} (écart réduit). Cet écart réduit répond à la question : de combien d'écarts types les deux valeurs sont-elles séparées ? Sous réserve d'une distribution normale des distances dans la population de référence, ce calcul ramène le raisonnement dans le schéma de principe présenté au précédent chapitre (tableau 3).

[1] Desrosières Alain. La partie pour le tout : comment généraliser ? *Cinq contributions à l'histoire de la statistique*. Paris : Economica, 1988.

Un cas limite

Nous allons illustrer ce raisonnement à l'aide d'un exemple limite. En effet, pour mettre à l'épreuve une théorie, il faut trouver des cas qui peuvent la mettre en défaut. Aussi a-t-on recherché dans le corpus du théâtre du XVIIe une tragédie particulièrement proche de celles de P. Corneille mais présentée par un autre écrivain (dont le corpus a été contrôlé selon les procédures du chapitre précédent). Il s'agit de *Stilicon* de son frère T. Corneille (1660). Le tableau 4 du chapitre II donne les distances entre cette tragédie et les autres de T. Corneille (moyenne intra : 0.189). Les distances entre cette pièce et les dix tragédies contemporaines de P. Corneille sont données dans le tableau 2.

Tableau 2. Distances de *Stilicon* (1660) de T. Corneille aux dix dernières tragédies de P. Corneille (classement chronologique).

P. Corneille	Stilicon (1660)
Œdipe (1659)	0,220
Toison d'Or (1661)	0,233
Sertorius (1662)	0,245
Sophonisbe (1663)	0,245
Othon (1664)	0,236
Agésilas (1666)	0,252
Attila (1667)	0,244
Tite et Bérénice (1670)	0,243
Pulchérie (1672)	0,238
Suréna (1674	0,231
Moyenne	0,239

La probabilité pour que la différence entre $\overline{D}_{(A\text{-}TPC)}(0.239)$ et $\overline{D}_{TPC}(0.181)$ ne soit pas anormale est donnée par l'écart réduit :

$$u = \frac{\overline{D}_{(A\text{-}TPC)} - \overline{D}_{TPC}}{\sigma_{TPC}} = \frac{0.239 - 0.181}{0.015} = 3{,}77 \qquad (1)$$

La table de l'écart réduit (loi normale centrée réduite) donne la probabilité α pour que l'écart-réduit égale ou dépasse, en valeur absolue, une valeur donnée. Ci-dessous les principales valeurs seuils (tableau 3).

Tableau 3. Extraits de la table de l'écart réduit[1]

α	0,05	0.01	0.001	0.000 1	0.000 01	0.000 001	0.000 000 1	0.000 000 01	0.000 000 001
u	1.96	2.58	3.29	3.89	4.42	4.89	5.33	5.73	6.11

Un écart réduit de 3,77 indique qu'il y a moins d'une chance sur mille de se tromper en affirmant que *Stilicon* n'est pas une pièce de P. Corneille ou encore qu'elle est l'œuvre d'un autre écrivain.

Ce calcul appelle trois remarques

Premièrement, l'adoption de H_1 (deux écrivains différents) ne permet pas d'écarter une influence possible de Pierre sur cette pièce, voire une collaboration occasionnelle. Cette possibilité est suggérée par la proximité remarquable avec *Œdipe*, contemporaine de la composition de *Stilicon* (seule distance comprise dans l'intervalle à 99% pour les deux corpus).

Rappelons que les frères Corneille ont toujours vécu ensemble, qu'ils ont fait les mêmes études, qu'ils ont épousé deux sœurs, que les deux ménages ont fait bourse commune jusqu'au décès de Pierre (1684) et vécu sous le même toit. Il n'y a pas de conditions plus favorables à des influences mutuelles, voire à des collaborations. C'est pourquoi, ils ont été choisis comme "cas limite" (comme les sœurs Brontë pour la littérature anglaise). Ces cas définissent la proximité maximale entre écrivains différents dans le cadre d'influence de l'un envers l'autre ou de collaboration ponctuelle entre les deux[2]. En-dessous de cette limite, l'hypothèse de la collaboration ou de l'influence pourra donc être rejetée au profit de celle d'un auteur unique.

Deuxièmement, il y a deux risques d'erreur. Le risque de rejeter H_0 alors qu'elle est vraie, ou "risque de première espèce" noté α que nous venons de calculer. Mais il y a aussi le risque d'accepter H_1 alors qu'elle est fausse, ou risque de "seconde espèce" (noté β). Si l'on considère que les deux hypothèses sont alternatives – un ou deux écrivains et rien d'autre - le risque de rejeter H_0 alors qu'elle est vraie est le même qu'accepter H_1 alors qu'elle est fausse, soit ici moins de 1‰. En revanche, au moins une autre hypothèse est concevable étant donné les relations entre les deux écrivains (une ou plusieurs collaborations) : le fait de rejeter H_0 n'implique pas que l'on puisse conclure à deux écrivains travaillant *indépendamment l'un de l'autre* (il faut pouvoir rejeter l'hypothèse d'une collaboration). Nous présentons plus bas le calcul de cette erreur de second

[1] D'après Fisher & Yates. *Statistical Tables for Biological, Agricultural and Medical Research* (1949).

[2] Nous reviendrons dans un prochain ouvrage sur la vie et les œuvres des deux frères Corneille.

type, dans le cas des hypothèses composites, calcul qui n'est pas possible pour un seul texte et un corpus de référence de moins de 30 textes[1].

Troisièmement, on peut souhaiter tester non pas les pièces une à une mais des corpus entiers. La question devient : toutes ces pièces peuvent-elles sortir d'un même moule ?

II. UN MÊME MOULE ?

L'hypothèse à tester est la suivante. Sans nier l'existence d'un style personnel, de nombreux spécialistes de la littérature affirment que les différents écrivains ne peuvent être reconnus car, en choisissant un genre donné, ils se sont coulés dans le même moule. Ces spécialistes pensent que ce serait spécialement le cas pour le théâtre du XVIIe. On peut résumer ainsi cette thèse : la loi du genre effacerait les différences individuelles et l'attribution d'auteur serait impossible même par ordinateur[2].

Une population unique : la tragédie classique ?

L'hypothèse du moule unique est notée H_o (hypothèse nulle). Elle signifie ici que, au seuil choisi, on ne peut écarter l'idée selon laquelle les distances "inter" entre deux corpus A et B ne s'écartent pas significativement des distances "intra" ($\overline{D}_{intra} \approx \overline{D}_{inter}$) et que les différences sont dues à la variabilité normale du caractère au sein de la population parente dont seraient issus tous les textes sous revue. S'il y a effectivement plusieurs écrivains, ils n'auraient pas une manière significativement différente d'employer le vocabulaire du français de leur époque, ou encore leurs œuvres sortiraient toutes d'un même "moule" (celui de la tragédie classique).

A l'inverse, si l'on dépasse le seuil α, on devra rejeter H_0 et accepter l'hypothèse inverse, notée H_1 : les auteurs peuvent être identifiés. Naturellement, rejeter H_0 c'est aussi considérer que l'hypothèse d'un moule unique - plus fort que les différences individuelles - n'est pas vérifiée.

---

[1] Une solution est donnée dans le dernier chapitre.

[2] Cette thèse a été formulée par : Etienne Brunet et Charles Muller. La statistique résout-elle les problèmes d'attribution ? *Strumenti critici,* 1988-III, n°3, p. 367-387. On trouve également cette idée, formulées de diverses manières, chez de nombreux littéraires. Par exemple, Georges Couton. *Richelieu et le théâtre*. Paris : PUF, 1986, p. 21-25. G. Couton considère que la question de savoir qui a collaboré aux pièces financées par le Cardinal, est une question sans importance car, dit-il, ces écrivains n'ont été que des "ouvriers" au service des desseins de Richelieu. C'est l'attitude de la plupart des littéraires face à la question des "plumes de l'ombre". Nous avons répondu en utilisant les mêmes textes que Brunet et Muller : Labbé Dominique. Identification de l'auteur d'un texte (Hugo, Lamartine, Musset et Vigny). Conférence invitée au séminaire *L'œuvre et son auteur : problèmes d'attribution*. Lille : Université de Lille-Nord de la France, 21 mai 2014.

H_o ($\overline{D}_{intra} \approx \overline{D}_{inter}$) contre H_1 ($\overline{D}_{intra} \neq \overline{D}_{inter}$)

Soit : $\overline{D}theo$ la moyenne des distances intra des corpus A et B et $\sigma theo$ la dispersion standard de ces distances autour de $\overline{D}theo$. $\overline{D}theo$ et $\sigma theo$ sont les moyennes des paramètres observés dans A et B, moyennes simples quand il y a le même nombre de pièces dans A et B ($N_{Aintra} = N_{B\,intra}$) et moyennes pondérées dans le cas inverse.

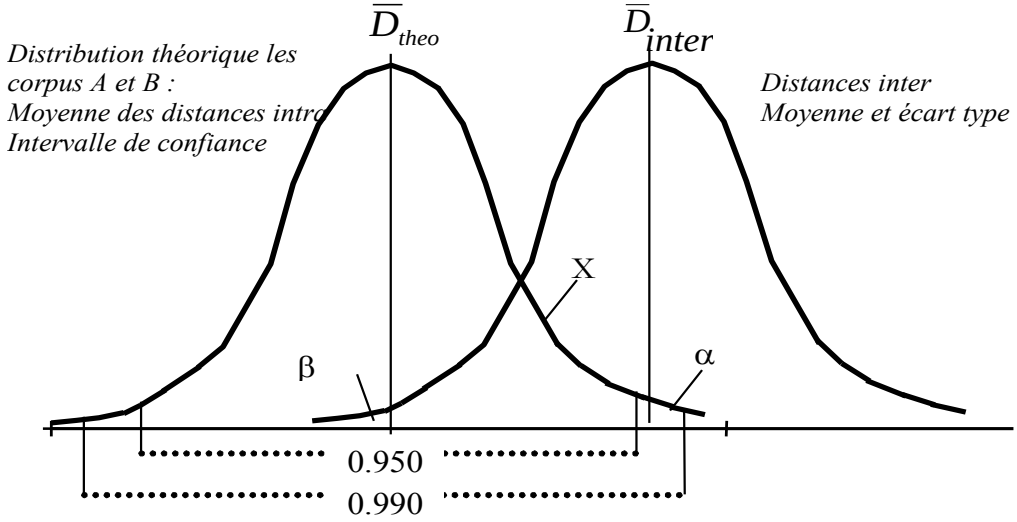
Avec $N_{A\,intra} \neq N_{B\,intra}$:

$$\overline{D}theo = \frac{\overline{D}_{A intra}\ N_{A\,intra} + \overline{D}_{B intra}\ N_{B\,intra}}{N_{A\,intra} + N_{B\,intra}} \quad (2)$$

$$\sigma theo = \frac{\sigma_{A intra}\ N_{A\,intra} + \sigma_{B\,intra} N_{B\,intra}}{N_{A\,intra} + N_{B\,intra}} \quad (3)$$

Le raisonnement est résumé dans le schéma de principe ci-dessous.

Tableau 4. Schéma de principe d'un test utilisant l'intervalle de fluctuation normale des distances intra, la moyenne et l'écart type des distances inter pour accepter ou refuser les hypothèses H_0 et H_1 et estimer les risques d'erreur de première et deuxième espèces.



Le risque de rejeter H_0 alors qu'elle est vraie, ou "risque de première espèce", noté α, est calculé grâce à la formule de l'écart réduit ci-dessus (formule 1). Dans le cas d'hypothèses composites - par exemple, il peut y avoir deux écrivains ou une collaboration ponctuelle entre les deux, voire avec un tiers, etc. – le fait de rejeter H_0 n'implique pas nécessairement que H_1 (deux écrivains travaillant indépendamment l'un de l'autre) soit vraie et, avant de l'accepter, il faut estimer l'erreur de seconde espèce (β) (accepter H_1 alors qu'elle est au moins partiellement fausse). Formuler des hypothèses composites revient à admettre que plusieurs facteurs – en tout ou partie absents des distances intra donc de la mesure de α - interviennent dans la détermination

des distances inter. L'écart type des distances inter enregistre l'impact de ces facteurs supplémentaires. L'erreur de seconde espèce (β) – accepter H_1 alors qu'elle est fausse - sera donc déterminée à partir des intervalles de variation normale des distances inter (schéma et formule ci-dessus). Pour mesurer (β), il faut pouvoir associer un intervalle aux distances inter, donc disposer d'au moins une trentaine de ces distances.

La probabilité P(X = $\overline{D}_{inter}$) – ou probabilité que la moyenne des distances inter soit égale à un point X de la courbe théorique déduite des distances intra - est donnée par la surface sous la courbe de gauche qu'il faut parcourir pour atteindre cette valeur. Elle est maximale lorsque $\overline{D}_{inter} = \overline{D}_{theo}$. Elle décline de manière asymptotique. Aux bornes de l'intervalle à 95% elle est égale à 0,025 et à la borne 99%, elle est égale à 0.005. L'erreur de première espèce s'en déduit directement.

A condition que N_{theo} et N_{inter} soient au moins égaux à 30, la probabilité pour que la différence entre $\overline{D}_{theo}$ et $\overline{D}_{inter}$ ne soit pas anormale (au seuil choisi) est donnée par l'écart réduit. Dans le cas d'hypothèses composites, l'influence des différents facteurs supplémentaires pouvant expliquer les distances inter est mesurée dans la fluctuation standard de ces distances et l'erreur de seconde espèce (β) est la surface comprise sous la courbe des distances inter située à gauche de la moyenne des distances intra (schéma ci-dessus) :

$$u' = \frac{\overline{D_{theo}} - \overline{D_{inter}}}{\sigma_{inter}} \quad (4)$$

Une remarque : la procédure consiste à examiner les différences entre deux populations recensées exhaustivement. Il n'est donc pas nécessaire de prendre en compte les fluctuations d'échantillonnage comme on devrait le faire si l'on considérait les deux corpus comme des échantillons issus d'une même population parente inconnue. Le calcul serait le même, mais il faudrait associer aux paramètres théoriques l'écart type estimant ces fluctuations d'échantillonnage. Cette complication est ici inutile et ne changerait pas les conclusions étant donné le caractère massif des écarts entre distances intra et inter.

Trois écrivains et trois œuvres distinctes

Les formules (1) à (4) sont d'abord appliquées à la comparaison entre P. Corneille et J. Racine ($\overline{D}_{theo}$ = 0,1936 ; $\sigma theo$ = 0,0154) :

$$u = \frac{0{,}2643 - 0{,}1936}{0{,}0154} = 4{,}57$$

En se reportant au tableau 3 ci-dessus (extraits de la table de l'écart-réduit), l'hypothèse H_0 (un seul écrivain) est rejetée avec un risque d'erreur de première espèce (α) inférieur à un pour dix mille.

Si l'on admet que H_1 est composite :

$$u' = \frac{0{,}2643 - 0{,}1936}{0{,}0185} = 3{,}82$$

Le risque de seconde espèce β (choisir H_1 alors qu'elle serait au moins partiellement fausse) est inférieur à 1 pour mille.

Pour la confrontation entre les pièces de T. Corneille et J. Racine, les résultats sont : u = 3,79 et u' = 4,08. On rejette H_0 et l'on retient H_1 avec des risques d'erreur de première et de seconde espèces inférieurs à 1 pour mille.

Pour la confrontation entre les pièces de T. Corneille et P. Corneille, u = 3,26 et u' = 3,15. On rejette H_0 et l'on retient H_1 avec des risques d'erreur de première et de seconde espèces inférieurs à 1%. Ici, u et u' sont plus faibles que dans les deux autres comparaisons. Bien que la vie et l'œuvre de ces deux écrivains soient très proches et que certaines pièces comme *Stilicon* portent la marque d'une influence de P. Corneille, les deux écrivains sont donc bien distincts et le calcul les départage sans problème.

Le poids de l'écrivain est donc prépondérant et l'hypothèse du moule unique peut être rejetée avec un risque d'erreur négligeable.

III. LE POIDS DE L'ÉCRIVAIN

Comment mesurer ce poids du facteur "auteur" dans la distance entre tragédies ? Si cette mesure est possible, les résultats sont-ils suffisants pour identifier le véritable auteur d'un texte ?

Calcul du poids de l'écrivain

Selon le modèle présenté à l'issue du premier chapitre (distance fonction de quatre facteurs distincts), le corpus utilisé annule le facteur "genre" et minimise le temps et les thèmes. On peut donc estimer le poids du facteur auteur de la manière suivante. Soit, pour deux corpus :

- les *n* distances inter (D_{inter}),

- les *m* distances intra (D_{intra})

$$\text{Poids du facteur auteur} = \frac{\frac{1}{n}\sum_{1}^{n} D_{inter}}{\frac{1}{m}\sum_{1}^{m} D_{intra}}$$

Tableau 5. Poids du facteur auteur pour trois corpus (distances inter comparées aux distances intra)

	Facteur auteur
P. Corneille – J. Racine	1,329
T. Corneille – J. Racine	1,298
P. Corneille – T. Corneille	1,249
Moyenne	1,292

Entre ces trois écrivains (P. Corneille, T. Corneille et J. Racine), les distances inter (distances entre textes d'auteurs différents) sont donc en moyenne de 30% supérieures aux distances intra (textes d'un seul écrivain) et ce ratio varie entre +25% dans le cas des frères Corneille et 33% pour la comparaison entre les corpus P. Corneille et J. Racine.

De nombreuses expériences – dont certaines sont évoquées en annexe 1 – ont montré que, suivant les genres, le poids du facteur auteur varie entre 25 et 40% de la distance totale avec une moyenne de 33%. Autrement dit, deux textes contemporains écrits dans un même genre par deux écrivains différents sont en moyenne d'un tiers plus distants que s'ils ont été composés tous les deux par un seul écrivain. Les frères Corneille donnent la proximité maximale : +25%.

Il ne s'agit que d'une estimation car temps et thèmes ne sont jamais totalement neutralisés. Par exemple, dans le cas présent, certaines pièces sont séparées par une quinzaine d'années et chacun des trois écrivains présente une "versatilité" thématique différente et une évolution plus ou moins nette au cours de la période, du plus stable (P. Corneille) au plus versatile (J. Racine).

Cette réserve admise, le poids de ce facteur auteur est suffisant pour permettre de reconnaître les tragédies contemporaines des frères Corneille et de J. Racine.

Examen direct des intervalles

Les calculs présentés dans la section précédente ne sont pas indispensables. Il suffit de consulter les différentes plages de fluctuation normale (tableau 6) pour s'assurer que, dans tous les cas, l'hypothèse H_1 doit être acceptée avec un risque d'erreur de première et de deuxième espèces inférieur à 1% (seuil choisi *a priori*).

Les données sont récapitulées dans le tableau 6. N_{intra} et N_{inter} sont les nombres de distances sur lesquelles ont porté les calculs.

Tableau 6. Paramètres des trois ensembles de distances intra et des deux ensembles de distances inter.

Distances intra	Moyenne ($\overline{D}_{intra}$)	Ecart type (σ_{intra})	N_{intra}
P. Corneille :	0,1812	0,0154	45
T. Corneille :	0,1919	0,0177	36
J. Racine :	0,2204	0,0155	21

Distances inter	Moyenne ($\overline{D}_{inter}$)	Ecart type (σ_{inter})	N_{inter}
J. Racine/P. Corneille	0,2644	0,0184	70
J. Racine/T. Corneille	0,2664	0,0157	63

Plages de fluctuations des distances intra et inter.

	Moyennes	Intervalles : $\alpha = 5\%$	$\alpha = 1\%$
Distances intra :			
Pierre Corneille	0.1812	0,1510 - 0,2113	0,1418 - 0,2205
Thomas Corneille	0.1919	0,1573 - 0,2265	0,1467 - 0,2371
Jean Racine	0.2204	0,1902 - 0,2502	0,1809 - 0,2600
Distances inter :	Moyennes		
J.Racine / P. Corneille	0.2644	0,2291 - 0,2995	0,2199 – 0.3088
J. Racine / T. Corneille	0,2664	0.2366 - 0.2962	0.2271 – 0.3056

* Rappel : seules les deux premières décimales sont significatives, les suivantes indiquent l'arrondi.

Toutes les moyennes inter ($\overline{D}_{inter}$) sont situées en dehors des intervalles de variation normale à 95%, voire à 99% des distances intra pour les trois corpus. L'examen direct de ces données apprend donc que l'on a moins de 5%, voire moins de 1% de chances de se tromper en affirmant qu'il y a trois populations distinctes, donc trois écrivains différents (sans qu'il soit nécessaire de calculer un écart réduit).

L'examen graphique complète le raisonnement.

Examen graphique

Les tableaux 7 et 8 ci-dessous montrent que les trois séries sont distinctes. Le tableau 7 est la superposition des tableaux du chapitre précédent : 2 (intra-P. Corneille) et 11 (inter P. Corneille-J. Racine). Le 8 superpose les tableaux 9 (intra-J. Racine) et 11 (inter P. Corneille-J. Racine).

Tableau 7. Histogramme des distances intra-P. Corneille et inter P. Corneille/J. Racine

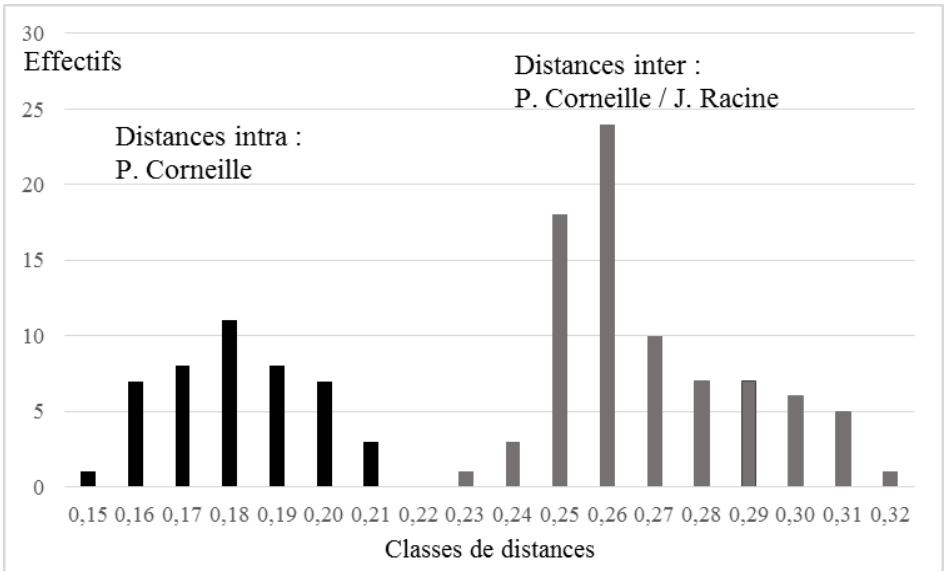
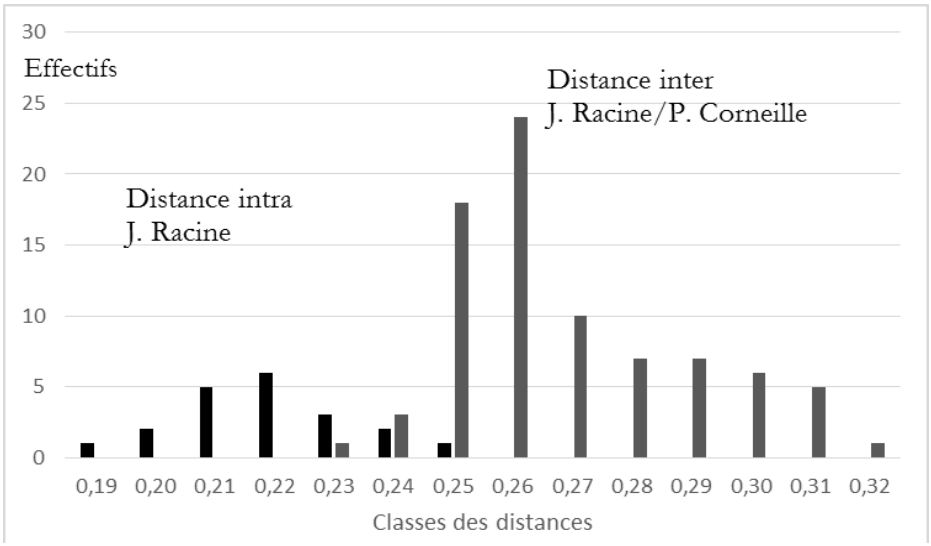


Tableau 8. Histogramme des distances intra-J. Racine et inter P. Corneille/J. Racine

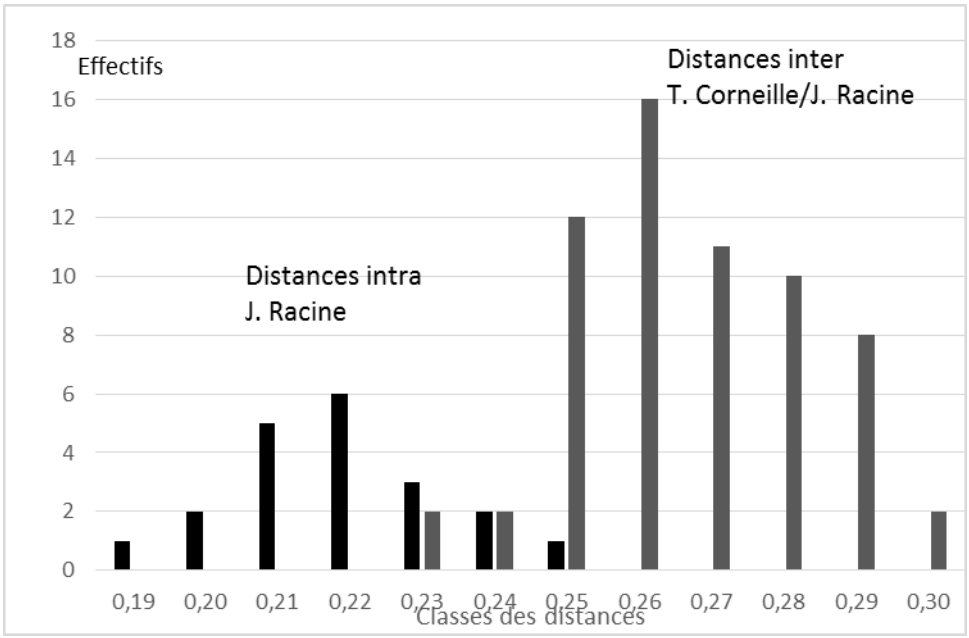


Dans le tableau 7, la classe 0.22 est vide : aucun recouvrement entre la série intra P. Corneille et la série inter P. Corneille-J. Racine. Dans le second graphique (intra J. Racine et inter

P. Corneille-J. Racine), le recouvrement des deux séries concerne moins de 9% du total des distances. Les deux courbes peuvent être considérées comme disjointes. L'hypothèse nulle est donc rejetée et l'hypothèse H_1 retenue, avec des risques négligeables d'erreurs de première et de seconde espèces.

Le même raisonnement s'applique à la comparaison entre J. Racine et T. Corneille avec les mêmes résultats. La moyenne des distances inter (0,266) est supérieure à la borne haute de l'intervalle de variation normale à 99% des distances intra T. Corneille (0.236) comme pour celles des distances intra J. Racine (0,260). En utilisant T. Corneille comme repère, on aboutit à un graphique non reproduit car semblable au tableau 21. Le recouvrement est plus sensible en utilisant J. Racine comme point de comparaison (tableau 9), sans remettre en cause la conclusion.

Tableau 9. Histogramme des distances intra-J. Racine et inter T. Corneille et J. Racine



Conclusions du chapitre

Dans un corpus homogène (genre unique, thèmes proches, écart chronologique faible), les distances entre les textes se distribuent de manière gaussienne autour de leur moyenne. Le coefficient de variation relative (des distances autour de cette moyenne) est généralement inférieur à 10%[1]. Dans ce cas, 95% des distances sont comprises dans un intervalle inférieur à ± 20% autour de la moyenne. Or nous venons de voir que deux textes – contemporains, dans un même genre et sur des thèmes proches - écrits par deux écrivains différents – et sans interférences entre eux – sont séparés par une distance en moyenne supérieure de 33% à ce qu'elle serait en cas d'écrivain unique et qu'on rencontre moins de 1% des écarts inférieurs à 25% de $\overline{D}_{intra}$. Par conséquent :

Deux textes écrits par deux écrivains contemporains, dans un même genre, sont *significativement* plus distants entre eux que ne le sont deux textes comparables écrits par un seul écrivain.

On objectera sans doute qu'on sait déjà que les pièces des frères Corneille n'ont pas été écrites par J. Racine. C'est oublier que :

- lorsque nous avons publié nos résultats sur P. Corneille et Molière, beaucoup de critiques ont affirmé que le moule du genre est si fort qu'il efface l'empreinte de l'écrivain, de telle sorte que l'ordinateur ne peut le reconnaître. On a vu qu'il n'en est rien et que, malgré les contraintes incontestables du genre, les distances intertextuelles discriminent bien les écrivains… quand ils sont différents ;

- l'automate travaille en aveugle et n'utilise aucune information sur les écrivains ni sur les titres des textes qu'il traite. Il est donc important de vérifier qu'il est bien capable de discriminer des écrivains différents – sans rien connaître d'eux - et de le faire sans erreur et pour tous leurs textes et ceci même lorsque ces deux auteurs sont remarquablement proches.

On a également objecté que l'acceptation de H_1 (deux écrivains différents) n'a pas le même statut que celle de H_0 (un seul écrivain). Dans le premier cas, cette décision se voit associer un certain risque d'erreur alors que le second (non rejet de H_0 : un seul écrivain) est une décision par

[1] Par exemple, nous avons trouvé plus haut : 8% pour le corpus P. Corneille, 9% pour le corpus T. Corneille et 7% pour le corpus J. Racine. Plus bas, J-G. Campistron : 8%. Ce sont les moyennes observées sur la plupart des corpus littéraires.

défaut qui ne signifie pas pour autant que H_0 est vraie. Cette présentation des choses est exacte lorsqu'on travaille avec des échantillons. Dans ce cas, il faut ajouter, à la variabilité naturelle du phénomène étudié, l'incertitude liée aux fluctuations d'échantillonnage, incertitude qui varie en raison inverse de $\sqrt{n}$ (avec n : nombre d'observations). Une augmentation du nombre d'observations réduit donc la plage de variation et cela peut suffire pour faire passer une valeur donnée de la zone d'acceptation à celle du rejet de l'hypothèse nulle. D'où, dans les études par échantillonnages, l'acceptation prudente de l'hypothèse nulle *en l'état des observations disponibles*.

Ce schéma ne s'applique pas à notre recherche car nous ne travaillons pas sur des échantillons mais sur des populations recensées exhaustivement. Toutes les tragédies, parues sous le nom d'un auteur pour une période donnée, ont été dépouillées en suivant exactement la même procédure. Tous les mots entrent dans le calcul pour leurs effectifs respectifs. Le calcul enregistre donc la variabilité naturelle des distances au sein de ces collections de textes sans ajouter aucune incertitude qui devrait être prise en compte dans le test et dans la décision. Si pour un texte n'appartenant pas au corpus de référence, la moyenne de ses distances à tous les textes composant un corpus est comprise dans l'intervalle de variabilité normale, autour de la moyenne des distances intra de ce corpus, alors ce texte appartient au corpus en question et il n'y a pas d'incertitude sur cette appartenance car aucune étude ultérieure ne peut venir réduire cet intervalle de fluctuation normale et remettre en cause l'appartenance de l'"étranger" à cette population de référence.

Plutôt que de raisonner de manière dichotomique, on peut souhaiter le faire en termes de "probabilité d'appartenance" et estimer qu'elle décroîtrait lorsque $\overline{D}_{inter}$ s'éloigne de $\overline{D}_{intra}$. Cette idée est contraire à celle de variabilité normale intrinsèque à tout phénomène naturel. Certains individus peuvent être plus centraux (*Iphigénie*), d'autres plus décalés (*Phèdre*), ils appartiennent pourtant à la même population. On admettra cependant que l'attribution est d'autant plus sûre que l'écart réduit est faible et qu'un grand nombre de distances - entre le texte discuté et les textes composant le corpus – sont comprises dans l'intervalle de variation normale.

Le raisonnement qui vient d'être présenté dans ce chapitre n'est possible qu'à deux conditions (outre le recensement exhaustif et sans erreur de tous les mots).

Premièrement, la distribution de *D* au sein du corpus suit une loi normale, comme cela a été vérifié à la fois grâce à la proximité des trois paramètres centraux et par la distribution « en cloche » des observations autour de ces paramètres centraux.

Deuxièmement, le test est effectué lorsque l'on dispose d'au moins une trentaine d'observations (distances).

Ces conditions peuvent sembler rigides mais il s'agit de conclure de manière solide, quitte à laisser des textes non attribués.

Puisque cette procédure permet de reconnaître l'écrivain, elle va maintenant être employée pour répondre aux questions posées au début de ce rapport. Qui a écrit *Juba* et *Tachmas* ? Plus largement : J.-G. Campistron et J. de La Chapelle ont-ils composé "leurs" pièces ou bien se sont-ils contentés de négocier avec les troupes et avec les éditeurs des textes composés par un autre ?

CHAPITRE IV
J. RACINE – J.-G. CAMPISTRON ET J. DE LA CHAPELLE

La méthode présentée dans le chapitre précédent permet d'identifier le véritable auteur d'un texte. Elle a été appliquée à plus de deux cent pièces du XVIIe siècle (annexe 4) auxquelles ont été confrontées *Juba* et *Tacmas* afin de déterminer l'écrivain qui les a composées. Ces expériences ont fait apparaître, notamment, des anomalies concernant les tragédies présentées par J. de La Chapelle et J.-G Campistron. Ces pièces sont séparées entre elles par des distances indiquant un écrivain unique (première section). Quand elles sont confrontées avec des tragédies comparables présentées par certains écrivains – comme les frères Corneille – on obtient les résultats attendus pour des écrivains différents et notamment des distances intra nettement disjointes des distances inter (deuxième section), alors que ce n'est pas du tout le cas avec les pièces présentés par J. Racine entre 1664 et 1677. Les résultats indiquent au contraire que la même plume a composé ces tragédies (d'*Andromaque* à *Phèdre*), les trois présentées par J. de la Chapelle et toutes celles parues sous le nom de J.-G. Campistron ainsi que *Tachmas* et *Juba* (troisième section).

I. DISTANCES INTRA

Le calcul fait apparaître l'unité des œuvres dramatiques parues sous les noms de J. de La Chapelle et de J.-G Campistron. Il permet aussi de conclure que la main qui a composé ces deux œuvres est la même.

J. de La Chapelle

La petite comédie des *Carrosses d'Orléans* (1680) fera l'objet d'un traitement à part.

Quatre tragédies ont été jouées sous le nom de J. de La Chapelle, toutes à la Comédie française. Ces pièces ont été des succès. J. de La Chapelle a même obtenu un triomphe avec *Cléopâtre* (1681). Trois de ces tragédies seulement ont été publiées. Nous les avons retranscrites en français contemporains en suivant les mêmes conventions que pour les pièces des frères Corneille ou de J. Racine.

Dans le tableau 1 ci-dessous, elles sont rangées par ordre chronologique et leurs distances mutuelles sont calculées.

Tableau 1. Distances entre les tragédies jouées sous le nom de J. de La Chapelle

	Cléopâtre	Téléphonte	Zaide
Cléopâtre (12 décembre 1681)	-	0,220	0,228
Téléphonte (26 décembre 1682)	0,220	-	0,216
Zaïde (26 janvier 1681)	0,228	0,216	-

La distance moyenne est de 0,221. La dispersion est faible mais il y a trop peu de valeurs pour calculer une variance. Ces distances sont un peu élevées pour des tragédies exactement contemporaines – du moins à l'aune de ce que l'on observe sur d'autres corpus comme ceux de P. et T. Corneille, mais elles sont dans la ligne de celles qu'on obtient sur le noyau de l'œuvre présentée par J. Racine.

J.-G. Campistron

Les comédies et les pièces lyriques – présentées par J.-G. Campistron - seront traitées dans un autre ouvrage. Les archives de la Comédie française permettent de dater avec précision chacune des pièces sauf *Tachmas* et *Juba*. Pour cette dernière, on dispose de l'allusion de Colonia dans un texte paru en 1695 (texte cité au début de ce rapport).

Les tragédies sont rangées par ordre chronologique et les distances sont reproduites dans le tableau 2 avec, en dernière ligne, la distance moyenne de chaque pièce par rapport aux autres. En gras, les valeurs inférieures ou égales à 0.2 qui indiquent un même écrivain sur des thèmes très proches.

La médiane est égale à 0.208, le mode à 0.21, la moyenne 0.210, l'écart-type à 0,0168[1]. De la proximité des trois valeurs centrales, on déduit que la série est "gaussienne" et que 95% des valeurs sont inscrites dans un intervalle de variation de 0,177 à 0.243, ce qui est effectivement le cas.

[1] Valeurs correspondantes sur le corpus J. Racine. Moyenne : 0.22 ; mode : 0.22 ; médiane : 0.22 ; écart type : 0.0156.

Tableau 2. Distances entre les tragédies du corpus J.-G. Campistron (classement chronologique)

	Virginie (1683)	Arminius (1684)	Andronic (1685)	Alcibiade (1685)	Phocion (1688)	Adrien (1690)	Tiridate (1691)	Pompéia (1692)	Aétius (1693)	Juba (1695)
Virginie	0,000	**0,184**	**0,197**	**0,195**	**0,201**	0,224	0,210	**0,203**	0,211	0,242
Arminius	**0,184**	0,000	**0,188**	**0,187**	**0,197**	0,208	0,213	**0,196**	**0,200**	0,218
Andronic	**0,197**	**0,188**	0,000	**0,191**	**0,203**	0,218	**0,195**	**0,200**	**0,194**	0,237
Alcibiade	**0,195**	**0,187**	**0,191**	0,000	**0,194**	0,217	0,209	0,213	**0,191**	0,228
Phocion	**0,201**	**0,197**	**0,203**	**0,194**	0,000	0,213	0,206	0,207	0,209	0,240
Adrien	0,224	0,208	0,218	0,217	0,213	0,000	0,227	**0,202**	0,211	0,236
Tiridate	0,210	0,213	**0,195**	0,209	0,206	0,227	0,000	**0,203**	0,209	0,261
Pompéia	**0,203**	0,196	**0,200**	0,213	0,207	**0,202**	**0,203**	0,000	**0,201**	0,237
Aétius	0,211	**0,200**	**0,194**	**0,191**	0,209	0,211	0,209	**0,201**	0,000	0,231
Juba	0,242	0,218	0,237	0,228	0,240	0,236	0,261	0,237	0,231	0,000
Moyenne	0,207	**0,199**	**0,202**	**0,203**	0,208	0,217	0,215	0,207	0,206	0,237

* NB : la troisième décimale n'est pas significative : elle indique comment arrondir la seconde

La distribution des distances présente une courbe proche de la normale avec une légère asymétrie à gauche. Cette déformation ne remet pas en cause l'appartenance de ces pièces à un ensemble unique. Elle provient essentiellement des distances entre les premières pièces, *Juba* et secondairement *Adrien* et *Tiridate*. L'asymétrie de la distribution semble donc provenir du temps séparant les premières des dernières pièces.

Tableau 3. Histogramme des distances entre les pièces présentées par J.-G. Campistron

Ce corpus est moins homogène que ceux de P. et T. Corneille mais un peu plus resserré que celui des pièces composant le noyau de l'œuvre présentée par J. Racine. Cela permet de conclure que toutes les tragédies présentées par J.-G. Campistron ont bien un même auteur.

Par conséquent, le même écrivain a écrit aussi *Aétius*, *Pompéia* et *Juba* – pour lesquelles il existait un doute puisqu'elles étaient inédites à la mort de J.-G. Campistron. *Juba* est légèrement décalée comme l'indique une moyenne plus élevée (dernière ligne) mais dans les limites de la plage de fluctuation normale et pour une pièce chronologiquement décalée par rapport aux autres. En effet, ce corpus présente une dimension chronologique : de *Virginie* à *Phocion*, les valeurs les plus faibles se trouvent au plus près de la diagonale du tableau (pièces chronologiquement les plus proches). En revanche, *Aétius*, *Pompéia* et *Juba* semblent antérieures. En effet, ces trois pièces sont plus proches des premières, notamment *Arminius* (1684), *Andronic* (1685) et d'*Alcibiade* (1685) que des dernières notamment *Tiridate* (1691). Il est donc possible que ces trois pièces aient été composées avant 1693 et la dernière laissée inachevée. Dans ce cas, l'arrêt de la production tragique, sous le nom de J.-G. Campistron, se situerait au plus tard en 1693.

II. DISTANCES INTER

Puisque les distances intra obtenues sur les pièces présentées par J. de La Chapelle et par J.-G. Campistron indiquent des populations homogènes à auteur unique, la procédure présentée dans le chapitre précédent leur est appliquée. Elles sont d'abord confrontées aux pièces des frères Corneille puis au noyau des tragédies présentées par J. Racine. Selon les résultats obtenus précédemment, les distances inter doivent être de 25 à 40% plus élevées que les distances intra obtenues ci-dessus et les valeurs centrales doivent être significativement différentes (au seuil de choisi).

J. de La Chapelle et les frères Corneille

Les trois tragédies présentées par J. de La Chapelle, confrontées aux dix dernières tragédies de P. Corneille, donnent trente distances inter, soit la limite inférieure de validité de la loi normale (tableau 4). Avec les neuf tragédies de T. Corneille cela donne seulement vingt-sept distances inter (tableau 5). On touche donc à la limite du raisonnement, mais les résultats étant exactement dans la ligne de ce que laissaient attendre ceux présentés dans le chapitre précédent, il est malgré tout possible de conclure à des écrivains différents.

Tableau 4. Distances inter tragédies présentées par J. La Chapelle et P. Corneille (classement chronologique).

La Chapelle / P. Corneille	Zaïde	Téléphonte	Cléopâtre	Moyenne
Oedipe	**0,241**	0,254	0,250	0,248
Toison d'or	**0,243**	0,257	0,249	0,250
Sertorius	0,249	0,265	0,256	0,257
Sophonisbe	0,249	0,270	0,255	0,258
Othon	0,258	0,278	0,271	0,269
Agésilas	0,259	0,283	0,293	0,278
Attila	0,259	0,280	0,268	0,269
Tite	0,255	0,275	0,272	0,267
Pulchérie	0,252	0,278	0,281	0,270
Suréna	0,246	0,270	0,282	0,266
Moyenne	0,251	0,271	0,268	0,263

Toutes les valeurs correspondent à ce qui est attendu dans le cas de deux écrivains différents. La moyenne (0.263), médiane et mode sont confondues, ce qui indique une population normale. Puisqu'il y a une trentaine de valeurs dans le tableau 4, on peut utiliser le raisonnement issu de la loi normale centrée et réduite. La plage de fluctuation des distances inter à 1% est de 0.230-0.298. Les moyennes des distances intra P. Corneille (0.181) et intra La Chapelle (0.220) étant situées hors de cette plage, on rejette l'hypothèse de l'écrivain unique, avec un risque d'erreur de seconde espèce inférieur à 1%. En revanche, le risque de première espèce ne peut être examiné (l'effectif du corpus La Chapelle étant insuffisant pour doter la distance moyenne intra d'un écart type et d'un intervalle de variation normale) mais ce n'est pas nécessaire : toutes les valeurs du tableau 4 sont supérieures à la borne haute de l'intervalle à 99% défini sur les œuvres de P. Corneille (0.142-0221). On a donc moins d'une chance sur 100 de se tromper en affirmant que les deux écrivains sont différents.

Les distances inter sont en moyenne de 32% supérieures aux distances intra, ce qui correspond exactement à la moyenne de l'écart attendu pour deux écrivains différents dans ce type de situation (même genre, thèmes proches, écarts temporels pas trop grands).

Les marges du tableau indiquent que *Zaïde* est la plus "cornélienne" des pièces présentées par J. de La Chapelle et que les deux autres sont nettement plus éloignées (dernière ligne) mais que curieusement ce sont les pièces de P. Corneille les plus anciennes (*Œdipe* et *La Toison d'Or*) qui auraient eu le plus d'influence sur ces trois pièces qui paraissent vingt ans après. Le chapitre précédent a montré que ces deux pièces de P. Corneille ont également inspiré certaines pièces présentées par J. Racine.

Les mêmes conclusions peuvent être tirées de la confrontation entre ces trois pièces et celles de T. Corneille (tableau 5). La moyenne et la médiane sont confondues avec le mode (0,26). Malgré une légère asymétrie à gauche (due à *Zaïde*), il est possible de conclure à deux écrivains différents, dans les mêmes conditions que ci-dessus, puisque la totalité des valeurs du tableau 5 sont plus élevées que la borne haute de l'intervalle de variation normale des distances intra de T. Corneille (0.157-0.225). Il y a donc deux écrivains différents.

Les distances inter sont en moyenne de 27% supérieures aux distances intra, ce qui correspond à l'écart attendu pour deux écrivains différents travaillant dans le même genre, sur des thèmes proches, à la même époque.

Tableau 5. Distances inter : tragédies présentées par J. La Chapelle et par T. Corneille (classement chronologique).

La Chapelle	Zaïde	Téléphonte	Cléopâtre	Moyenne
Stilicon	0,245	0,274	0,279	0,266
Camma	0,238	0,259	0,269	0,256
Persée	0,237	0,266	0,262	0,255
Maximian	0,243	0,278	0,278	0,266
Pyrrhus	0,233	0,263	0,270	0,255
Annibal	0,241	0,258	0,261	0,253
Ariane	0,245	0,278	0,287	0,270
Achille	0,236	0,268	0,254	0,253
Essex	0,247	0,283	0,279	0,270
Moyenne	0,240	0,270	0,271	0,260

Cependant, la première pièce présentée par J. de La Chapelle (*Zaïde)* pose problème. Sa proximité moyenne avec les pièces de T. Corneille (0.24) est remarquable alors que les deux suivantes sont nettement plus éloignées. Du fait du nombre restreint des distances (9), il est impossible de choisir entre deux hypothèses possibles (influence ou collaboration de T. Corneille à cette pièce). Comme facteur de proximité entre *Zaïde* et T. Corneille, il faut signaler le thème qui a été souvent traité par T. Corneille[1] mais cela ne peut expliquer la proximité de *Zaïde* avec des pièces où ce thème n'est pas abordé. Le dernier chapitre de ce rapport reviendra sur ces questions.

[1] *Zaïde* est une pièce sur le travestissement et les quiproquos, notamment amoureux, qu'il entraîne. C'est un thème traité à plusieurs reprises par T. Corneille (notamment *la Femme juge et partie* ou *le Feint Alcibiade*) et par P. Corneille dans le *Dépit amoureux*.

J.-G. Campistron et les frères Corneille

Les distances inter Campistron-P. Corneille sont présentées dans le tableau 6. En gras les deux distances inférieures à 0.25.

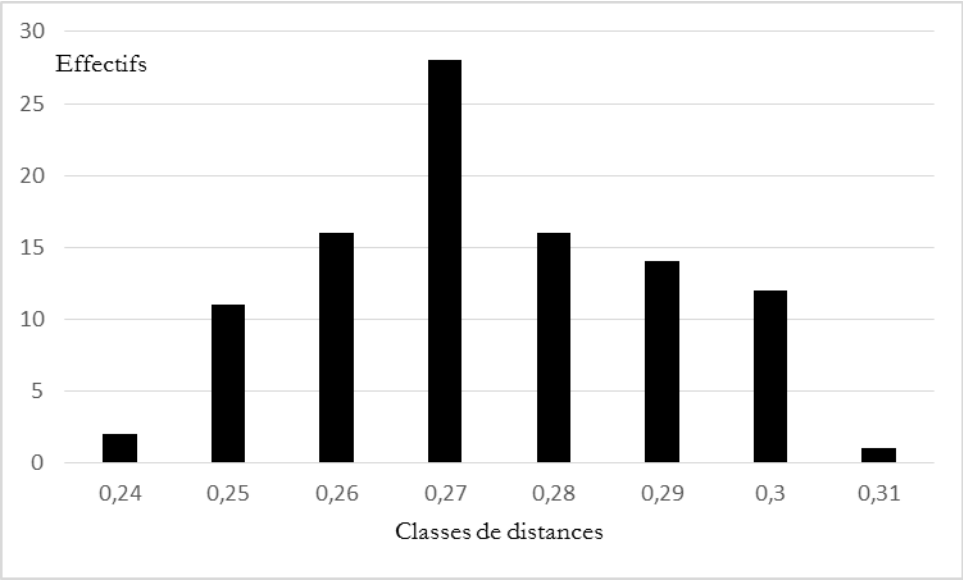
Tableau 6. Distances entre les pièces présentées par J.-G. Campistron et les dernières tragédies de P. Corneille (classement chronologique).

Campistron P. Corneille	Virginie	Arminius	Andronic	Alcibiade	Phocion	Adrien	Tiridate	Pompéia	Aétius	Juba
Œdipe	0,253	0,248	0,246	0,249	0,266	0,277	0,266	0,268	0,261	0,274
Toison d'or	0,251	**0,242**	0,247	**0,244**	0,268	0,267	0,266	0,267	0,255	0,275
Sertorius	0,258	0,246	0,251	0,252	0,283	0,297	0,280	0,279	0,268	0,277
Sophonisbe	0,259	0,254	0,260	0,262	0,285	0,297	0,277	0,278	0,270	0,278
Othon	0,267	0,269	0,262	0,272	0,300	0,307	0,289	0,290	0,275	0,294
Agésilas	0,277	0,279	0,270	0,276	0,306	0,316	0,287	0,301	0,285	0,310
Attila	0,268	0,260	0,261	0,261	0,297	0,292	0,289	0,290	0,272	0,287
Tite	0,265	0,269	0,258	0,270	0,294	0,306	0,276	0,288	0,269	0,305
Pulchérie	0,269	0,271	0,260	0,272	0,300	0,304	0,280	0,289	0,267	0,299
Suréna	0,263	0,268	0,254	0,256	0,290	0,301	0,270	0,281	0,266	0,292
Moyenne	0,263	0,260	0,257	0,261	0,289	0,296	0,278	0,283	0,269	0,289

Les paramètres de la série sont les suivants : moyenne : 0.275 ; médiane : 0.271 ; mode : 0.27 ; écart-type : 0.0172. La proximité des trois paramètres centraux et la distribution des distances autour de ces valeurs centrales indiquent une population normale (tableau 7 ci-dessous).

Etant donné le nombre des distances inter (100) et leur distribution, l'écart type (0.171) permet de définir un intervalle de variation - 95% des distances sont comprises entre 0.242 et 0.308 – et de choisir entre les deux hypothèses alternatives : un ou deux écrivains différents ? Ici les moyennes intra des deux corpus sont nettement situées en dehors de cet intervalle. On peut donc conclure directement à deux écrivains différents avec des risques d'erreur de première et deuxième espèces négligeables.

Tableau 7. Histogramme des distances inter P. Corneille - J.-G. Campistron



Tous les paramètres de la série correspondent à ce qui est attendu pour deux écrivains différents (tels qu'obtenus sur les frères Corneille ou J. Racine). Les distances inter sont en moyenne de 41% supérieures aux distances intra. Pour des écrivains contemporains travaillant dans un même genre et sur des thèmes voisins, moins de 5 distances inférieures à 0.25 sont attendues. De fait, il y en a deux (en gras sur le tableau). Cependant, dans une distribution aléatoire, ces distances "anormales" ne devraient pas être concentrées sur la *Toison d'or* (avec *Arminius* et *Alcibiade*). Le chapitre précédent a montré que cette même pièce de P. Corneille a également inspiré certaines pièces présentées par J. Racine.

Il est possible de reprendre le raisonnement présenté dans le chapitre précédent. Si les deux corpus sont issus de la même plume, leurs distances inter respectives ne doivent pas s'écarter significativement de la moyenne des distances intra. Le calcul donne :

$$u = \frac{0.2633 - 0.1961}{0.0161} = 4{,}1$$

De la même manière, on obtient $u' = 3{,}9$.

En se reportant aux extraits de la table de la loi normale (tableau 3, chapitre III), on peut conclure - avec des risques d'erreur de première et seconde espèce inférieurs à un pour dix mille - que les deux écrivains sont différents.

Les résultats de la comparaison entre les tragédies présentées par J.-G. Campistron et celles de T. Corneille sont semblables. Les distances inter sont en moyenne de 34% supérieures aux distances intra, exactement la proportion moyenne attendue pour deux écrivains différents dans un même genre. Parmi les 90 distances, seules trois sont inférieures à 0.25 (en gras sur le tableau 8).

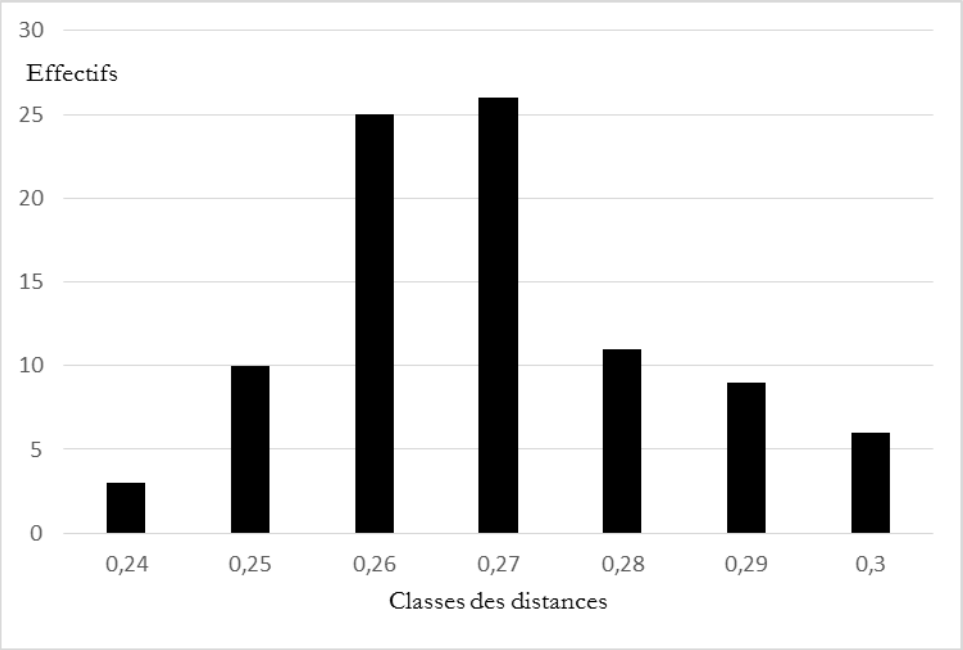
Tableau 8. Distances entre les pièces présentées par J.-G. Campistron et les tragédies de T. Corneille (classement chronologique).

Campistron T. Corneille	Virginie	Arminius	Andronic	Alcibiade	Phocion	Adrien	Tiridate	Pompéia	Aétius	Juba
Stilicon	0,254	0,266	0,258	0,260	0,271	0,284	0,269	0,269	0,268	0,285
Camma	**0,241**	0,260	0,253	0,258	0,265	0,279	0,265	0,256	0,267	0,293
Persee	0,246	0,248	0,250	0,247	0,267	0,281	0,260	0,260	0,259	0,277
Maximian	0,255	0,267	0,255	0,265	0,277	0,294	0,273	0,267	0,267	0,300
Pyrrhus	0,251	0,261	0,257	0,262	0,272	0,293	0,264	0,271	0,270	0,301
Annibal	0,255	0,246	**0,242**	0,255	0,280	0,292	0,263	0,270	0,261	0,285
Ariane	0,257	0,280	0,263	0,267	0,291	0,306	0,258	0,274	0,271	0,316
Achille	**0,244**	0,248	0,253	0,255	0,271	0,282	0,255	0,269	0,259	0,291
Essex	0,261	0,277	0,264	0,267	0,287	0,308	0,268	0,282	0,280	0,313
Moyennes	0,251	0,261	0,255	0,260	0,276	0,291	0,264	0,269	0,267	0,296

Les paramètres des distances inter : médiane : 0.263 ; moyenne 0.269 ; mode : 0.27 et l'écart type : 0.0163 – indiquent une distribution normale qui est vérifiée par l'histogramme des distances (tableau 9).

La plage de fluctuation des distances inter est comprise entre 0.238 et 0.300, c'est-à-dire clairement au-dessus des distances moyennes intra T. Corneille (0.1919) et intra Campistron (0.2101). Les écarts réduits (4.73 et 3.66) indiquent qu'il y a moins de une chance sur mille de se tromper en considérant que les deux écrivains sont différents.

Tableau 9. Histogramme des distances inter, tragédies présentées par T. Corneille et J.-G. Campistron



Pour les deux frères Corneille, confrontés à J.-G. Campistron, les histogrammes regroupant les distances intra et inter font apparaître des distributions disjointes non reproduites car semblables à celles des tableaux 18 à 20 du chapitre précédent.

Ces expériences étaient importantes car elles ont confirmé que la méthode est capable de discriminer des écrivains différents et qu'il y a, dans les résultats, une nette régularité qui interdit de considérer que ceux présentés au chapitre précédent sont de simples "accidents". On s'attend donc à ce que la confrontation entre les pièces présentées par J. Racine, J. de La Chapelle et J.-G. Campistron aboutisse à des résultats semblables : distances intra significativement inférieures aux distances inter (de 25 à 40%), moyennes des distances inter situées au-dessus de l'intervalle de variation des distances intra.

J.-G. Campistron et J. Racine

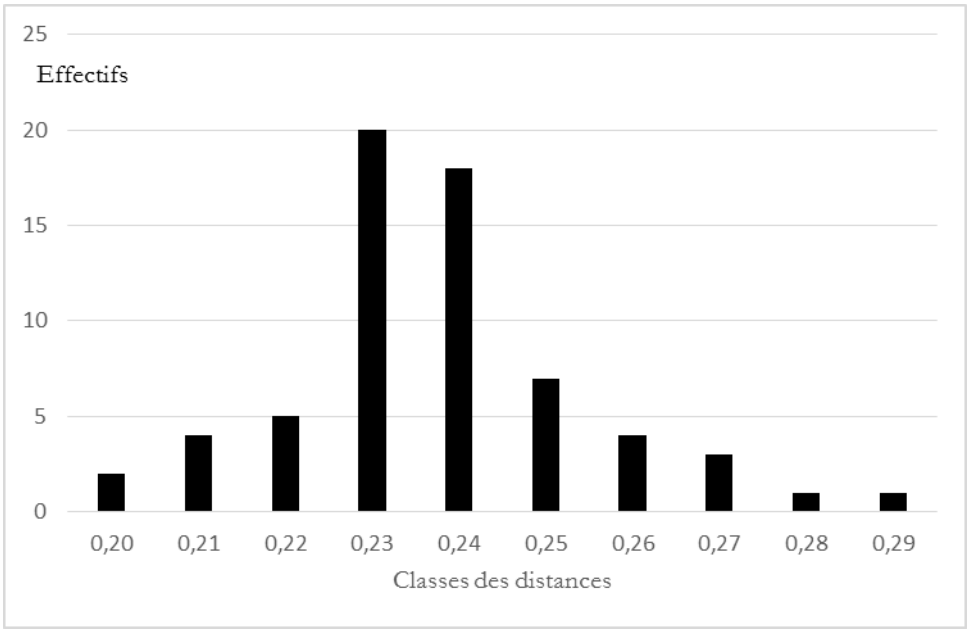
Les distances inter (Campistron - Racine) sont présentées dans le tableau 10 ci-dessous. Les dates de création des pièces sont indiquées car le décalage temporel est parfois important : en moyenne 17 ans séparent les tragédies présentées sous les noms de J.-G. Campistron et de J. Racine (cette question est traitée dans la troisième partie).

Tableau 10. Distances inter tragédies présentées J. Racine et J.-G. Campistron (classement chronologique)

Racine / Campistron	Andromaque (1667)	Britannicus (1669)	Bérénice (1670)	Bajazet (1672)	Mithridate (1672)	Iphigénie (1674)	Phèdre (1677)	Moyenne
Virginie (1683)	0,237	0,236	0,240	0,234	**0,209**	**0,212**	0,235	0,229
Arminius (1684)	0,237	0,232	0,228	0,230	**0,201**	**0,211**	0,232	0,225
Andronic (1685)	0,233	0,229	0,226	0,227	**0,203**	0,223	0,236	0,225
Alcibiade (1685)	0,241	0,233	0,236	0,233	**0,209**	0,216	0,230	0,228
Phocion (1688)	0,249	0,246	0,257	0,238	0,223	0,232	0,227	0,239
Adrien (1690)	0,265	0,270	0,271	0,265	0,244	0,237	0,238	0,256
Tiridate (1691)	0,242	0,234	0,231	0,248	0,224	0,234	0,236	0,236
Pompéia (1692)	0,251	0,251	0,249	0,242	0,216	0,234	0,237	0,240
Aétius (1693)	0,265	0,240	0,233	0,248	0,228	0,241	0,255	0,244
Juba (1695)	0,290	0,271	0,284	0,258	0,242	0,259	0,277	0,269
Moyenne	0,251	0,244	0,246	0,242	0,220	0,230	0,240	0,239

Moyenne des distances inter : 0.239 ; mode : 0.23 médiane : 0.236. Les trois valeurs centrales sont quasiment confondues. L'examen de l'histogramme des distances classées (voir tableau 11 ci-dessous) confirme la normalité de la distribution avec une légère asymétrie à gauche et une queue de distribution à droite qui s'explique par l'écart chronologique considérable entre certaines pièces (chapitre VII).

Tableau 11. Histogramme des distances inter tragédies présentées par J. Racine - J.-G. Campistron



L'écart-type, égal à 0,0180 indique que 95% des valeurs sont comprises entre 0.205 et 0.273. Les distances intra - J. Racine (0.220) et intra - J.-G. Campistron (0.211) sont donc situées dans cet intervalle. Les écarts réduits sont les suivants.

$$u = \frac{0.2390 - 0.2160}{0.0163} = 1{,}40$$

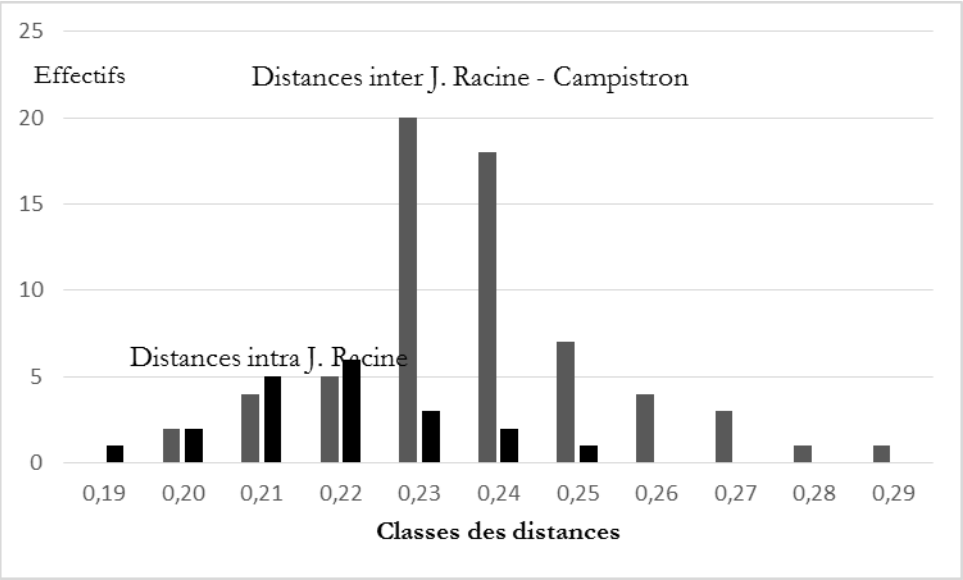
$$u' = \frac{0.2390 - 0.2160}{0.180} = 1{,}28$$

L'hypothèse de deux écrivains différents (H_1) est rejetée. L'hypothèse H_0 (pas de différence significative entre les deux corpus, c'est-à-dire un seul écrivain) est validée. Ce calcul n'était pas indispensable puisque dans le tableau ci-dessus, six distances (en gras) sont si faibles qu'elles permettent de rejeter l'hypothèse de deux écrivains collaborant ponctuellement (cas *Stilicon-Œdipe* examiné plus haut) et ceci d'autant plus que les pièces concernées ne sont pas contemporaines.

On remarquera que, en moyenne, les distances inter sont de 11% supérieures aux distances intra dans ces deux corpus alors qu'on attendait au moins 25% (d'autant que ces textes sont séparés par un écart temporel important). Nous renvoyons le lecteur à ce qui a été indiqué plus haut à propos du poids de l'écrivain : une différence aussi faible ne se rencontre pas entre textes d'écrivains différents. Même chez les frères Corneille, elle est deux fois plus importante.

Dans les comparaisons entre P. et T. Corneille d'une part et J. Racine d'autre part, moins de 5% des distances étaient inférieurs à 0.25. Ici, il y en a 90%. Et les huit distances supérieures à 0.25 séparent des pièces très décalées chronologiquement. Par exemple, 28 ans séparent *Andromaque* de *Juba* et 25 ans *Andromaque* d'*Adrien* (ces deux pièces sont responsables des distances les plus fortes). Ces distances figurent le plus à droite dans l'histogramme ci-dessus. Sauf cette queue de distribution, les deux séries sont extrêmement proches (tableau 12) alors que, sur des écrivains différents, elles sont disjointes ou se recouvrent très faiblement (voir tableaux 17, 18 et 19 du chapitre précédent).

Tableau 12. Histogramme des distances intra tragédies présentées par J. Racine (noir) et inter (comparaison des tragédies présentées par J. Racine / J.-G. Campistron (gris)



Sous réserve de la discussion à venir sur les probabilités d'appartenance (chapitre V), *Juba* - comme toutes les autres tragédies présentées par J.-G. Campistron - est du même écrivain que les pièces présentées par J. Racine entre 1667 (*Andromaque*) et 1677 (*Phèdre*).

Qu'en est-il de *Tachmas* ?

Tachmas

Le fragment de *Tachmas* ne comporte que 4 000 mots. Pour les longueurs de textes inférieures à 5 000 mots, le procédé de calcul présenté dans le premier chapitre ne s'applique qu'imparfaitement (les distances peuvent être instable et sont plus élevées). Un procédé présenté en 2007 permet de neutraliser cet "effet-longueur"[1]. Il consiste à découper des fragments, de la longueur de *Tachmas*, dans les autres textes et à mesurer leurs distances avec *Tachmas*, puis à en faire la moyenne.

Appliqué à *Tachmas*, ce procédé donne les distances suivantes : *Tachmas - Aétius* (1693) : 0.206 ; *Tachmas - Alcibiade* (1685) : 0.218 ; Tachmas – *Zaïde* (1681) (0.219) ; *Tachmas - Tiridate* (1691) : 0.219 ; *Tachmas -Virginie* (1683) : 0.221 ; *Tachmas-Bérénice* (1670) : 0.223, etc. La moyenne des distances de *Tachmas* aux pièces présentées par J.-G. Campistron est de 0.228 ; à celles de J. La Chapelle : 0,232 et à celles de J. Racine : 0.237. Toutes ces valeurs tombent dans les intervalles de variation normale présentés ci-dessus et rattachent donc *Tachmas* au corpus Campistron - La

[1] Labbé Dominique. Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*. 14-1, 1, April 2007, p. 33-80.

Chapelle – Racine. Sa proximité avec *Aétius* (1693) laisse supposer une création contemporaine. Mais comme pour *Aétius*, les autres plus proches voisins suggèrent une création plus ancienne.

Cette estimation n'a pas la même précision que le calcul présenté dans le premier chapitre de cette note – tel qu'il a été appliqué à toutes les autres pièces - mais la convergence des résultats permet de conclure que l'auteur de ce brouillon est bien le même que celui de toutes les pièces, notamment celles de J. de la Chapelle qui vont maintenant être examinées.

J. de la Chapelle, J. Racine et J.-G. Campistron

Les trois tragédies présentées par J. de La Chapelle sont comparées au noyau central de J. Racine et à toutes les tragédies de J.-G. Campistron (tableau 13).

En ce qui concerne la comparaison entre les pièces présentées par J. de La Chapelle et par J. Racine, le nombre des distances étant inférieur à 30, la loi normale ne peut en toute rigueur être appliquée. Il est malgré tout possible de conclure directement en utilisant l'intervalle obtenu sur les pièces présentées par J. Racine (0.190-0.250). Toutes les valeurs du premier cadre du tableau 13 sont comprises dans cet intervalle. On peut donc repousser l'hypothèse H_1 (deux écrivains différents) et retenir H_o (un seul écrivain).

Il en est de même pour le second cadre du tableau 13 (La Chapelle – Campistron). Neuf valeurs sur 10 sont inscrites dans l'intervalle à 95%, malgré des décalages chronologiques importants. L'écrivain est le même.

Ici les distances inter ne sont que de 6% plus élevées que les distances intra. On peut affirmer que cette différence est au moins quatre fois trop faible pour deux écrivains différents. La plume est la même.

Cette conclusion est confirmée par l'examen graphique de l'ensemble des valeurs contenues dans le tableau 13. Le tableau 14 présente cet examen.

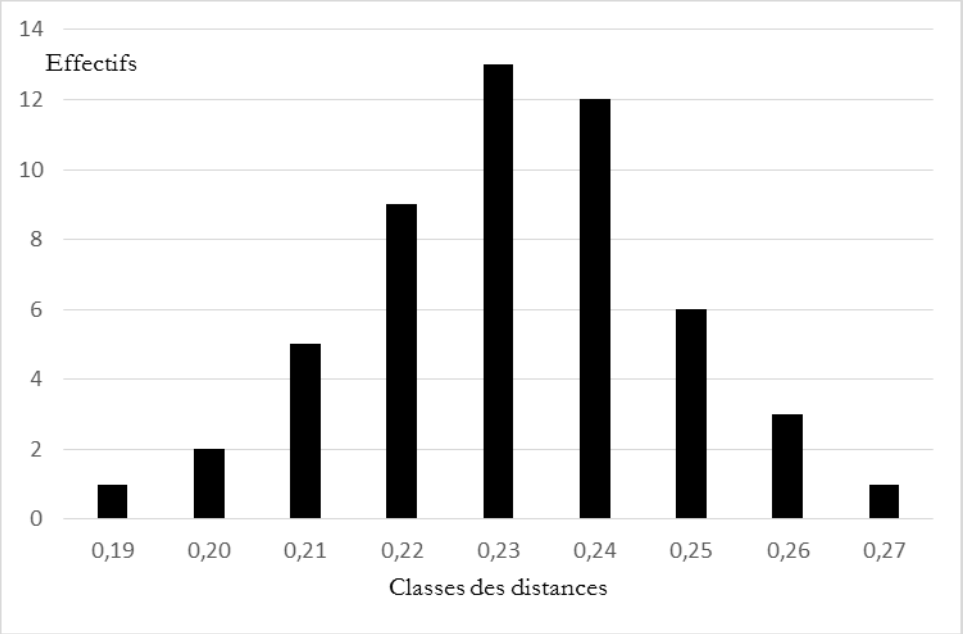
Tableau 13. Distances entre les trois tragédies présentées par J. de La Chapelle et celles présentées par J. Racine et J.-G. Campistron.

La Chapelle	Zaïde (1681)	Téléphonte (1683)	Cléopâtre (1683)	Moyenne
J. Racine				
Andromaque (1667)	0,241	0,240	0,246	0,242
Britannicus (1669)	0,230	0,239	0,236	0,235
Bérénice (1670)	0,240	0,234	0,249	0,241
Bajazet (1672)	0,235	0,242	0,238	0,238
Mithridate (1672)	0,225	**0,204**	0,221	0,217
Iphigénie (1674)	0,234	0,207	0,231	0,224
Phèdre (1677)	0,238	0,238	0,246	0,240
Moyenne Racine	0.235	0.229	0.238	**0.234**
J.-G. Campistron				
Virginie (1683)	0,212	0,211	0,228	0,217
Arminius (1684)	0,213	0,211	0,221	0,215
Andronic (1685)	0,218	**0,194**	0,236	0,218
Alcibiade (1685)	0,205	0,221	0,223	0,216
Phocion (1688)	0,230	0,226	0,246	0,234
Adrien (1690)	0,242	0,233	0,248	0,241
Tiridate (1691)	0,228	0,220	0,251	0,233
Pompéia (1692)	0,226	0,222	0,244	0,230
Aétius (1693)	0,218	0,228	0,237	0,228
Juba (1695)	0,258	0,267	0,259	0,261
Tachmas (?)	0,219	0,227	0,249	0.232
Moyenne Campistron	0.225	0.223	0.259	**0.229**
Moyenne générale	0,229	0,225	0,239	**0,231**

* Seules les deux premières décimales sont significatives. La troisième indique les arrondis

La moyenne (0.231) et la médiane (0.230) sont confondues dans la classe modale (0.23), la distribution des distances est normale (tableau 14). C'est le profil en cloche attendu pour un corpus de textes par un seul écrivain dans un même genre et sur des thèmes proches alors que l'existence de trois prétendus auteurs laissait attendre une courbe multimodale avec des paramètres centraux nettement plus élevés.

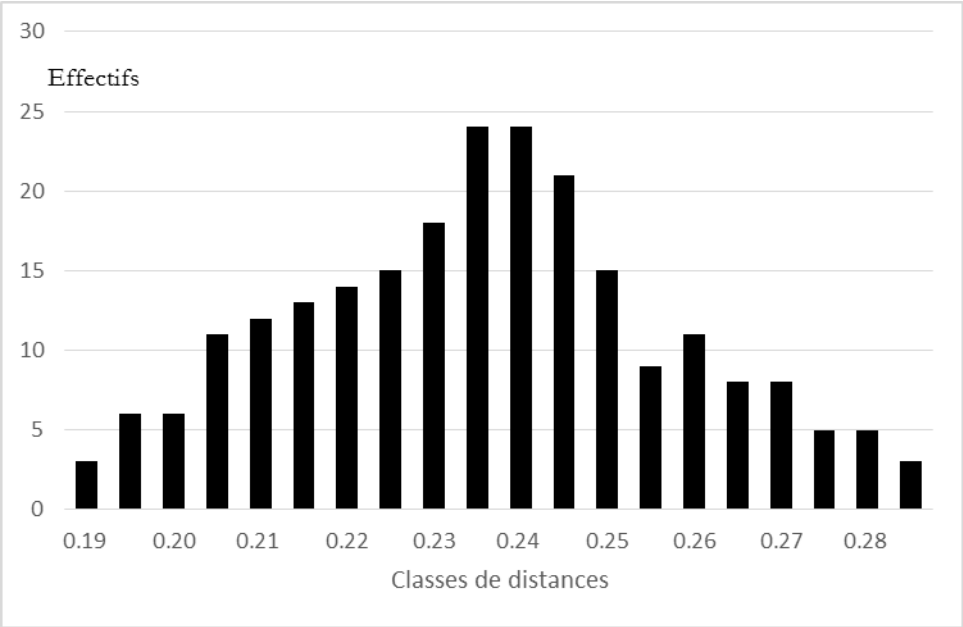
Tableau 14. Histogramme des distances inter tragédies présentées par J. de La Chapelle, par J. Racine et J.-G. Campistron (tableau 13).



L'écart-type (0.0147) indique que 95% des distances sont comprises entre 0.201 et 0.258. Non seulement les trois moyennes intra et la moyenne inter La Chapelle-Campistron-Racine sont toutes situées dans cet intervalle mais 94% des valeurs du tableau 13 ci-dessus tombent dans ce même intervalle. L'hypothèse de trois auteurs différents est rejetée. Toutes ces pièces forment une population homogène et l'écrivain unique est la seule explication possible.

Cette explication est vérifiée en confrontant toutes ensembles les 21 tragédies présentées par J. Racine, J. de la Chapelle et J.-G. Campistron. Cela donne 210 distances intra et inter qui sont reproduites dans le tableau 15 ci-dessous. Ce tableau doit être comparé aux tableaux 18 et 20 ci-dessus (comparaison entre J. Racine et les frères Corneille). Le tableau 15 ne présente pas du tout le profil bimodal attendu.

Tableau 15. Histogramme des fréquences des distances intra et inter tragédies présentées par J. Racine, J. de La Chapelle et J.-G. Campistron.



La courbe présente un profil en cloche, quasiment symétrique et à mode unique : les distances les plus nombreuses sont comprises entre 0.235 et 0.245. La moyenne des distances (0.233) et l'écart type (0.022) tombent d'ailleurs dans l'intervalle de variation normale calculé sur le corpus J. Racine seul. La dispersion entre ces 21 tragédies n'est pas plus forte que celle observée sur les 7 présentées par J. Racine, malgré le fait que leur création s'étend sur 28 ans. Un seul écrivain a composé ces 21 tragédies.

Conclusions de la première partie

Entre 1667 et 1693, trois "auteurs" différents – J. Racine, J. de La Chapelle puis J.-G. Campistron - présentent 19 pièces au public, mais toutes ces pièces, et les trois inédits des archives Campistron, sortent de la même plume. L'information donnée par le Père Colonia est donc vérifiée.

Cette attribution appelle trois commentaires.

Premièrement, plusieurs chercheurs avait fait le rapprochement entre les pièces présentées par J.-G. Campistron et J. Racine[1] mais c'était au titre des "parallèles" très en vogue chez les littéraires. La parenté entre toutes ces pièces n'avait jamais été détectée auparavant car il est impossible de déterminer la plume qui a composé un texte à l'aide des outils traditionnels de la critique littéraire. Cela concerne plus de la moitié des pièces de théâtre du XVIIe siècle qui n'ont pas été présentées sous le nom des écrivains qui les ont composées mais par des intermédiaires – souvent de riches comédiens comme Molière, Montfleury, Hauteroche ou Poisson - qui les négociaient avec les troupes. L'identification des écrivains qui ont composé ces pièces "orphelines" n'en est qu'à ses débuts.

Deuxièmement, l'expérience semblait *a priori* difficile (même sans retenir la théorie du "genre"). Beaucoup de contraintes pèsent sur ces tragédies du XVIIe : les alexandrins, les "règles", un marché étroit et la demande exprimée par les troupes qui choisissent les pièces qu'elles présentent au public. D'ailleurs, ces trois troupes n'en forment plus que deux après 1673 et une seule en 1680 (la Comédie française). Dans un cadre de concurrence imparfaite comme celui-ci, les producteurs – peu nombreux et en interaction constante - visent naturellement le "comédien" médian et ont tendance à calquer leurs produits sur les réussites concurrentes. Bref, tous les écrivains semblent condamnés à faire un peu la même chose. C'était d'ailleurs le principal argument en faveur de la thèse selon laquelle il serait impossible de reconnaître le véritable auteur d'une pièce de théâtre du XVIIe. On a vu que cette thèse est erronée et que, même dans ce contexte contraignant, les écrivains se distinguent fort bien.

Troisièmement, en effet, dans une collection de textes écrits dans un même genre, à la même époque, sur des thèmes proches, tous les facteurs qui concourent à la distance intertextuelle sont neutralisés – ou minimisés – sauf l'écrivain. La distance entre textes écrits par des écrivains différents est augmentée en moyenne de 33% par rapport à ceux sortis d'une seule

[1] Ces rapprochements sont cités dans : Jean-Charles Basson et Dominique Labbé. Un ami tient dignement sa place. Postface…

plume et cette moyenne est inscrite dans un intervalle de 0.25 à 0.40. C'est le poids moyen du "facteur auteur". Ce poids permet d'identifier – sans erreur – les différents écrivains. Dès lors, pour attribuer des pièces d'origine douteuse ou inconnue, il suffit de disposer de collections suffisantes de textes, écrits à la même époque, dans le même genre – et d'origine non douteuse.

Enfin que le fait de travailler sur des recensements exhaustifs (et non sur des échantillons) enlève toute incertitude aux résultats (hors la variabilité normale du phénomène étudié).

A l'issue de cet examen, deux questions se posent qui seront examinées dans la suite de cette note.

- Comment simplifier l'attribution d'auteur ? Deux voies complémentaires sont possibles : les classifications et l'utilisation d'une échelle standardisée de la distance intertextuelle, capable de fournir un premier jugement sur un texte sans avoir besoin de refaire toutes les analyses présentées dans cette première partie.

- Existe-t-il des indices complémentaires pouvant conforter cette attribution ? Les prochains chapitres en présentent deux (l'influence du temps et le style).

DEUXIEME PARTIE
CLASSIFICATIONS

Dans une recherche comme celle-ci, le recours aux classifications automatiques a plusieurs justifications.

On travaille sur 39 textes (40 en comptant *Tachmas*). Cela peut paraître peu. Pourtant, leur comparaison deux à deux génère (40*39)/2 = 780 distances différentes. Un tableau d'une telle dimension est impossible à reproduire et difficile à traiter à la main. Outre l'aspect fastidieux, le risque est grand de commettre des erreurs ou de passer à côté de phénomènes importants. Dès lors, le recours à des procédures automatiques est une nécessité.

Ces classifications aboutissent à des tableaux simplifiés et à des graphiques. Or un dessin est toujours plus parlant que des tableaux de chiffres. Effectivement, les graphiques du prochain chapitre donnent immédiatement les deux informations essentielles : les tragédies présentées par J. Racine, J. de La Chapelle et J.-G. Campistron entre 1667 et 1693 forment un seul groupe ; *Aétius*, *Juba* et *Tachmas* appartiennent à ce même groupe.

Ces procédures commencent à devenir courantes dans la gestion des grandes bases documentaires électroniques.

Deux méthodes sont possibles. D'une part, la classification hiérarchique ascendante (chapitre V) qui procède par agrégations successives jusqu'à ce que tous les textes soient rattachés à un groupe. Elle est donc exhaustive. D'autre part, une classification non-hiérarchique, en cours de mise au point, qui rend mieux compte de toute l'information disponible en acceptant de ne pas aboutir parfois à un classement complet (chapitre VI).

CHAPITRE V.
CLASSIFICATIONS HIERARCHIQUES

La question revient à demander à un automate de trouver – dans un temps limité - le "meilleur" ordre possible dans une population de n individus. Cette classification optimale est définie comme la partition qui minimise les distances internes (intra) dans chacun des groupes qui auront été constitués et qui maximise les distances entre ces groupes (inter)[1]. Autrement dit, les groupes doivent être aussi homogènes que possible mais aussi les plus contrastés entre eux.

Ce classement est réalisé par des automates (algorithmes) non supervisés et aveugles. La non-supervision signifie qu'il n'y a pas d'apprentissage préalable, pour leur apprendre à reconnaître les écrivains et pas d'intervention de l'opérateur, durant le processus, pour orienter l'automate. De plus, l'automate est "aveugle" : il ne dispose d'aucun renseignement autre qu'un numéro pour chaque texte et ses distances à tous les autres. Pour faciliter la tâche du lecteur, les noms des auteurs et les titres de pièces sont rétablis, mais il faut se souvenir que l'automate ne dispose pas de ces informations.

Dans les procédures classiques, la recherche se limite au(x) K-NN textes séparés par les plus courtes distances[2].

Enfin, la classification hiérarchique aboutit à une représentation graphique plus commode à consulter que les grands tableaux de chiffres.

Après un exemple permettant de comprendre comment fonctionne cette technique, elle est d'abord appliquée aux frères Corneille et à J. Racine (deuxième section), puis J. de La Chapelle et J.-G. Campistron sont introduits (troisième section).

---

[1] Sneath Peter & Sokal Robert. *Numerical Taxonomy*. San Francisco: Freeman, 1973. Roux Maurice. Algorithmes de classification. Paris : Masson, 1985. Ouvrage disponible en ligne : http://www.imep-cnrs.com/docu/mroux/algoclas.pdf. Voir

[2] Pour la classification à l'aide des plus proches voisins : Cover Thomas M. & Hart Peter E.. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*. 13, 1967, 21–27. Cover Thomas M. & Thomas J. A.. *Elements of Information Theory*. John Wiley & Sons, 1991. Hall P, Park BU, Samworth RJ (2008). "Choice of neighbor order in nearest-neighbor classification". *Annals of Statistics* 36 (5): 2135–2152. Pour une application aux corpus textuels : Han Eui-Hong, Karypis George & Kumar Vipin. Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification. *Lecture Notes in Computer Science.* Volume 2035, 2001, p. 53-65. Pour l'application à l'attribution d'auteur, voir aussi notre article pour *Images des mathématiques* (Art.cit).

I. UN EXEMPLE DE CLASSIFICATION : P. CORNEILLE

Prenons les tragédies de P. Corneille comme exemple (tableau 1, chapitre 2). L'algorithme agrège les pièces les unes après les autres selon leurs proximités. Ces opérations sont récapitulées dans un dendrogramme.

La classification

L'algorithme procède à la construction d'une classe en regroupant les deux textes séparés par la distance la plus faible ("plus proches voisins"), puis il recalcule les distances des autres textes par rapport à ce nouvel ensemble par la moyenne arithmétique simple des distances, etc. Et ceci jusqu'à la constitution d'un ensemble unique (classement exhaustif).

Dans le tableau 1 du chapitre 2, les deux pièces les plus proches sont *Tite et Bérénice* et *Pulchérie* (distance : 0.153). L'algorithme les agrège en un seul groupe en fusionnant les huitième et neuvième lignes et colonnes du tableau. Le remaniement de la matrice des distances est effectué de la manière suivante :

Tite et Bérénice – *Suréna* : 0.156 ; *Pulchérie* – *Suréna* : 0.158

La distance du nouveau groupe {*Tite et Bérénice- Pulchérie*} à *Suréna* est de :

$$D(\{Tite\ et\ Bérénice\text{- }Pulchérie\},Suréna) = \frac{(0.156+0.158)}{2} = 0.157$$

Les résultats de cette première étape sont récapitulés dans le tableau 1 ci-dessous. Ce tableau a une ligne et une colonne de moins que le tableau original.

Tableau 1. Première étape de la réduction de la matrice des distances par agrégation de *Tite et Bérénice* avec *Pulchérie* et nouveau calcul des distances des autres textes par rapport à ce groupe.

	Oedipe	Toison	Sertorius	Sophonisbe	Othon	Agésilas	Attila	**{Tite - Pulchérie}**	Suréna
Œdipe	0,000	0,194	0,196	0,190	0,194	0,211	0,196	**0,207**	0,194
Toison	0,194	0,000	0,187	0,194	0,199	0,201	0,191	**0,200**	0,201
Sertorius	0,196	0,187	0,000	0,159	0,177	0,173	0,177	**0,172**	0,173
Sophonisbe	0,190	0,194	0,159	0,000	0,171	0,175	0,188	**0,178**	0,180
Othon	0,194	0,199	0,177	0,171	0,000	0,179	0,169	**0,161**	0,174
Agésilas	0,211	0,201	0,173	0,175	0,179	0,000	0,186	**0,162**	0,162
Attila	0,196	0,191	0,177	0,188	0,169	0,186	0,000	**0,178**	0,178
{Tite -Pulchérie}	**0,207**	**0,200**	**0,172**	**0,178**	**0,161**	**0,162**	**0,178**	**0,153**	**0,157**
Suréna	0,194	0,201	0,173	0,180	0,174	0,162	0,178	**0,157**	0,000

L'algorithme conserve en mémoire la distance interne au groupe qu'il vient de former (0.153) – nous l'avons encadrée sur le tableau 1 - car elle fournit une information sur le degré d'homogénéité du groupe ainsi formé, information qui sera reportée sur le graphique récapitulant les opérations successives.

Ensuite, il recherche la plus petite distance dans ce nouveau tableau. Il s'agit de D({Tite - Pulchérie},Suréna) = 0.157. Par le même procédé que ci-dessus, il regroupe les deux dernières colonnes et lignes du tableau 1. Ce qui aboutit au tableau 2 ci-dessous qui a encore une ligne et une colonne de moins.

Tableau 2. Deuxième étape de la réduction de la matrice des distances par agrégation de *Suréna* au groupe {*Tite et Bérénice – Pulchérie*} et nouveau calcul des distances des autres textes par rapport à ce groupe.

	Oedipe	Toison	Sertorius	Sophonisbe	Othon	Agésilas	Attila	**{Tite-Pulchérie-Suréna}**
Œdipe	0,000	0,194	0,196	0,190	0,194	0,211	0,196	**0,201**
Toison	0,194	0,000	0,187	0,194	0,199	0,201	0,191	**0,201**
Sertorius	0,196	0,187	0,000	0,159	0,177	0,173	0,177	**0,172**
Sophonisbe	0,190	0,194	0,159	0,000	0,171	0,175	0,188	**0,179**
Othon	0,194	0,199	0,177	0,171	0,000	0,179	0,169	**0,167**
Agésilas	0,211	0,201	0,173	0,175	0,179	0,000	0,186	**0,162**
Attila	0,196	0,191	0,177	0,188	0,169	0,186	0,157	**0,178**
{Tite-Pulchérie-Suréna}	**0,201**	**0,201**	**0,172**	**0,179**	**0,167**	**0,162**	**0,178**	**0.157**

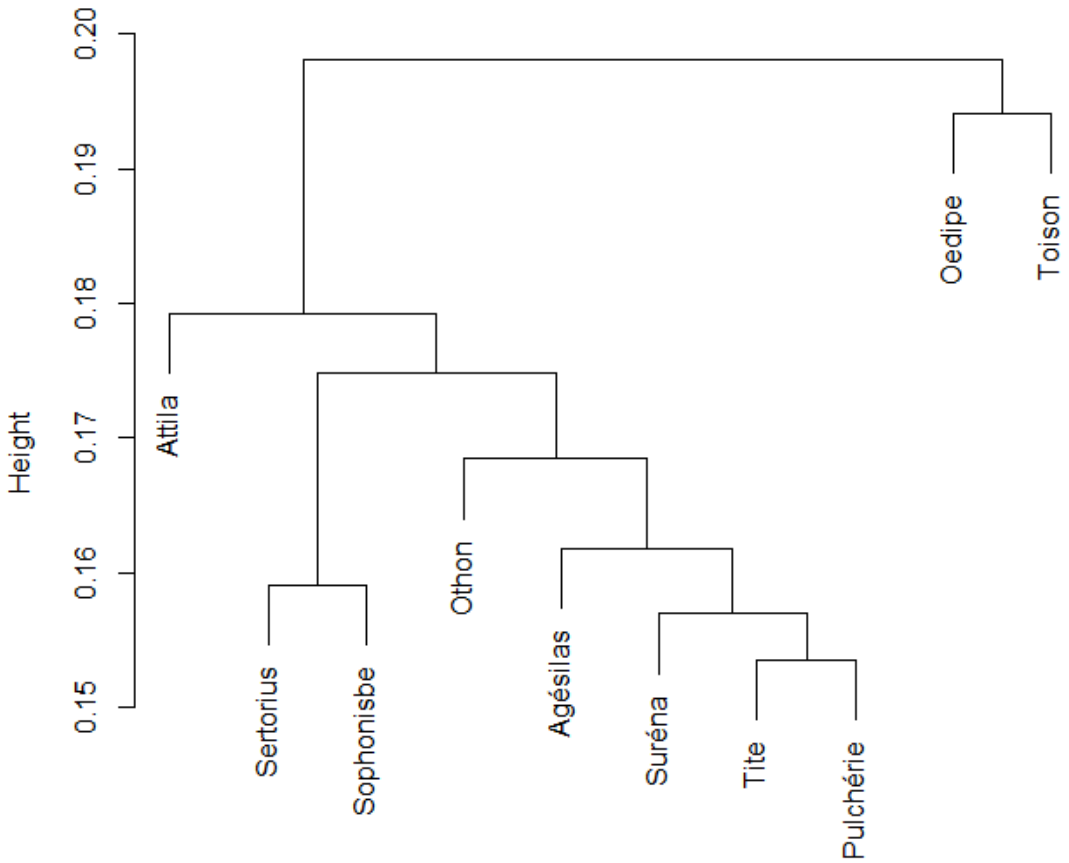
L'opération est répétée jusqu'à ce qu'il ne reste plus qu'un seul groupe.

Le dendrogramme

Ces regroupements successifs sont représentés par un arbre (ou dendrogramme) qui comporte, en ordonnées, les distances correspondantes aux niveaux d'agrégation successifs[1].

[1] Graphiques réalisés avec le logiciel R. Voir Meyer D., Hornik K & Feinerer I. (2008). *Text mining infrastructure in r*. 25(5):569–576.

Tableau 3. Dendrogramme de la classification hiérarchique ascendante sur le corpus des pièces de P. Corneille (méthode de la moyenne).



Il ne faut pas attacher d'importance au classement des textes de gauche à droite (le logiciel place les textes les plus proches au centre du graphe et les plus éloignés à droite mais c'est une convention arbitraire). Seul compte le niveau auquel l'agrégation est réalisée (ligne horizontale reliant les deux textes). Par exemple, *Pulchérie* et *Tite et Bérénice* se rejoignent à 0.153 et *Suréna* les rejoint à 0.157. Le classement s'achève légèrement en dessous de 0.20 qui est la distance moyenne d'*Œdipe* et de la *Toison d'Or* à tous les autres.

Ces niveaux donnent donc une idée de la proximité plus ou moins grande entre les éléments regroupés. En coupant le graphe horizontalement, on peut isoler les groupes de textes très proches, relativement proches, etc. Ce qui fait ressortir nettement l'aspect chronologique du classement, *Tite et Bérénice*, *Pulchérie* et *Suréna* – qui forment le noyau le plus resserré- sont les trois dernières pièces (1670-1674). En haut du graphe, les deux pièces les plus anciennes – *Œdipe* (1659) et *la Toison d'or* (1661) - forment le couple le plus hétérogène de ce corpus et paraissent un peu décalées (avec *Attila*) par rapport au noyau central des pièces de la dernière période de la vie théâtrale de P. Corneille.

Cette méthode a cependant un défaut majeur : pour une pièce particulière, le graphique indique précisément le groupe auquel elle appartient mais pas forcément sa proximité relative à

telle ou telle autre pièce composant ce groupe. Par exemple, le graphe peut laisser penser qu'*Othon* est proche d'*Agésilas* (chemin le plus court) alors qu'on voit dans le tableau 2 qu'*Othon* est très proche de *Pulchérie* (0.158) ou de *Tite et Bérénice* (0.163) et relativement éloignée d'*Agésilas* (0.179). Le procédé d'agrégation successive a effacé les liens les plus forts. Le graphique représente donc correctement les étapes de la classification, non pas certaines liaisons entre individus considérés par couples.

Au total, les œuvres de la dernière partie de la vie théâtrale de P. Corneille se rejoignent légèrement en dessous de 0.20 – au maximum moins d'un mot sur cinq est différent dans les différentes œuvres composant cet ensemble - ce qui est un signe de forte homogénéité. Un seuil aussi bas se rencontre rarement dans un ensemble comportant une dizaine d'œuvres dont la création s'étale sur 15 ans. Cette dernière remarque est importante car, contrairement au lecteur, l'automate n'est pas informé de l'identité de l'écrivain qu'il vient d'identifier à l'aveugle.

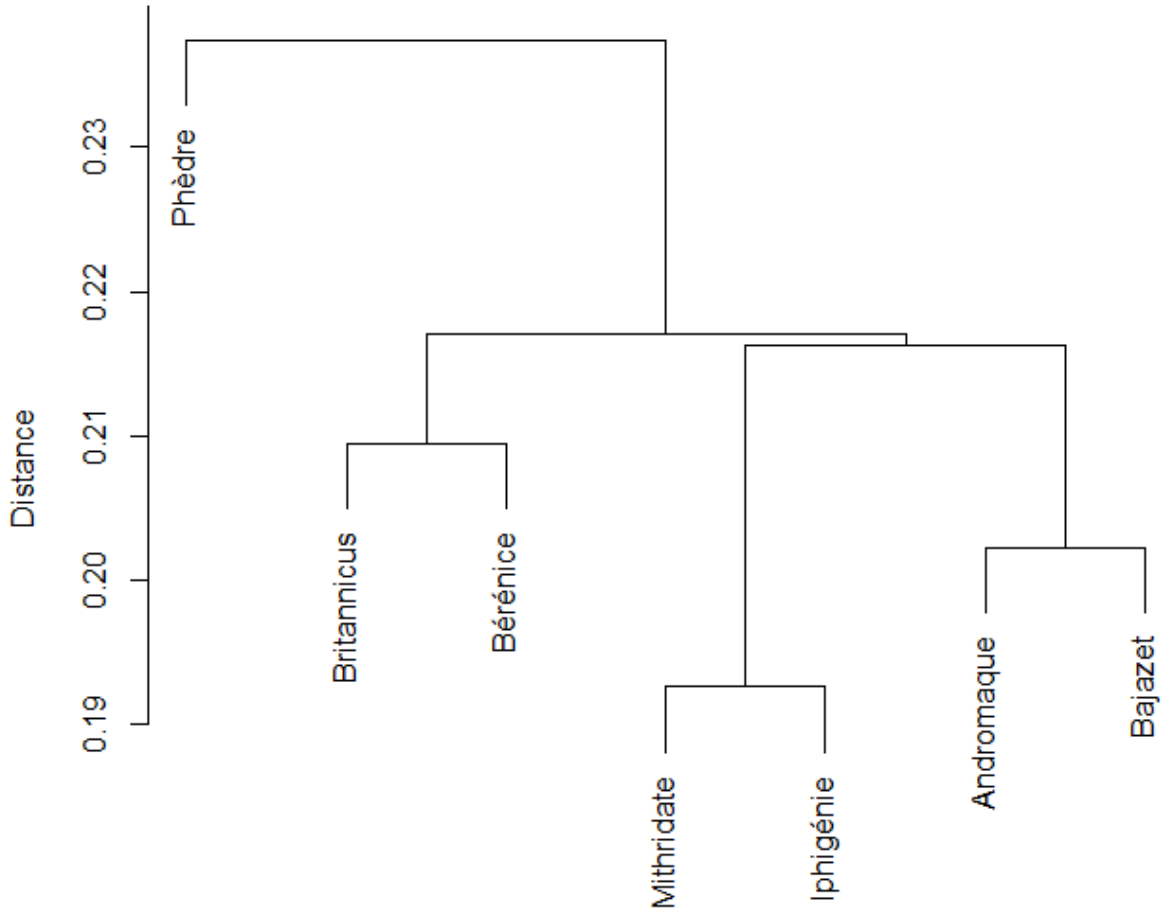
Naturellement, du point de vue de l'analyse littéraire, ce classement est un point de départ. Les groupes étant isolés, leurs vocabulaires et leurs styles sont étudiés pour comprendre ce qui les singularise dans l'œuvre d'un écrivain et d'une époque.

Ces classifications se sont également révélées être de bons outils au service de l'attribution d'auteur. Pour vérifier cette aptitude, la classification a d'abord été appliquée aux trois corpus (les frères Corneille et J. Racine) avant d'y adjoindre les pièces présentées par J. de La Chapelle et J.-G. Campistron.

Classification des œuvres présentées par J. Racine

Les mêmes opérations, appliquées sur le noyau central des tragédies présentées par J. Racine, donnent le graphique suivant (tableau 4).

Tableau 4. Dendrogramme de la classification hiérarchique ascendante sur le noyau central des pièces présentées par J. Racine (méthode de la moyenne).



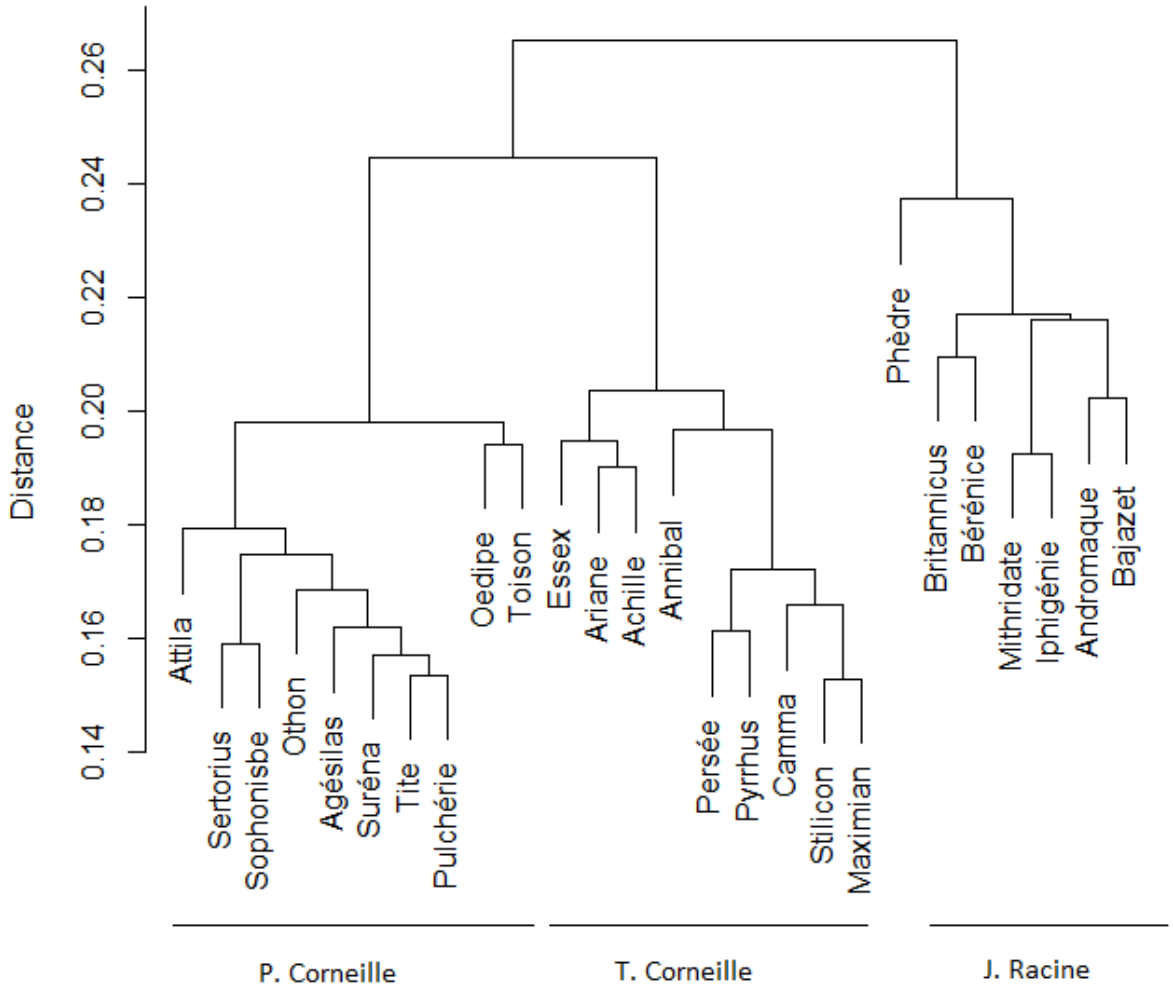
Mithridate et *Iphigénie* forment manifestement le noyau de ces pièces. Autour de ce couple, quatre autres viennent s'agréger également par couples : *Andromaque* et *Bajazet* d'une part, *Britannicus* et *Bérénice* d'autre part. Ces trois couples se rejoignent en dessous de 0.22, ce qui signale une homogénéïté beaucoup plus faible que chez P. Corneille. En revanche, *Phèdre* semble plus lointaine.

Rééditons l'opération en soumettant à l'automate l'ensemble des pièces de ce premier corpus (P. Corneille, J. Racine et T. Corneille).

Vérification sur les trois écrivains

Le résultat correspond à ce qui est attendu pour trois écrivains différents (tableau 5).

Tableau 5. Dendrogramme de la classification hiérarchique ascendante sur le corpus des pièces de J. Racine, P. et T. Corneille (méthode de la moyenne).



A gauche et à droite de la figure, on reconnait les deux diagrammes précédents. Entre eux deux, vient s'intercaler T. Corneille dont l'œuvre apparaît quasiment aussi homogène que celle de

son frère aîné et plus proche de ce dernier que de J. Racine. Les groupes des deux frères se joignent d'abord (au-dessus de 0.24), puis avec celui de J. Racine au-dessus de 0.26.

Les trois œuvres sont bien distinctes et il n'y a aucune erreur de classification, aucune pièce ambigüe. Cela peut sembler normal puisque conforme à la tradition littéraire. Pourtant, comme indiqué ci-dessus, l'épreuve est plus difficile qu'il paraît. Rappelons d'abord que l'automate ne connaît pas les écrivains ni le titre des pièces. Il est donc capable de les identifier sans problème. Deux éléments supplémentaires :

D'une part, alors que les deux frères Corneille font bourse commune, vivent et travaillent sous le même toit – toutes conditions propices à des collaborations - leurs deux œuvres sont proches mais elles se distinguent bien.

De plus, toutes ces pièces sont contemporaines, dans le genre le plus contraignant : la tragédie classique en cinq actes, en alexandrins, sur des thèmes proches et dans le carcan des "règles". Ces trois dramaturges sont en concurrence pour gagner les faveurs des mêmes acteurs et du même public.

En revanche, dix ans après, quand J. de La Chapelle et J.-G. Campistron apparaissent successivement sur la scène, P. Corneille puis J. Racine se sont retirés (du moins le proclament-ils) et le troisième (T. Corneille) ne produit plus de tragédies (sous son nom).

III. … J. DE LA CHAPELLE ET J.-G. CAMPISTRON

Les œuvres présentées sous le nom de J. de La Chapelle et de J.-G. Campistron sont ajoutées aux précédentes. Quels sont les résultats attendus ?

Les résultats attendus…

L'automate doit classer 40 pièces dont il ignore les titres et les auteurs. Mais le chercheur, lui, les connaît et s'attend à ce que l'automate retrouve cinq groupes puisque ces œuvres ont été présentées par cinq "auteurs" différents et que l'histoire littéraire lui affirme qu'il s'agit bien de cinq écrivains différents.

En effet, la première partie a démontré que l'écrivain est le principal facteur discriminant et l'on vient de voir que la classification hiérarchique ascendante discrimine correctement ces écrivains. Dès lors, s'il existe cinq écrivains, comme l'affirme l'histoire littéraire, l'automate doit parvenir à cinq groupes différents, comme il en a reconnu trois précédemment.

De plus, les deux groupes constitués par les œuvres de J. de La Chapelle et de J.-G. Campistron devraient logiquement rejoindre les autres à un seuil plus élevé que celui des pièces des deux frères Corneille qui donnent une sorte d'étalon de la proximité maximale.

Enfin, le décalage temporel doit encore élever ce seuil. En effet, contrairement à J. Racine et aux frères Corneille qui sont contemporains, il existe ici un écart temporel qui peut aller jusqu'au quart de siècle et qui doit augmenter les distances entre les deux nouveaux et les trois "anciens" (la troisième partie revient sur cette dimension temporelle).

- L'hétérogénéité du corpus augmentant, cela devrait se manifester par une élévation des niveaux ultimes d'agrégation (notamment le dernier situé pour l'instant légèrement au-dessus de 0,26).

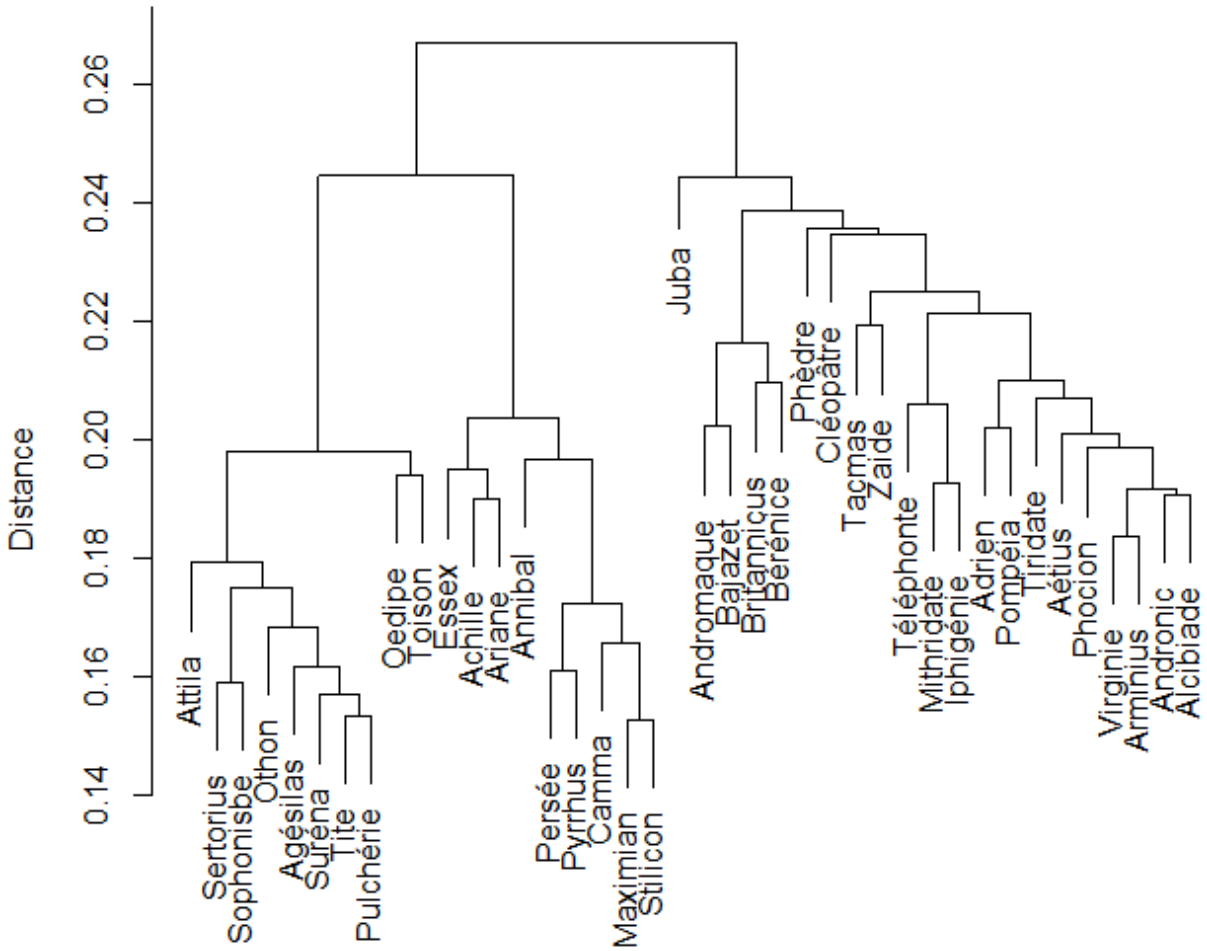
Toutes ces attentes se révèlent bien éloignées de la réalité.

… *loin de la réalité*

Les résultats obtenus sont exactement à l'opposé de ces attentes et soulignent l'étrangeté de la situation (tableau 6). Au lieu des cinq groupes attendus, l'automate n'en a constitué que trois et le stade d'agrégation ultime se situe au même niveau que précédemment (légèrement au-dessus de 0,26).

A gauche, les deux groupes des frères Corneille ne changent pas. A droite, le groupe des œuvres présentées par J. Racine est bouleversé.

Tableau 6. Dendrogramme de la classification hiérarchique ascendante sur le corpus des pièces présentées par J. Racine, P. et T. Corneille, J. de La Chapelle et J.-G. Campistron (méthode de la moyenne).



Le noyau des œuvres présentées sous le nom de Campistron (y compris les quatre actes d'*Aétius*) est amalgamé avec *Mithridate* et *Iphigénie* (les deux pièces les plus centrales de J. Racine) et avec *Téléphonte* (de J. La Chapelle) dont les deux autres pièces (*Zaïde*, *Cléopâtre*) sont rattachées à cet ensemble en même temps que… *Tachmas* et *Phèdre* ! Puis viennent *Andromaque*, *Bazajet*, *Britannicus*, *Bérénice* et enfin *Juba*. En quelque sorte, ce groupe ressemble à un sandwich : une couche de Campistron, une couche de Racine, une couche de La Chapelle, une couche de Racine et un nappage de Campistron ! Sauf *Juba*, toutes se rejoignent à un niveau inférieur à celui des deux frères Corneille, alors que le contraire était attendu.

J. Racine, J. de La Chapelle et J.-G. Campistron ne sont pas frères, n'ont pas eu les mêmes professeurs, n'ont pas épousé des sœurs ni vécu sous le même toit ni fait bourse commune toute leur vie… Il ne reste qu'une explication : un seul écrivain pour les œuvres regroupées dans la partie droite du tableau, dont la création s'étale sur près d'un quart de siècle et qui ont été

présentées sous trois noms différents mais sont pourtant plus proches entre elles que ne le sont les œuvres contemporaines des deux frères Corneille.

Conclusions du chapitre.

La classification donne une représentation fidèle des principaux groupes - existant dans un corpus et de leurs proximités relatives – grâce aux propriétés de la distance intertextuelle. En revanche, les autres métriques - qui n'ont pas les propriétés des distances euclidiennes - engendrent des distorsions d'autant plus grandes que l'on s'élève dans les agrégations successives. Ces défauts sont facilement détectables : au bas du graphique, les groupes semblent compressés et plus on s'élève, plus ils sont exagérément étirés.

Ces calculs n'ont de signification que si les textes ont été dépouillés en suivant la même norme afin de ne pas interpréter les fluctuations dans les graphies d'un même mot comme des différences réelles. Ces fluctuations peuvent entraîner des différences quantitativement significatives mais sans contenu lexical...

Concernant la classification, la méthode employée produit parfois des "effets de chaîne" qui se manifestent par un graphique en forme "d'escalier". Certaines proximités entre textes ne sont alors pas discernables, car les sommets qui les relient sont effacés par des agrégations effectuées à un niveau inférieur. Nous avons déjà donné l'exemple d'*Othon* correctement classée dans les œuvres de P. Corneille mais dont la position au sein de celle-ci est mal représentée.

Pour le corpus Racine-La Chapelle-Campistron, cela est notamment le cas de *Tachmas* et surtout de *Juba*. *Tachmas* est "marié" avec *Zaïde* dont il est séparé par une distance de 0.219 alors que ses plus proches voisins sont *Aétius* (0.206), *Alcibiade* (0.218) et *Tiridate* ex-aequo avec *Zaïde* (0.219). Mais ces trois-là sont groupés avec des textes un peu plus proches d'eux (*Andronic*, *Phocion*, *Adrien*, *Pompéia*…), de telle sorte que *Tachmas* semble décalé par rapport à ses voisins les plus proches, ce qui serait une interprétation erronée. Pour *Juba*, c'est encore pire puisqu'il n'est inclus dans le groupe de droite du tableau qu'à l'ultime étape à une distance proche de 0.25, ce qui efface totalement sa parenté avec *Arminius* (distance : 0.218), *Alcibiade* (0.228), *Aétius* (0.231), etc. L'effet de chaîne aboutit à exagérer son décalage par rapport aux autres pièces du groupe Racine-La Chapelle-Campistron.

L'arbre ne doit donc pas être utilisé aveuglément. L'appartenance de chacun des textes à un groupe est solide. En cela, la classification hiérarchique ascendante est un excellent auxiliaire pour une attribution d'auteur – sous réserve que les distances sur lesquelles elle travaille soient des distances euclidiennes et non de simples mesures de dissimilarité. En revanche, pour le détail, il

n'est pas mauvais de se reporter aux tableaux de distances surtout lorsque certains textes sont représentés comme décalés par rapport aux autres ou lorsqu'une partie du graphe a une structure en escalier.

Ceci conduit à envisager, en complément de cette première étape, des procédures qui restituent ces informations effacées ou déformées par la classification hiérarchique. Elles permettront aussi de répondre à une question : quelle confiance accorder à ces classifications ?

CHAPITRE VI.
CLASSIFICATIONS NON-HIERARCHIQUES

Il convient d'abord de rappeler que l'automate ne connait des textes que des numéros et leurs distances réciproques. Pour faciliter la tâche du lecteur, les auteurs et les titres sont conservés dans ce rapport, mais il faut se souvenir que l'ensemble de l'opération se déroule en aveugle (c'est-à-dire à l'abri de la doxa concernant les textes analysés). Comme dans le chapitre précédent, il s'agit de partitionner une vaste population en groupes aussi homogènes et aussi contrastés que possible. Mais ici, on cherche à conserver toutes les informations qu'efface la classification hiérarchique ascendante et dont nous avons vu l'importance notamment à propos de la place de chaque texte dans le groupe où il est affecté.

Pour cela, l'automate doit :

- considérer l'ensemble de l'information disponible. En particulier, le classement d'un texte dans un groupe ne doit pas détruire les liens entre ce texte et tous les autres, spécialement avec ceux qui sont classés dans d'autres groupes ;

- ne pas chercher à parfaire le classement en affectant tout individu quelque part mais au contraire, interrompre les opérations quand les deux individus à "marier" sont trop éloignés ;

- mesurer le degré d'homogénéité de chacun des groupes et la qualité globale de la classification ;

- affecter à chaque texte, un indice mesurant le nombre de liens unissant ce texte au groupe dans lequel il est classé. Dans le cas d'une recherche en paternité, cela permettra d'indiquer la solidité de l'éventuelle attribution et de répondre à l'objection évoquée dans le troisième chapitre à propos de l'évaluation des risques d'erreur quand on retient l'hypothèse nulle.

Tout cela a un prix : admettre la possibilité d'individus sécants (pouvant appartenir à plusieurs groupes) et, éventuellement, un résidu inclassable.

I. METHODES

Les travaux de Covert et Hart, et ceux qui ont suivi cette voie de recherche, permettent de déduire de la densité du voisinage, un indice d'appartenance au groupe auquel le texte est affecté par l'automate. Cet indice comporte toutefois quelques limites.

Voisinage et indice d'appartenance à un groupe

Selon Covert et Hart, dans une vaste population – au moins trente individus - de répartition gaussienne, le plus proche voisin d'un individu donné contient environ la moitié de l'information totale concernant cet individu. Cette information décroît de manière exponentielle en fonction du rang du voisin considéré. Autrement dit, dans une vaste population distribuée de manière gaussienne, l'information fournie par un voisin diminuerait selon un ratio d'environ 50% en fonction du rang de ce voisin : le deuxième voisin fournirait environ 25% de l'information - le troisième 12,5%, etc. Autrement dit, les trois premiers voisins apporteraient environ 88% de l'information totale et les quatre premiers : 94%, etc. Cinq voisins sont donc nécessaires pour dépasser 97% et six pour atteindre 99%.

Il est proposé de considérer ce ratio comme une approximation de la force des liens unissant un individu à un groupe. Par exemple, quand les quatre premiers voisins d'un texte sont classés dans le même groupe que lui, son indice d'appartenance à ce groupe serait d'environ 94%.

Le raisonnement comporte quelques limites.

Limites

En toute rigueur, ces ratios ne seraient valables qu'en cas de voisinage réciproque. Par exemple, A est le premier voisin de B et B est le premier voisin de A. Leur probabilité d'appartenir à un même groupe serait de 0.5. A et B partagent un deuxième voisin (C), leur indice d'appartenance à un même groupe serait de 0.75, etc. En pratique, cette rigueur n'est pas praticable. Cela tient d'abord à la seconde limite.

Seuls les deux premiers chiffres de la distance sont réellement significatifs. Il y a donc un assez grand nombre d'ex-aequo. Par exemple, dans la classification sur les œuvres de P. Corneille, la distance *Pulchérie-Suréna* (0.1580) est égale à *Pulchérie-Othon* (0.1581). Pourtant dans la procédure automatique, le mariage se fait entre *Pulchérie* et *Suréna*, à cause de la 4^{e} décimale, et la proximité aussi grande entre *Othon* et *Pulchérie* est effacée. On recherche donc une procédure qui n'efface

pas ces proximités et prenne en compte l'information fournie par les deux voisins ensemble. Par exemple deux premiers voisins ex-aequo fournissent environ 75% de l'information totale (soit 37% chacun), trois premiers voisins ex-aequo, 88% (soit chacun 29,3% etc). Autrement dit, la probabilité que A, B et C appartiennent à un même groupe se calcule sur le nombre de voisins communs sans tenir compte de leurs rangs mutuels.

Une troisième limite vient de la question : à partir de quelle distance deux textes cessent-ils d'être "proches" ? En l'absence de toute autre information, il faut utiliser celles contenues dans le tableau des distances. Pour le corpus {T. Corneille – P. Corneille – J. Racine}, la distance moyenne est de 0.233. Faute d'autres informations, il n'est pas raisonnable de poursuivre la recherche de voisinages au-delà de ce seuil (le dernier chapitre revient sur cette question en fournissant des seuils standardisés).

Il faut choisir le moment où la classification doit s'interrompre. Tout dépend d'abord du seuil retenu. Si l'on souhaite approcher les 95%, il faut considérer au moins les quatre premiers voisins ; pour approcher les 99%, au minimum les six premiers voisins de chaque texte sont nécessaires. Il ne semble pas raisonnable d'aller au-delà pour éviter des tableaux trop volumineux. Le choix dépend aussi du nombre d'individus à classer et de la dimension des groupes potentiels. Par exemple, le noyau des pièces de J. Racine ne comporte que 7 pièces. Tout texte appartenant à ce groupe a donc un nombre de voisin potentiel maximum de 6…. Le corpus La Chapelle compte trois pièces (donc théoriquement, on ne devrait trouver que 2 proches voisins). Au fond, une classification n'est intéressante que sur de vastes populations. En tous cas, si l'on souhaite aller jusqu'au sixième voisin, une trentaine d'individus à classer paraissent un minimum. L'expérience porte ici sur 40 pièces (en comptant *Tachmas*).

Lorsque l'information est contradictoire (voisins appartenant à plusieurs groupes différents), l'automate signale que l'individu est sécant, avec indication des groupes auxquels il peut appartenir et l'indice d'appartenance à chacun des groupes.

Dernière remarque, l'automate doit vérifier que la segmentation est non-contradictoire. Il faut notamment qu'aucun des individus les plus lointains, de chacun des membres d'un groupe donné, n'appartiennent à ce même groupe.

Cette procédure sera d'abord présentée sur les corpus T. Corneille – P. Corneille – J. Racine puis ceux de J. de La Chapelle et J.-G. Campistron seront ajoutés.

II. J. RACINE ET LES FRERES CORNEILLE…

L'algorithme commence par constituer les groupes les plus homogènes possibles, à l'aide des plus proches voisins de chaque texte, puis il vérifie la validité de cette classification en considérant les plus lointains.

Constitution de groupes homogènes

L'algorithme classe tous les couples de textes par distances croissantes. Il groupe les deux textes les plus proches de la manière suivante :

- *Tite et Bérénice* - *Pulchérie* (0.153) -> Groupe 1

- *Maximian* - *Stilicon* (0.153) -> Groupe 2

- *Suréna* et *Tite et Bérénice* (0.156) -> Groupe 1 {*Tite et Bérénice – Pulchérie – Suréna*}

Mais contrairement au chapitre précédent, l'algorithme conserve les distances internes aux groupes sous forme de tableaux en mémoire. Par exemple, l'état du groupe 1 à ce stade de la classification (tableau 1 ci-dessous).

Tableau 1. Tableau de classification : premier groupe, deux premiers voisins, classement par ordre d'agrégation.

	1er voisin	1er distance	2e voisin	2e distance
Tite et Bérénice	Pulchérie	0,153	Suréna	0,156
Pulchérie	Tite et Bérénice	0,153	Suréna	0,158
Suréna	Tite et Bérénice	0,156	Pulchérie	0,158

L'automate poursuit ses agrégations successives en suivant l'ordre des distances croissantes (et en complétant les tableaux en mémoire) :

- *Othon* et *Pulchérie* (0.158) -> Groupe 1 {*Tite et Bérénice – Pulchérie – Suréna – Pulchérie - Othon*}

- *Agésilas* et *Tite et Bérénice* (0.159) -> Groupe 1 {*Tite et Bérénice – Pulchérie – Suréna – Pulchérie – Othon - Agésilas*}

[…]

- *Persée et Démétrius* et *Pyrrhus* (0.161) -> Groupe 2 {*Maximian – Stilicon – Persée et Démétrius – Pyrrhus*}

[…]

- *Mithridate* et *Iphigénie* (0.193) -> Groupe 3

- *Andromaque* et *Bajazet* (0,202) -> Groupe 4

- *Mithridate* et *Bajazet* (0.204) -> Les groupes 3 et 4 sont fusionnés {*Mithridate – Iphigénie – Bajazet*} Etc.

L'ordre d'agrégation respecte les distances individuelles : chaque texte est effectivement classé avec ses plus proches voisins. Par exemple, *Othon*, qui était mal situé sur l'arbre de la classification hiérarchique, retrouve dans le tableau 2 sa vraie place au plus près du couple *Pulchérie* et *Tite et Bérénice*.

Au bout du compte, l'algorithme aboutit à trois groupes bien disjoints.

Le tableau 2 récapitule toutes les opérations successives jusqu'au sixième voisin pour le groupe 1 (le premier à être achevé).

Ce tableau indique que les 6 premiers voisins de ces 10 textes appartiennent tous à ce groupe et que les distances sont toutes inférieures au seuil (0.233). Les liens d'appartenance à ce groupe sont donc maximaux pour tous ces textes. Ou encore, ce classement peut être accepté avec une incertitude négligeable.

Le groupe 2 (T. Corneille) présente les mêmes caractéristiques (tableau non reproduit).

Le groupe 3 (J. Racine, tableau 3) fait apparaître quelques difficultés

Seuls *Mithridate* et *Iphigénie* atteignent l'indice d'appartenance maximal : les six voisins appartiennent au même groupe et toutes les distances jusqu'à la sixième sont inférieures au seuil choisi (0.233). Quatre autres dépassent 95%. Le plus faible score est obtenu par *Phèdre* (75%) : seules les distances avec les deux premiers voisins sont inférieures au seuil choisi.

On remarque également que pour *Andromaque* (1667) et *Britannicus* (1669)*, la Mort d'Achille* (T. Corneille, 1673) et *Sertorius* (P. Corneille, 1662) précèdent *Phèdre* (1677) qui n'est que le 7e voisin (non présent sur le tableau qui est limité à six). Autrement dit, pour *Andromaque* et *Britannicus* l'indice d'appartenance à ce groupe est de 97%. Les proximités avec les frères Corneille pouvant être négligées, à ce stade, puisque les distances sont supérieures ou égales au seuil choisi (0.233). Ces anomalies ne remettent pas en cause la classification, mais elles sont intéressantes pour l'histoire littéraire. La principale explication réside dans l'écart chronologique entre *Phèdre* d'une part, *Andromaque* et *Britannicus* d'autre part (comme le montrera le prochain chapitre). Cependant, on ne peut écarter la possibilité d'une influence mutuelle. La *Mort d'Achille* étant postérieure à *Andromaque*, la proximité de ces deux pièces indique une influence *d'Andromaque* sur le cadet des frères Corneille. En revanche, *Britannicus* est postérieur à *Sertorius*, l'influence est ici de P. Corneille sur l'auteur de *Britannicus*. Dans ces deux cas, la proximité des thèmes ajoute ses effets aux influences mutuelles.

Tableau 2. Le premier groupe (Tragédies de P. Corneille avec leurs six plus proches voisins). Classement par ordre d'agrégation.

	1e voisin	1e distance	2e voisin	2e distance	3e voisin	3e distance	4e voisin	4e distance	5e voisin	5e distance	6e voisin	6e distance
Tite et B.	Pulchérie	0,153	Suréna	0,156	Agésilas	0,159	Othon	0,163	Sertorius	0,171	Sophonisbe	0,177
Pulchérie	Tite et B.	0,153	Suréna	0,158	Othon	0,158	Agésilas	0,165	Sertorius	0,173	Attila	0,176
Suréna	Tite et B.	0,156	Pulchérie	0,158	Agésilas	0,162	Sertorius	0,173	Othon	0,174	Attila	0,178
Othon	Pulchérie	0,158	Tite et B.	0,163	Attila	0,169	Sophonisbe	0,171	Suréna	0,174	Sertorius	0,177
Agésilas	Tite et B.	0,159	Suréna	0,162	Pulchérie	0,165	Sertorius	0,173	Sophonisbe	0,175	Othon	0,179
Sertorius	Sophonisbe	0,159	Tite et B.	0,171	Pulchérie	0,173	Agésilas	0,173	Suréna	0,173	Attila	0,177
Sophonisbe	Sertorius	0,159	Othon	0,171	Agésilas	0,175	Tite et B.	0,177	Pulchérie	0,179	Suréna	0,180
Attila	Othon	0,169	Pulchérie	0,176	Sertorius	0,177	Suréna	0,178	Tite et B.	0,180	Agésilas	0,186
Toison d'or	Sertorius	0,187	Attila	0,191	Sophonisbe	0,194	Œdipe	0,194	Tite et B.	0,197	Othon	0,199
Œdipe	Sophonisbe	0,190	Toison	0,194	Othon	0,194	Suréna	0,194	Sertorius	0,196	Attila	0,196
Moyenne		0,164		0,170		0,173		0,176		0,179		0,182

Les dix textes atteignent le score maximal (98%) : leurs six premiers voisins appartiennent au même groupe qu'eux avec un indice d'appartenance égal.

Tableau 3. Les six plus proches voisins des tragédies composant le noyau des tragédies de J. Racine (par ordre d'agrégation)*.

Texte	1e voisin	1e distance	2e voisin	2e distance	3e voisin	3e distance	4e voisin	4e distance	5e voisin	5e distance	6e voisin	6e distance
Mithridate	Iphigénie	0,193	Bajazet	0,204	Bérénice	0,206	Andromaque	0,208	Britannicus	0,218	Phèdre	0,224
Iphigénie	Mithridate	0,193	Phèdre	0,216	Britannicus	0,222	Andromaque	0,222	Bérénice	0,226	Bajazet	0,230
Andromaque	Bajazet	0,202	Mithridate	0,208	Britannicus	0,214	Iphigénie	0,222	Bérénice	0,227	***Achille***	***0,233***
Bajazet	Andromaque	0,202	Mithridate	0,204	Britannicus	0,206	Bérénice	0,220	Iphigénie	0,230	*Phèdre*	*0,244*
Britannicus	Bajazet	0,206	Bérénice	0,209	Andromaque	0,214	Mithridate	0,218	Iphigénie	0,222	***Sertorius***	***0,246***
Bérénice	Mithridate	0,206	Britannicus	0,209	Bajazet	0,220	Iphigénie	0,226	Andromaque	0,227	*Phèdre*	*0,250*
Phèdre	Iphigénie	0,216	Mithridate	0,224	*Bajazet*	*0,244*	*Andromaque*	*0,245*	*Britannicus*	*0,246*	*Bérénice*	*0,250*
		0,203		0,211		0,218		0,223		0,228		*0,240*

* En gras, les classements "anormaux" : ce voisin est déjà classé dans un autre groupe (ici avec au moins 6 distances inférieures)

* En italiques, les voisins séparés par des distances supérieures à la moyenne (0,233) qui ne doivent pas être pris en compte dans le calcul d'un indice d'appartenance.

- *Mithridate* et *Iphigénie* ont un indice maximal d'appartenance à ce groupe.

- *Andromaque*, *Bajazet*, *Britannicus* et *Bérénice* ont un indice d'appartenance de 95% à ce groupe (cinq premiers voisins).

- *Andromaque* est sécante mais avec un fort indice d'appartenance au même groupe que ses cinq premiers voisins (J. Racine).

- les distances entre *Andromaque* et *la Mort d'Achille* (T. Corneille) et entre *Britannicus* et *Sertorius* (P. Corneille) sont égales ou supérieures au seuil et ne doivent pas être considérée.

- l'indice d'appartenance de *Phèdre* est de 75%.

L'opération confirme que l'homogénéité des deux premiers groupes (les pièces des frères Corneille) est plus forte que celle des pièces présentées par J. Racine (moyennes en dernières lignes des tableaux)

Dans la classification hiérarchique traditionnelle, ces informations sont perdues alors que, sans remettre en cause le découpage en trois groupes principaux, elles sont susceptibles d'apporter des éléments intéressants à l'analyste. Les contreparties sont évidentes : la masse des données retournées par l'automate peut être considérable et le classement peut ne pas être exhaustif.

Les textes les plus lointains

Dans le cas présent tous les textes sont classés et, sauf *Phèdre*, ils le sont avec des indices d'appartenance supérieurs à 95%. Dans ce cas, il est légitime de considérer que l'on a abouti au meilleur classement possible et l'étape suivante n'a pas de raison d'être. Dans le cas contraire (grand nombre de groupes, classification avec résidu ou appartenance douteuse pour certains textes), l'automate vérifie que la disjonction entre les groupes aboutit à la meilleure coupure possible : les pièces séparées par les plus grandes distances doivent être affectées à des groupes différents.

On choisit un seuil déterminé (par exemple, toutes les distances supérieures à 0.233). Généralement, il n'est pas nécessaire d'aller jusqu'aux six plus lointains pour cette vérification. A titre d'exemple, le tableau 3 donne les trois plus lointains pour les pièces de J. Racine et P. Corneille.

Toutes les pièces les plus lointaines appartiennent bien à d'autres groupes. De plus, la seconde condition (maximiser les distances inter) est vérifiée grâce aux moyennes (dernières lignes des deux cadres du tableau 4). De nouveau, le "contraste" est légèrement meilleur avec les tragédies de T. Corneille qu'avec celles de J. Racine.

Tableau 4. Textes les plus lointains pour chacune des tragédies de J. Racine et de P. Corneille

	1e lointain	1e distance	2e lointain	2e distance	3e lointain	3e distance
J. Racine						
Andromaque	Maximian	0,264	Agésilas	0,263	Pulchérie	0.263
Britannicus	Camma	0.278	Ariane	0.272	Stilicon	0,270
Bérénice	Stilicon	0,298	Maximian	0,295	Attila	0.289
Bajazet	Agésilas	0.271	Attila	0.271	Stilicon	0.262
Mithridate	Essex	0.267	Maximian	0.266	Stilicon	0.265
Iphigénie	Ariane	0.292	Essex	0.290	Agésilas	0,285
Phèdre	Agésilas	0.318	Pulchérie	0.306	Attila	0.305
Moyennes		0,284		0,280		0,277
P. Corneille						
Oedipe	Bérénice	0,276	Phèdre	0,264	Bajazet	0,257
Toison d'or	Bérénice	0,272	Phèdre	0,271	Britannicus	0,262
Sertorius	Phèdre	0,297	Iphigénie	0,260	Bérénice	0,258
Sophonisbe	Phèdre	0,296	Iphigénie	0,264	Bérénice	0,262
Othon	Phèdre	0,303	Iphigénie	0,275	Bérénice	0,274
Agésilas	Phèdre	0,318	Iphigénie	0,285	Bérénice	0,278
Attila	Phèdre	0,306	Bérénice	0,289	Iphigénie	0,275
Tite et Bérénice	Phèdre	0,302	Iphigénie	0,281	Bajazet	0,262
Pulchérie	Phèdre	0,306	Iphigénie	0,278	Bérénice	0,271
Suréna	Phèdre	0,298	Iphigénie	0,275	Bérénice	0,264
Moyennes		0,297		0,274		0,266
T. Corneille						
Achille	Phèdre	0,274	Bérénice	0,269	Britannicus	0,267
Annibal	Phèdre	0,291	Iphigénie	0,264	Bérénice	0,263
Ariane	Iphigénie	0,292	Phèdre	0,292	Alexandre	0,282
Camma	Bérénice	0,285	Phèdre	0,279	Britannicus	0,278
Essex	Phèdre	0,291	Iphigénie	0,289	Bérénice	0,284
Maximian	Bérénice	0,295	Phèdre	0,285	Iphigénie	0,282
Persée	Bérénice	0,282	Phèdre	0,277	Iphigénie	0,270
Pyrrhus	Phèdre	0,286	Bérénice	0,282	Iphigénie	0,275
Stilicon	Bérénice	0,298	Iphigénie	0,283	Phèdre	0,273
Moyennes		0.289		0.280		0.275

La méthode ayant fait ses preuves sur de très nombreux corpus - comme ceux des frères Corneille ou des pièces présentées par J. Racine - elle est appliquée aux tragédies présentées par J. de La Chapelle et J.-G. Campistron qui sont mélangées aux œuvres présentées par les trois précédents.

III. … J. DE LA CHAPELLE ET J.-G. CAMPISTRON

Le corpus comporte maintenant 40 pièces dont les titres et les auteurs ne sont pas pris en compte par l'algorithme. Mais l'opérateur s'attend de nouveau à ce que l'automate retrouve cinq groupes puisque ces œuvres ont été présentées par cinq "auteurs" différents et que l'histoire littéraire affirme qu'il s'agit d'écrivains différents.

Trois auteurs

Comme l'algorithme de classification hiérarchique, l'automate isole trois groupes et non pas cinq.

Il retrouve les groupes 1 (P. Corneille) et 2 (T. Corneille) sans changement par rapport à ce qui vient d'être présenté. Autrement dit, il distingue bien ces deux écrivains de J. de La Chapelle et de J.-G. Campistron. Chacun des groupes 1 et 2 est homogène (les six premiers voisins de chacun des textes appartiennent bien au même groupe). Il n'y a aucun mélange.

En revanche, l'automate ne fait qu'un seul groupe des trois autres (tableau 5), c'est-à-dire qu'il confirme que les tragédies présentées par J. Racine, J. de La Chapelle et J.-G. Campistron entre 1666 et 1693 ont été composées par un seul écrivain. Ce n'est donc pas par hasard que la classification hiérarchique ascendante avait également mélangé ces oeuvres.

Dans ce troisième groupe, tous les textes atteignent l'indice maximal d'appartenance à un même groupe, sauf *Phèdre* (97%) et *Juba* (88%). La classification est donc quasiment parfaite.

On remarque également que les moyennes en dernière ligne du tableau 5 sont inférieures à celles du tableau 3 ci-dessus (obtenues sur les seules tragédies parues sous le nom de J. Racine). Autrement dit, paradoxalement, l'introduction des pièces présentées par J. de La Chapelle et J.-G. Campistron augmente l'homogénéité de ce groupe alors que l'histoire littéraire – et la chronologie - laissait attendre le contraire.

Les lignes concernant *Mithridate* et *Iphigénie* expliquent pourquoi ces deux pièces figurent dans le groupe que les pièces présentées par J.-G. Campistron - classifiées avant elles (*Arminius*, *Virginie*, *Alcibiade*, *Andronic*, *Aétius*) – où elles entraînent avec elles les autres tragédies présentées par J. Racine. *Mithridate* et *Iphigénie* sont des voisins réciproques, de telle sorte que l'automate se reporte à leurs deuxièmes voisins : *Arminius* (J.-G. Campistron) puis *Téléphonte* (J. de La Chapelle). Cette dernière pièce est rattachée au même groupe via son premier voisin : *Andronic* (présentée par J.-G. Campistron). Les trois "auteurs" n'en font qu'un ! Et cette unité est vérifiée jusqu'à l'ultime voisin pour les 21 textes.

Tableau 5. Groupe 3 : J. Racine, J. de La Chapelle et J.-G. Campistron. Classification sur les six plus proches voisins, classement par ordre d'agrégation au groupe.

Texte	1e voisin	1e distance	2e voisin	2e distance	3e voisin	3e distance	4e voisin	4e distance	5e voisin	5e distance	6e voisin	6e distance
Arminius	Virginie	0,184	Alcibiade	0,187	Andronic	0,188	Pompeia	0,196	Phocion	0,197	Aétius	0,200
Virginie	Arminius	0,184	Alcibiade	0,195	Andronic	0,197	Phocion	0,201	Pompéia	0,203	**Mithridate**	**0,209**
Alcibiade	Arminius	0,187	Andronic	0,191	Aétius	0,191	Phocion	0,194	Virginie	0,195	**Zaïde**	**0,205**
Andronic	Arminius	0,188	Alcibiade	0,191	Aétius	0,194	Tiridate	0,195	Virginie	0,197	**Téléphonte**	**0,199**
Aétius	Alcibiade	0,191	Andronic	0,194	Arminius	0,200	Pompéia	0,201	Phocion	0,209	Tiridate	0,209
Mithridate	Iphigénie	0,193	**Arminius**	**0,201**	**Andronic**	**0,203**	Bajazet	0,204	**Téléphonte**	**0,205**	Bérénice	0,206
Iphigénie	Mithridate	0,193	**Téléphonte**	**0,207**	**Arminius**	**0,211**	**Virginie**	**0,212**	**Alcibiade**	**0,216**	Phèdre	0,216
Phocion	Alcibiade	0,194	Arminius	0,197	Virginie	0,201	Andronic	0,203	Tiridate	0,206	Pompéia	0,207
Tiridate	Andronic	0,195	Pompéia	0,203	Phocion	0,206	Alcibiade	0,209	Aétius	0,209	Virginie	0,210
Pompéia	Arminius	0,196	Andronic	0,200	Aétius	0,201	Adrien	0,202	Virginie	0,203	Tiridate	0,203
Téléphonte	**Andronic**	**0,199**	**Mithridate**	**0,205**	**Iphigénie**	**0,207**	**Virginie**	**0,211**	**Arminius**	**0,211**	Zaïde	0,215
Adrien	Pompéia	0,202	Arminius	0,208	Aétius	0,211	Phocion	0,213	Alcibiade	0,217	Andronic	0,218
Bajazet	Andromaque	0,202	Mithridate	0,204	Britannicus	0,206	Bérénice	0,220	**Andronic**	**0,227**	**Arminius**	**0,230**
Andromaque	Bajazet	0,202	Mithridate	0,208	Britannicus	0,214	Iphigénie	0,222	Bérénice	0,227	**Andronic**	***0,233***
Zaïde	**Alcibiade**	**0,205**	**Virginie**	**0,212**	**Arminius**	**0,213**	Téléphonte	0,215	**Aétius**	**0,218**	**Andronic**	**0,218**
Tachmas	Aétius	0,206	Alcibiade	0.218	**Zaïde**	**0,219**	Tiridate	0,219	Virginie	0,221	**Bérénice**	**0,223**
Britannicus	Bajazet	0,206	Bérénice	0,209	Andromaque	0,214	Mithridate	0,218	Iphigénie	0,222	**Andronic**	**0,229**
Bérénice	Mithridate	0,206	Britannicus	0,209	Bajazet	0,220	**Andronic**	**0,226**	Iphigénie	0,226	Andromaque	0,227
Phèdre	Iphigénie	0,216	Mithridate	0,224	**Phocion**	**0,227**	**Alcibiade**	**0,230**	**Arminius**	**0,232**	***Virginie***	***0,235***
Juba	Arminius	0,218	Alcibiade	0,228	Aétius	0,231	*Adrien*	*0,236*	*Andronic*	*0,237*	*Pompéia*	*0,237*
Cléopâtre	Téléphonte	0,220	**Arminius**	**0,221**	**Mithridate**	**0,221**	**Alcibiade**	**0,223**	Zaïde	0,227	Virginie	0,228
moyennes		0,199		0,205		0,208		0,211		0,214		0,217

* en gras les anomalies (par rapport à l'hypothèse de trois écrivains distincts), en italiques les distances supérieures au seuil (0.233).
Sauf *Juba* et *Phèdre*, toutes les pièces atteignent le score maximum.

J. de La Chapelle à la charnière

Une objection pourrait être faite : pourquoi considérer les six plus proches voisins des pièces parues sous le nom de J. de La Chapelle alors qu'il n'en a produit que trois ? Ne faudrait-il pas, dans leur cas, limiter l'examen au deux premiers voisins ? Rappelons que les pièces sont anonymées et que l'automate a été programmé pour examiner les six voisins tant que leurs distances mutuelles sont inférieures à un certain seuil à partir duquel on ne peut plus affirmer qu'il s'agit d'un voisinage significatif (ici : 0.233). Si J. de La Chapelle était un écrivain – comme les deux frères Corneille – les trois pièces parues sous son nom devraient être leurs plus proches voisins réciproques et les suivantes devraient présenter des distances supérieures au seuil. Le tableau 6 ci-dessous permet de voir que la situation est bien différente.

Tableau 6. Plus proches voisins des trois tragédies présentées par J. de La Chapelle confrontées à celles de J.-G. Campistron et J. Racine (distances entre parenthèses)

La Chapelle	Premier voisin	Deuxième voisin	Troisième voisin
Zaïde (1681)	Campistron Alcibiade (0,205)	Campistron Virginie (0,212)	Campistron Arminius (0,213)
Téléphonte (1682)	Campistron Andronic (0,199)	Racine Mithridate (0,205)	Racine Iphigénie (0,207)
Cléopâtre (1683)	Campistron Arminius (0,220)	**La Chapelle Téléphonte (0,221)**	Racine Mithridate (0,221)

* NB : la troisième décimale n'est pas significative : elle indique comment arrondir la seconde

Il est construit comme les précédents tableaux mais il ne présente que les trois textes les plus proches de chacune des trois tragédies présentées par J. de La Chapelle en rappelant les noms sous lesquels ces pièces sont parues. Si J. de La Chapelle les avaient composées, sur chaque ligne, le premier et deuxième plus proches voisins devraient être ses deux autres tragédies et le troisième voisin se situer à une distance supérieure au seuil. En gras, le seul cas (sur neuf) qui corresponde à cette attente. Au total, sept pièces différentes apparaissent dans ce tableau de neuf cases – outre celle présentée par J. de La Chapelle, il s'agit des quatre premières pièces parues sous le nom de J.-G. Campistron - *Virginie*, *Arminius* (2 fois) *Andronic* et *Alcibiade* - et des deux pièces les plus centrales du noyau central des œuvres présentées par J. Racine : *Mithridate* (2 fois) et *Iphigénie*. Etant donné les faibles distances, ces neuf pièces sont du même écrivain… Par transitivité, puisque toutes les pièces présentées par J.-G. Campistron sont bien de la même main, il n'y a qu'un seul écrivain pour l'ensemble.

Le graphique le plus adapté à cette méthode est la classification arborée[1]. Elle présente un double avantage : elle représente, le mieux possible, toutes les distances individuelles et la qualité de cette représentation peut être mesurée[2]. Elle n'est pas reproduite ici pour ne pas lasser le lecteur. Les conclusions sont rigoureusement les mêmes : les tragédies présentées sous les trois noms de J. Racine – d'*Andromaque* à *Phèdre* - J. de La Chapelle et J. G. Campistron sont regroupées ensemble. La qualité de ce regroupement est la plus haute possible.

[1] Xuan Luong. *Méthodes d'analyse arborée. Algorithmes, applications*. Thèse pour le doctorat ès sciences. Université de Paris V, 1988. Cyril Labbé & Dominique Labbé. A Tool for Literary Studies. Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*. 2006, 21-3, p. 311-326.

[2] Cyril Labbé & Dominique Labbé. Peut-on se fier aux arbres ? In Serge Heiden et Bénédicte Pincemin (Eds). *9e Journées internationales d'analyse statistique des données textuelles (Lyon, 12-14 mars 2008)*. Lyon : Presses universitaires de Lyon, 2008, volume 2, p. 635-645.

Conclusions de la deuxième partie

Les méthodes de classification s'appliquent aux grandes collections de textes. Elles procèdent de manière objective, contrôlable, sans "apprentissage" et sans intervention humaine. La procédure exposée dans le dernier chapitre ne vise pas à l'exhaustivité mais à la fiabilité. Il ne s'agit pas de classer tous les textes, mais de le faire avec un risque d'erreur négligeable et de permettre de comprendre les problèmes posés par les individus sécants ou inclassables.

L'attribution d'auteur n'est qu'une application particulière de ces techniques.

En suivant les procédures décrites dans ces deux chapitres, il devient possible d'identifier celui qui a écrit un texte anonyme ou d'origine douteuse. Il suffit de disposer de textes contemporains de la plume potentielle, dans le même genre, sur des thèmes pas trop éloignés.

Dans le cas d'une attribution en aveugle (textes anonymés), l'analyste peut prendre cette décision grâce aux informations suivantes : textes contemporains, écrits dans un même genre et sur des thèmes proches. Dans ce cas simple, où l'écrivain est l'unique facteur déterminant, la classification non-hiérarchique permet de prendre une décision en y associant un degré d'appartenance du (ou des) texte(s) au groupe de rattachement.

Ainsi, en utilisant un seuil très restrictif (0.233) :

- l'algorithme sépare, en trois groupes, sans erreur ni incertitude, les tragédies présentées par P. Corneille, T. Corneille et J. Racine ;

- les pièces présentées par J. Racine, J. de La Chapelle et J.-G. Campistron sont attribuées à un seul et même écrivain avec des indices d'appartenance d'environ 99%, sauf *Phèdre* (94%) et *Juba* (88%).

Comme le montrera le dernier chapitre, la distance seuil raisonnable est 0.25 (à ce niveau, les tests évoqués en première partie donnent $\alpha = 5\%$). Avec ce seuil, tous les textes de J. Racine, J. de La Chapelle et J.-G. Campistron atteignent l'indice maximal d'appartenance à un même groupe. On en conclut qu'ils ont tous été composés par le même écrivain.

De multiples expériences comme celle-ci – organisées selon des procédures rigoureuses – ont confirmé la fiabilité de la méthode et la validité de l'échelle de la distance (présentée dans le dernier chapitre).

Ces expériences ont beaucoup d'autres intérêts. Par exemple, du point de vue de l'histoire littéraire, elles mettent en lumière l'influence de P. Corneille sur certaines pièces présentées par J. Racine ou les influences mutuelles entre certaines pièces du cadet des frères Corneille et de J. Racine.

TROISIEME PARTIE
COMPLEMENTS POUR UNE ATTRIBUTION D'AUTEUR

L'examen des distances intertextuelles et les classifications aboutissent à deux conclusions.

Premièrement, les tragédies présentées par J. de La Chapelle puis par J.-G. Campistron, entre 1680 et 1693 – ainsi que les manuscrits d'*Aétius*, *Juba* et *Tachmas* que nous publions -, sortent de la même main que les pièces présentées par J. Racine entre 1666 (*Andromaque*) et 1677 (*Phèdre*).

Deuxièmement, l'écrivain n'est pas le seul facteur déterminant la distance entre textes écrits dans un même genre. Leurs thèmes et l'époque de leur création jouent aussi un rôle qui peut compliquer l'attribution. En particulier, lorsque l'œuvre d'un écrivain s'étend sur un grand nombre d'années, le temps peut alors peser assez lourd, voire brouiller une question d'attribution.

Les pièces présentées par J. de La Chapelle et les premières par J.-G. Campistron sont bien contemporaines (1681 – 1685), ce qui neutralise l'influence du temps, mais *Andromaque* est de 1667, *Mithridate* de 1672 et *Iphigénie* de 1674, de telle sorte que leurs distances d'avec *Téléphonte* (1682), *Cléopâtre* et *Virginie* (1683), *Arminius* (1684), *Andronic* 1685, etc. peuvent sembler "trop" faibles par rapport à ce que laisserait attendre la tendance observée sur le seul corpus J. Racine. C'est surtout le cas pour les pièces présentées par J.-G. Campistron, ce qui renforce l'attribution mais laisse aussi supposer que les dates de composition ne seraient peut-être pas contemporaines de leur présentation au public.

Nous allons donc présenter une méthode permettant d'isoler assez précisément l'influence du temps sur les distances entre textes produits par les mêmes écrivains puis par des écrivains différents (Chapitre VII).

Ces expériences ont également permis de détecter d'autres indicateurs qui peuvent aider une attribution d'auteur mais aussi révéler les singularités d'un écrivain, notamment son style (Chapitre VIII).

CHAPITRE VII.
LE TEMPS

La plupart des écrivains changent de vocabulaire au cours de leur vie créatrice. De plus, la langue est un organisme vivant dont le composant sémantique (le "lexique") évolue constamment et ce flux emporte les écrivains sans qu'ils en soient forcément conscients.

Pour mesurer le poids de la dimension temporelle dans la distance entre textes, il faut neutraliser les trois autres variables (genre, écrivain, thème). Pour cela, il est nécessaire de disposer de textes d'un même écrivain, dans un même genre, portant à peu près sur les mêmes thèmes mais écrits à des époques différentes. Parmi les corpus de ce genre, l'un des plus intéressants est la correspondance de V. Hugo, pendant son exil dans les Iles anglo-normandes (de 1853 à 1870)[1]. C'était son seul lien avec la France et, d'un bout à l'autre, cette correspondance a été dominée par un certain nombre de thèmes et de considérations pratiques qui reviennent régulièrement. 1 126 lettres ont été conservées, envoyées à ses amis, sa famille, ses éditeurs, des journalistes et des écrivains... Dans ce corpus, la distance minimale, entre deux lettres – à un même destinataire et sur des thèmes comparables mais non contemporaines -, augmente en moyenne de 1.4% par an, soit environ 15% en dix ans. La même expérience - répétée sur d'autres écrivains et sur des romans, des pièces de théâtre, des poésies – montre que, plus deux textes d'un même écrivain sont éloignés dans le temps, plus sera forte la distance entre eux. Pour la plupart des écrivains, et selon les genres, cet accroissement s'inscrit dans un intervalle de 0.5 à 2% par année (soit un doublement en 35 ans).

Ces proportions se vérifient-elles dans le théâtre du XVIIe ? Quelles en sont les conséquences sur l'attribution des pièces présentées par J. Racine, J. de La Chapelle et J.-G. Campistron ?

[1] Cyril Labbé & Dominique Labbé. Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. In Banks David (ed.). *Le texte épistolaire du XVIIe siècle à nos jours*. Paris : L'Harmattan, 2013, p. 53-85. Le détail du calcul sur V. Hugo est présenté dans Cyril Labbé & Dominique 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 images*. (http://images.math.cnrs.fr/La-classification-des-textes.html). 28 mars 2011.

I. L'EFFET DU TEMPS DANS LES ŒUVRES D'UN ÉCRIVAIN

La chronologie est une dimension essentielle pour l'attribution d'auteur, du moins dès que les textes discutés ne sont pas exactement contemporains. En premier lieu, quand on détecte des proximités anormales entre des œuvres – comme celles relevées entre J. La Chapelle, J.-G. Campistron et J. Racine -, il s'agit d'abord d'établir l'antériorité de l'un des écrivains en se référant aux dates de création des textes et à la biographie des supposés auteurs. De plus, il faut tenir compte d'un éventuel décalage temporel entre ces créations, pour en tenir compte dans les calculs. Enfin, il faut s'assurer de l'importance du temps dans les œuvres, indiscutées, des écrivains concernés.

Chronologie et attribution d'auteur

Le tableau 1 ci-dessous rappelle quelques données de base concernant les trois "auteurs" que tous les indices désignent comme n'en faisant qu'un et qui sont donc candidats pour une paternité unique des œuvres parues sous les trois noms.

Tableau 1 Données biographiques concernant J. Racine, J. La Chapelle et J.-G. Campistron

	Naissance – décès	Age en 1667 (*Andromaque*)	Arrivée à Paris
J. Racine	1639 -1699	28 ans	1660
J. de La Chapelle	1651-1723	16 ans	1679 ou 1680
J.-G. Campistron	1656-1723	11 ans	1682

Puisque toutes les pièces d'*Andromaque* (1667) à *Juba* (composition probable vers 1695) sortent de la même plume, deux des trois candidats se trouvent éliminés puisque trop jeunes pour avoir composé les premières des pièces présentées par J. Racine et d'ailleurs absents de Paris à ces dates. Il n'en reste donc plus qu'un.

Ceci rappelle également que la création de ces pièces s'est étalée sur plus d'un quart de siècle. Quelle conséquence cela peut-il avoir sur une attribution d'auteur ?

Les pièces présentées par J. Racine sont rangées par ordre chronologique ; on calcule l'espace temporel les séparant et on y associe les distances correspondantes. Naturellement, la date de création et la date d'écriture peuvent être plus ou moins éloignées ; aussi utilise-t-on l'année civile et non le mois, sauf dans les cas limites où les créations coïncident avec le début ou la fin d'une année[1]. Par exemple, le tableau 2 donne le calcul pour les pièces dont la création est séparée d'un an.

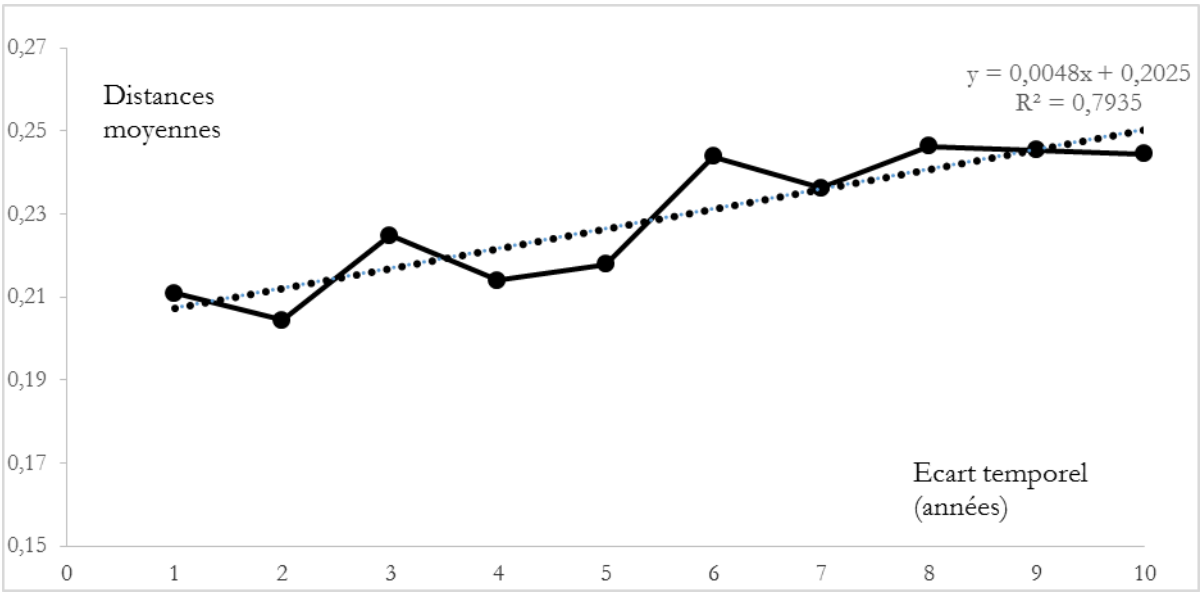
Tableau 2. Calcul de la distance moyenne entre les pièces séparées d'une année (corpus J. Racine)

Ainée	Année création	Cadette	Année de création	Distance
Britannicus	1669	Bérénice	1670	0,209
Bérénice	1670	Bajazet	1672	0,219
Bajazet	1672	Mithridate	1672	0.204
Mithridate	1672	Iphigénie	1674	0,193
Moyenne				0.207

Pour les pièces séparées d'environ une année, la distance moyenne est de 0.207) (légèrement inférieure à 0,21). On répète le même calcul pour chaque écart temporel et on reporte le résultat sur un graphique (tableau 3). On notera que sur, ce graphique, l'origine de l'axe vertical a été déplacée de 0.0 à 0.15 afin de grossir le phénomène. Dans la réalité les accidents de la courbe sont moins prononcés et la pente de la droite plus horizontale.

[1] Par exemple, *Bérénice* (21 novembre 1670) et *Bajazet* (1er janvier 1672) sont séparés par un an et non deux en utilisant l'année civile. De même, un an sépare *Bajazet* (1er janvier 1672) de *Mithridate* (23 décembre 1672), etc. Malgré ces précautions, il est évident que les écarts temporels ne sont pas très précis.

Tableau 3. Accroissement de la distance en fonction de l'écart temporel séparant les pièces (tragédies de J. Racine d'Andromaque à Phèdre, valeurs observées et ajustées)



Sur l'axe vertical sont portées les distances moyennes pour les empans temporels mesurés sur l'axe horizontal. Le trait gras relie les points observés (distances moyennes pour chaque écart temporel). A part de légères encoches, la tendance à l'augmentation des distances moyennes semble relativement régulière et elle présente un profil linéaire (figuré par la ligne pointillée sur le graphique). Autrement dit, les distances augmenteraient en fonction du temps et les petites ruptures de pente dans la courbe seraient l'effet de perturbations – notamment le léger flou concernant les écarts chronologiques - que l'on propose de négliger.

Pour vérifier cette hypothèse, on procède en trois temps.

Premièrement, le calcul de la droite dite d'ajustement (trait pointillé sur le graphique). Cette droite a pour principale caractéristique de passer au plus près des observations. Elle représente le profil du phénomène en l'absence des petites perturbations que l'on néglige pour l'instant. C'est pourquoi on parle "d'ajustement" - ou de "régression" - des distances (en fonction du temps).

Cette droite est calculée de la manière suivante (sous réserve de disposer d'un nombre suffisant d'observations). On note :

t le rang des intervalles (t variant ici de 1 à $T = 10$ années)

$\bar{t}$ le "milieu" de la série temporelle (5 années)

$\overline{D_t}$ la distance moyenne des pièces séparées par t années

$\overline{D}$ la moyenne de ces moyennes :

$$\overline{D} = \frac{\sum_{t=1}^{T} \overline{D_t}}{T} = .2204$$

Le point central de la droite d'ajustement a pour coordonnées ($\bar{t}$ et $\overline{D}$).

A chaque valeur observée ($\overline{D_t}$) correspond un point théorique (D'_t) de même abscisse *t* et dont l'ordonnée donne la valeur que l'on aurait observée s'il n'y avait pas eu les petites perturbations. En joignant ces points théoriques, on obtient la droite d'ajustement (pointillés sur le graphe). Ces valeurs théoriques sont données par l'équation suivante :

$D'_t = at + b$

Le calcul fournit quatre informations :

- L'origine de la droite d'ajustement (deuxième partie de l'équation inscrite sur le tableau : b = 0,2025) ou valeur théorique de la variable pour l'année zéro (ici deux textes exactement contemporains) que l'on peut considérer comme une estimation de *Dmin* pour le cas de J. Racine.

- La pente de la droite d'ajustement (*a*) par rapport à l'horizontale (l'axe des abscisses), calculée pour minimiser la somme des carrés des écarts entre les valeurs observées (D_t) et les valeurs théoriques correspondantes (D'_t) en utilisant la formule suivante :

$$a = \frac{\sum_{t=1}^{T}\left(t - \bar{t}\right)\left(D_t - \overline{D}\right)}{\sum_{t=1}^{T}\left(t - \bar{t}\right)^2} = 0.0048$$

Ce coefficient varie entre 0 (droite horizontale) et ± ∞ (droite proche de la verticale). Lorsqu'il est égal à 1, la droite est parallèle à la première diagonale ; et parallèle à la seconde diagonale quand a = -1.

Cette pente permet de mesurer directement l'influence de la variable temps : lorsque l'intervalle entre pièces présentées par J. Racine augmente d'un an, la distance moyenne augmente de 48 mots (pour 10 000). C'est peu mais sur 10 ans, cela donne une croissance de 480 mots par rapport à la distance d'origine (année 0). Par exemple, on peut prévoir que deux tragédies présentées par J. Racine à dix ans d'intervalle seront séparées d'une distance de 0.2025 + 0,0480 = 2505, soit une augmentation de +24% par rapport à deux textes contemporains et un rythme annuel de + 2,2% qui est tout à fait exceptionnel.

- R, le coefficient de liaison de la distance au temps (0,89 : racine carrée de 0.79). Ce coefficient est une application particulière du coefficient de corrélation (dit Bravais-Pearson) aux séries temporelles :

$$R^2 = \frac{\textit{somme des carrés de la régression}}{\textit{racine carrée du produit des variances}} = \frac{\sum_{t=1}^{T}\left(D_t - \overline{D}\right)\left(D'_t - \overline{D}\right)}{\sqrt{\sum_{t=1}^{T}\left(D_t - \overline{D}\right)^2 \sum_{t=1}^{T}\left(D'_t - \overline{D}\right)^2}} = 0.79$$

Si le temps était seul en cause, ce coefficient serait égal à 1. A l'inverse, si le temps n'avait aucune relation avec la distance, ce coefficient serait égal à 0.

L'appréciation concernant la puissance de la liaison dépend du nombre de mesures sur lesquelles s'appuie le calcul (qui détermine le nombre de "degrés de liberté"). Si l'on s'en tient au nombre de variables (les lignes du tableau de calcul : 10 années), cela donne 8 ou 9 degrés de liberté (ddl) selon que l'on considère qu'il y a une année "contrainte" ou deux. Si l'on calcule les ddl sur le nombre de distances entrant dans le calcul (21 distances), cela donne 19 ou 20 ddl. Les tables établies par Fisher et Yates (extraits ci-dessous), indiquent que – avec 8 ddl - l'on a moins de 1% de chances de se tromper en considérant que la tendance à l'accroissement des distances entre pièces de J. Racine est bien liée au nombre d'années qui les sépare. Ce qui signifie que deux pièces écrites par J. Racine sont d'autant plus différentes qu'elles sont éloignées dans le temps.

Tableau 4. Extraits de la table du coefficient de corrélation (valeur minimale de R pour un risque d'erreur (α) et un certain nombre de degrés de liberté (ddl)[1].

ddl α	0.05	0.01
5	0.7545	0.8745
10	0.5760	0.7079
15	0.4821	0.6055
20	0,4227	0,5368
25	0,3809	0,4869
30	0,3494	0,3932
35	0,3246	0,4182
40	0,3044	0,3932
45	0,2875	0,3721
50	0,2732	0,3541
60	0.2500	0.3248
70	0.2319	0.3017
100	0.1946	0.2540

- Le coefficient de détermination de la distance par le temps (R^2 ou carré du coefficient de liaison ci-dessus). L'interprétation conventionnelle est la suivante : 79% de l'augmentation de la distance entre tragédies non contemporaines de J. Racine s'expliquent par leur éloignement temporel plus ou moins important. Là encore, la liaison est exceptionnelle. Outre le problème des dates de création ne correspondant pas à des années civiles, le résidu (un cinquième du phénomène) peut s'expliquer par trois raisons. Premièrement, l'utilisation des années civiles et non des mois et jours de création (les écarts sont des approximations). Deuxièmement, le décalage plus ou moins long entre l'époque de la rédaction et la première représentation. Enfin, les modifications apportées aux éditions successives, la version de référence étant la dernière

[1] Fisher & Yates. *Statistical Tables for Biological, Agricultural and Medical Research* (1949).

parue avant la mort de J. Racine et non celle qui a suivi immédiatement la première série de représentation des pièces.

Sous réserve de ces remarques, l'équation de la droite d'ajustement permet de calculer la distance attendue en fonction de la date de création de deux textes par la même plume. Par exemple, pour des pièces créées à 14 ans d'écart, l'équation donne une distance théorique de 0,0048*14+0,2025 = 0,27. La distance entre *Phèdre* (1677) et *Athalie* (1691) est de 0,275. L'ajustement semble donc parfait, mais comme le montrera le dernier chapitre de cette note, des distances aussi élevées ne permettent plus de conclure à un écrivain unique. Autrement dit, des décalages chronologiques importants peuvent mettre en échec une attribution d'auteur, surtout si les thèmes et le vocabulaire se renouvellent assez rapidement comme c'est le cas dans les tragédies présentées par J. Racine entre 1667 et 1677. A l'inverse, dans un cas pareil, de faibles distances cumulées avec un écart temporel important renforcent l'attribution. C'est ce qui se produit avec les pièces présentées sous le nom de J. de La Chapelle et de J.-G. Campistron.

Pour les pièces présentées par J. Racine, on conclut que, en tendance moyenne, chaque année se traduit par une augmentation des distances entre œuvres supérieure à 2% (doublement en 35 ans).

Cependant, le fait qu'une variable évolue régulièrement au cours du temps ne signifie pas que le temps est l'explication de cette variation. Le temps peut être simplement la condition permettant à d'autres variables de s'exprimer. Par exemple, chez certains écrivains, la découverte de nouveaux thèmes ou des changements de "registre"… Plus l'écart chronologique est important, plus il y a de chances pour que le thème change, comme cela se produit d'ailleurs chez J. Racine, spécialement entre 1677 (*Phèdre*) et 1688 (*Esther*).

Le temps dans les œuvres des frères Corneille

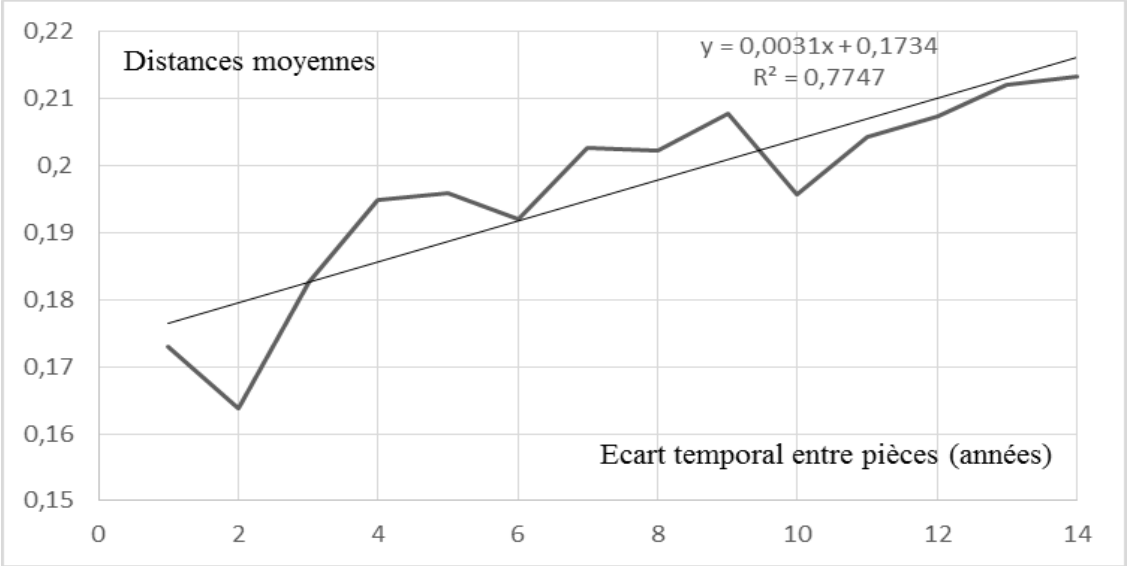
Dans les pièces de la maturité de P. Corneille, on observe une tendance à l'augmentation de la distance de 12% en 10 ans[1]. Autrement dit, une distance de 0.19 entre deux pièces contemporaines de P. Corneille équivaut à une distance supérieure à 0.21 entre deux pièces de cet écrivain mais séparées par 10 ans et de 0.24 pour un intervalle de 20 ans. Dans la dernière partie de sa vie créatrice, P. Corneille fait preuve d'une stabilité plus grande que la majorité des écrivains, notamment J. Racine.

On constate la même tendance chez T. Corneille (tableau 5). Par rapport aux pièces de J. Racine, la droite d'ajustement a une origine beaucoup plus basse (0,17 au lieu de 0,20), une pente

[1] Calcul présenté dans : Cyril Labbé & Dominique Labbé. La classification des textes. *Art. Cit.* Nous n'en reproduisons pas à nouveau le détail.

moins forte mais un ajustement au moins aussi bon (coefficient de corrélation égal à 0,88 pour 12 ou 34 degrés de liberté selon les interprétations choisies).

Tableau 5. Accroissement de la distance en fonction de l'écart temporel séparant les pièces (œuvres de T. Corneille, valeurs observées et ajustées)



En conséquence, on a moins de 1% de chances de se tromper en affirmant que plus les dates de création des pièces de T. Corneille sont éloignées, plus ces pièces sont séparées par des distances importantes.

II. L'EFFET DU TEMPS DANS LA COMPARAISON ENTRE PLUSIEURS ÉCRIVAINS

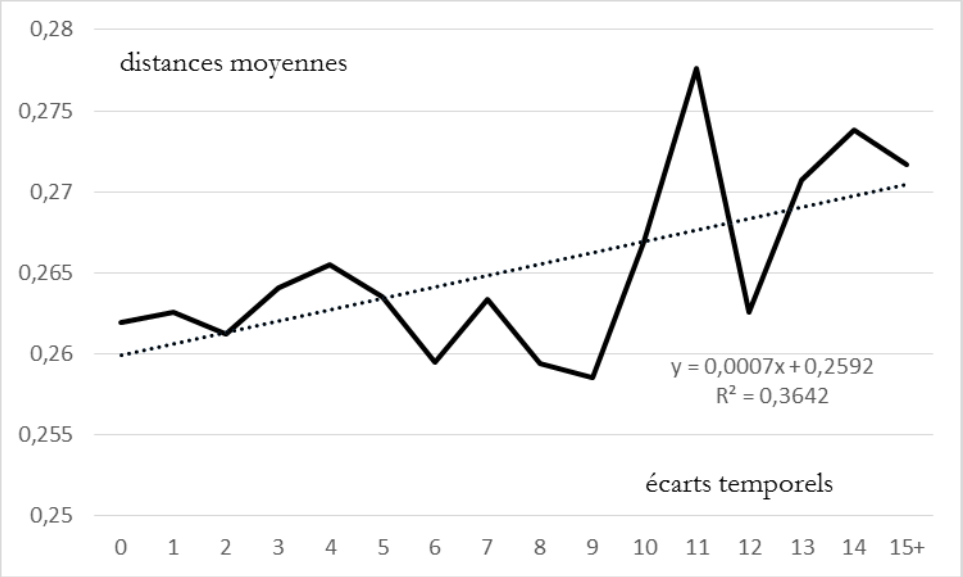
Les distances intertextuelles peuvent rarement être calculées entre œuvres exactement contemporaines. La comparaison entre J. Racine, P. et T. Corneille porte sur une période relativement limitée : 18 ans séparent *Œdipe* (1659) de *Phèdre* (1677) et la majorité des couples sont séparés par moins de 10 ans. Cela permet de vérifier si ces écarts temporels sont bien sans conséquences dans la comparaison entre deux écrivains différents.

Le temps dans la comparaison entre les pièces des frères Corneille et celles présentées par J. Racine

Pour la comparaison entre les pièces présentées par P. Corneille et J. Racine, le calcul fait apparaître une légère tendance à l'accroissement des distances avec le nombre d'années séparant les pièces comparées (tableau 6) mais la pente de la droite est très faible (sur le tableau 4 l'origine est déplacée à 0.25) et, même si le coefficient de corrélation est significatif (avec 68 degrés de

liberté mais pas avec 13 ddl et un seuil à 1%), à peine plus du tiers des fluctuations de la distance peuvent s'expliquer par l'écart temporel entre les pièces.

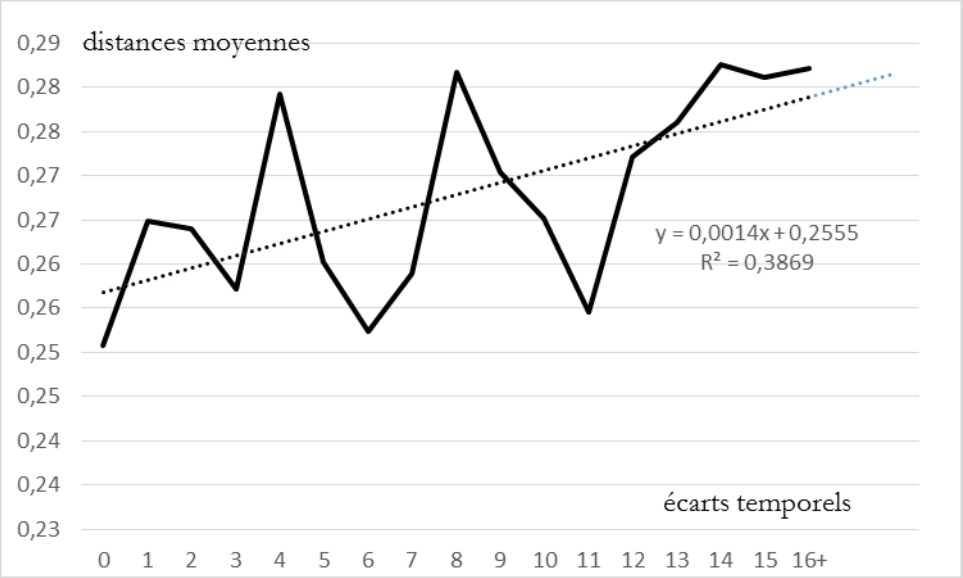
Tableau 6. Variation de la distance entre les tragédies présentées par J. Racine et P. Corneille en fonction de l'écart temporel séparant ces pièces (valeurs observées et ajustées)



L'ajustement est médiocre notamment parce qu'un creux important apparaît au milieu de la série et concerne quelques pièces de P. Corneille – spécialement la *Toison d'or* (1660), *Sertorius* (1662), *Sophonisbe* (1663) et *Othon* (1664) – proches de certaines pièces présentées par J. Racine - notamment *Andromaque* (1667), *Britannicus* (1669) et *Mithridate* (1672). Si l'on voulait creuser d'avantage l'influence "cornélienne" sur ces pièces, il faudrait donc refaire l'expérience en tenant compte de ce décalage. Mais au fond, c'est une évidence que l'influence possible d'un écrivain sur un autre se manifeste nécessairement avec un certain retard temporel.

Le même calcul appliqué à la comparaison entre T. Corneille et J. Racine fait apparaître une pente de la droite d'ajustement légèrement plus prononcée, avec une liaison faiblement significative, mais aussi une détermination de la distance par le temps également médiocre (tableau 7).

Tableau 7. Variation de la distance entre les tragédies présentées par J. Racine et T. Corneille en fonction de l'écart temporel séparant ces pièces (valeurs observées et ajustées)



Deux conclusions s'imposent.

1. Dans la comparaison entre les frères Corneille et J. Racine, le facteur temps – dix-huit ans au maximum - a donc un poids minime.

2. L'expérience démontre que la distance intertextuelle attribue sans problème des textes à peu près contemporains (ou des œuvres séparées par un laps de temps pas trop grand), écrits dans un même genre, sur des thèmes proches, pour les mêmes comédiens et le même.

Est-ce le cas pour J. Racine, J. de la Chapelle et J.-G. Campistron alors que, cette fois, les œuvres sont décalées chronologiquement ?

Le temps dans la comparaison entre les pièces présentées par J. Racine, J. de la Chapelle et J.-G. Campistron

Dans cette étude, deux dimensions sont à prendre en compte :

- Entre *Andromaque* (1667) et *Phèdre* (1677), il s'écoule dix ans.

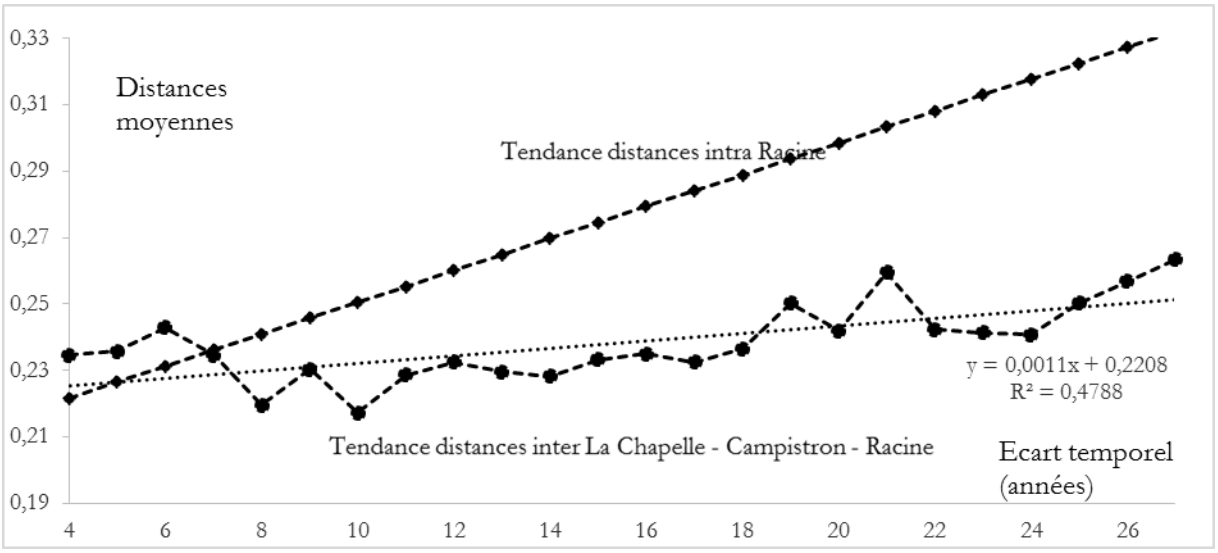
- La dernière pièce présentée par J. Racine *(Phèdre* (1677) et *Zaïde* (La Chapelle début 1681) sont séparées de 4 ans. Entre *Phèdre* (1677) et *Virginie* (*1683*), il s'écoule 6 ans. Ce sont les intervalles minimaux. L'intervalle le plus important sépare *Andromaque* (1667) d'*Aétius* (1693) : 26 ans.

Pour mesurer le poids de ce décalage temporel, on répète l'opération présentée au paragraphe précédent : calcul des écarts temporels séparant les pièces considérées par couples que l'on associe aux moyennes des distances correspondantes puis on réalise un ajustement linéaire de

ces données. L'opération est effectuée uniquement sur les couples formés par le rapprochement entre ces pièces et celles de J. Racine (tableau 8).

Ces écarts et les distances correspondantes sont portés sur le diagramme ci-dessous), avec le prolongement de la tendance propre au corpus Racine sur ce même laps de temps, telle qu'elle a été calculée dans la section précédente (pointillé gras).

Tableau 8. Accroissement de la distance en fonction du temps (distances entre les œuvres de J. Racine et celles présentées sous les noms La Chapelle et de Campistron)



Le trait gras pointillé est l'accroissement théorique des distances calculé d'après la tendance observée sur les pièces présentées par J. Racine (tableau 3). Si le même rythme de renouvellement, observé entre 1667 et 1677, s'était poursuivi jusqu'au bout, la distance entre *Aétius* et *Andromaque* aurait été de 0.33 alors qu'elle n'est que de 0.26. Toutes les observations sont inférieures à la tendance, sauf les trois premières (distances entre *Phèdre* et les pièces parues sous le nom de J. La Chapelle). En dehors de cette exception, les pièces présentées par J. de La Chapelle et par J.-G. Campistron sont "trop" proches de celles de J. Racine par rapport à ce que les écarts temporels laisseraient attendre.

La tendance à l'accroissement des distances en fonction du temps après 1677 est donc beaucoup plus lente qu'auparavant (quatre fois moins rapide). La pente de la droite d'ajustement est d'ailleurs presque nulle (0.0011). Mais la liaison demeure. Avec 24 degrés de liberté, le coefficient de corrélation (0.69) est même très significatif : on a moins de 1 chances sur 1000 de se tromper en acceptant une augmentation faible mais continue des distances entre les pièces présentées sous le nom de J. Racine et celles présentées par J. de La Chapelle puis par J. G. Campistron, mais le temps n'explique que la moitié des fluctuations de la variable, fluctuations au demeurant très faibles.

Plusieurs interprétations sont possibles (et cumulables) :

- à part sans doute *Adrien* (1690), les tragédies présentées sous le nom de J. de La Chapelle puis de J.-G. Campistron auraient, pour la plupart, été composées plusieurs années avant leur parution et n'auraient pas été présentées dans l'ordre de leur composition. Ceci vaut principalement pour les dernières pièces (*Tiridate*, *Pompéia*, *Aétius*) qui semblent nettement antérieures. *Juba* lui-même aurait été commencée bien avant 1695 ;

- les pièces qui ont assuré les plus grands succès avant 1677 (*Iphigénie*, *Andromaque*…) ont servi de modèles pour les pièces présentées sous les noms de J. de La Chapelle puis de J.-G. Campistron. Autrement dit, les premiers succès ont servi de matrice pour les pièces de l'ombre ;

- le renouvellement du vocabulaire et des thèmes est plus faible après 1677, ce qui est l'idée ci-dessus formulée différemment.

Enfin, ce constat renforce l'attribution à une même plume de toutes les tragédies présentées, entre 1666 et 1693, sous les noms de J. Racine, de J. de La Chapelle puis de J.-G. Campistron puisque les distances constatées sont inférieures aux valeurs attendues pour un auteur unique et étant donné le laps de temps qui séparent leurs dates de création.

D'autres dimensions peuvent caractériser un écrivain et aider à lui attribuer un texte d'origine inconnue ou douteuse. En ce qui concerne le théâtre en vers du XVIIe, la longueur et la structure de la phrase sont des indices stylométriques particulièrement intéressants.

CHAPITRE VIII.
PHRASE ET CHOIX STYLISTIQUES

D'autres indices peuvent conforter l'attribution d'auteur ou, à l'inverse, mettre en doute cette paternité lorsque les distances sont suffisamment élevées pour que deux hypothèses soient envisageables (plume de l'ombre, plagiat ou collaboration entre deux écrivains ou influence passagère de l'un sur l'autre). Outre les classifications déjà évoquées, il s'agit des combinaisons de mots les plus fréquents, du sens des principaux mots[1], des longueurs et des structures des phrases[2].

Ces dernières sont particulièrement intéressantes pour l'étude du théâtre du XVIIe et spécialement dans le cas des pièces présentées par J. Racine et ses deux épigones. Le raisonnement présenté en première partie leur sera appliqué.

I. PHRASE ET STYLISTIQUE

Comme pour la reconnaissance des mots, présentée dans le premier chapitre, la délimitation des phrases et leur analyse pose une série de problèmes qu'il convient de résoudre si l'on veut fonder l'analyse sur des données fiables. La méthode sera illustrée à l'aide des tragédies présentées par J. Racine.

Délimitation de la phrase et décompte

La stylistique s'est beaucoup penchée sur cette question[3] car la longueur et la construction des phrases sont des dimensions essentielles du style d'un écrivain. Les principales caractéristiques *théoriques* de la phrase françaises sont connues[4]. Mais, hormis quelques travaux

[1] Labbé Cyril & Labbé Dominique. How to measure the meanings of words? Amour in Corneille's work. *Language Resources Evaluation*. 39, p. 335-351 (article consultable en ligne sur les archives ouvertes du CNRS).

[2] Labbé Cyril & Labbé Dominique. Ce que disent leurs phrases. In Bolasco Sergio, Chiari Isabella, Giuliano Luca (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, 2010, Vol 1, p. 297-307 (communication consultable en ligne sur les archives ouvertes du CNRS).

[3] Molinié Georges (1986). *Eléments de stylistique française*. Paris : PUF, 1986.

[4] Voir notamment : Le Goffic Pierre. *Grammaire de la phrase française*. Paris : Hachette, (1999). Et pour une actualisation : Charolles Michel, Fournier Nathalie, Fuchs Catherine, Lefeuvre Florence (2007). *Parcours de la phrase: Mélanges offerts à Pierre Le Goffic*. Paris : Ed. Ophrys.

pionniers - comme ceux de F. Richaudeau[1], de J. Milly sur Proust[2] ou de R. Garette sur Racine[3] - les études *empiriques* sur de vastes corpus manquent. Cela s'explique notamment par le fait que les dépouillements sont manuels et que l'étude des phrases par informatique s'est heurtée jusqu'à maintenant à des obstacles sous-estimés[4].

Premièrement, la phrase est définie comme l'espace de texte compris entre deux ponctuations fortes. La longueur de la phrase est mesurée par le nombre de mots compris dans cet espace. Une ponctuation forte est l'un des signes suivants : '.' '…' ' ?' ' !', quand ils sont suivis d'un mot dont l'initiale est en majuscule. Si un nom propre suit un point, l'opérateur doit se substituer à l'automate et trancher entre deux possibilités : début d'une nouvelle phrase ou simple abréviation (par exemple "M. Racine", "etc."). Le respect de ces conventions est indispensable comme l'a montré la controverse autour de la longueur des phrases chez Proust[5].

Deuxièmement, pour étudier la construction de la phrase, il faut d'abord identifier chacun de ses composants (notamment les groupes verbaux et nominaux) et donc pour cela étiqueter sans erreur tous les mots qui la composent…

Troisièmement, comme pour la graphie des mots, il se pose un problème de standardisation. Au XVIIe siècle, la plupart des écrivains ont abandonné la ponctuation aux compositeurs[6]. C'est manifestement le cas pour les pièces publiées par J. Racine[7]. Aussi nous n'utilisons pas la ponctuation des éditions originales mais celle établie par J. Mesnard selon des conventions syntaxico-sémantiques qui ont également été utilisées pour l'édition moderne des œuvres de Molière et de P. Corneille (avec, pour ce qui concerne J. Racine quelques rectifications

---

[1] Richaudeau François. *Ce que révèlent leurs phrases*. Paris : Retz, 1988.

[2] Milly Jean. *La Phrase de Proust*. Paris : Larousse, 1975 (Réédition Paris : Champion, 1983). Et : Milly Jean. *La longueur des phrases dans "Combray"*. Paris-Genève : Champion-Slatkine, 1986.

[3] Garette Robert (1995). *La phrase de Racine. Etude stylistique et stylométrique*. Toulouse : Presses universitaires du Mirail.

[4] Pour les principes d'une stylométrie (statistique appliquée à l'étude du style) : Monière Denis & Labbé Dominique. Essai de stylistique quantitative. Duplessis, Bourassa et Lévesque. In Morin Annie et Sébillot Pascale (Eds). *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002)*. Rennes : IRISA-INRIA, 2002, vol. 2, p. 561-569. Labbé Cyril, Labbé Dominique & Monière Denis. Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest. *Canadian Journal of Political Science / Revue canadienne de science politique*. 41:1, mars 2008, p. 43-69. Textes consultables en ligne sur les archives ouvertes du CNRS.

[5] Milly Jean, *Op.cit.*, 1986, p. 165-167. En l'occurrence, le dépouillement manuel de J. Milly était juste et celui effectué par ordinateur gravement biaisé (pratiquement 10% de phrases en trop !)

[6] Riffaut Alain. *La ponctuation du théâtre imprimé au XVIIe siècle*. Genève : Droz, 2007. Cet ouvrage comporte à la fin les principaux principes applicables aux textes anciens quand leur ponctuation est manifestement lacunaire ou défectueuse. Ce sont ceux que nous suivons depuis le début de nos recherches dans les années 1990.

[7] Giraud Yves. Lire Racine, vraiment ? *Revue d'Histoire Littéraire de la France*. 2001-2, 101, p. 303-311.

mineures par R. Picard pour son édition de la Pléïade). Nous avons suivi les mêmes principes pour les transcriptions sur support électronique de toutes les pièces du XVIIe, notamment celles publiées par J. de La Chapelle, J.-G. Campistron et T. Corneille utilisées dans ce rapport de recherche.

La phrase dans les pièces présentées par J. Racine

Pour étudier la longueur des phrases, une mesure unique, par exemple la longueur moyenne, ne suffit pas. Une série de valeurs significatives - présentées dans le tableau 4 ci-dessous – permettent de comprendre la singularité du phénomène.

Tableau 4. Caractéristiques des phrases dans les tragédies publiées par J. Racine (en mots, classement par ordre chronologique) comparée à aux caractéristiques moyennes des tragédies de T. Corneille

Texte	Longeur (mots)	N Phrases	Ponctuations par phrase	Mode	Médiane	Moyenne	Médiale	Ecart type
Thébaide	13 813	740	2,07	16	16,5	19,7	22,8	12,0
Alexandre	13 864	726	1,99	9	16,3	19,1	25,4	13,6
Andromaque	15 076	1 054	1,49	9	10,2	14,3	17,9	10,8
Britannicus	15 387	1 014	1,69	9	10,5	15,2	19,4	12,5
Bérénice	13 242	981	1,50	9	9,1	13,5	17,8	11,9
Bajazet	15 297	1 069	1,53	8	9,8	14,3	18,7	12,1
Mithridate	15 091	946	1,77	9	11,1	16,0	22,4	13,2
Iphigénie	15 782	1 096	1,49	9	10,0	14,4	17,8	11,9
Phèdre	14 394	1 024	1,36	8	10,0	14,1	17,4	11,2
Esther	11 147	763	1,36	9	10,0	14,6	18,4	11,9
Athalie	15 492	1 078	1,37	8	9,6	14,4	18,3	12,2
Moyenne Racine	14 417	954	1,59	9	10,5	15,1	19,2	12,2
MoyenneT. Corneille	17 384	768	2,15	8	19,2	21,8	32,4	16,3

.

Les valeurs centrales (mode, médiane et moyenne) sont fortement décalées, alors que dans le cas des distances intertextuelles elles étaient très proches ou confondues. De plus, l'écart type est anormalement élevé (dans les pièces de J. Racine, 81% de la moyenne). Tout cela indique qu'il existe non pas une seule population mais plusieurs types de phrases.

Examinons d'abord chacun des paramètres.

Dans les tragédies publiées par J. Racine, les phrases les plus fréquentes (**mode**) comportent huit à neuf mots. Il en est de même chez la quasi-totalité des auteurs de tragédies en vers, contemporains de J. Racine, notamment chez T. Corneille. En effet, un vers alexandrin compte en moyenne 8 à 9 mots. Les écrivains cherchent logiquement à caler autant que possible

la fin de la phrase sur une fin de vers. Mais dans les pièces publiées par J. Racine, la plupart du temps la phrase couvre un seul vers alors que, chez les frères Corneille, c'est souvent deux voire plus, ce qui entraîne une construction plus complexe, le premier indice de cette complexité étant la nécessité d'utiliser plus de ponctuations internes à la phrase.

La **médiane** indique que, dans les pièces publiées par J. Racine, la moitié des phrases comportent 10 mots et moins, ce qui traduit une forte prédominance des phrases courtes (de un vers à un vers et demi). Chez T. Corneille, cette moitié n'est atteinte qu'à 18 mots (deux vers), soit 80% de plus que chez J. Racine.

La **moyenne** est fortement décalée par rapport au mode et à la médiane. Un tel décalage se produit lorsque le caractère (ici les mots) est réparti de manière très inégalitaire entre les individus (ici les phrases). Dans ce cas, la moyenne est tirée vers le haut par quelques individus riches (comme pour les revenus). C'est ce qui se passe dans le théâtre du XVIIe où figurent parfois des phrases très longues. Dans les pièces publiées par J. Racine, cette longueur moyenne est de 15,1 mots (presque deux vers). Chez T. Corneille, elle est de 22,7 mots (66% de plus que chez J. Racine, ce qui est considérable).

La **médiale** indique que, dans les tragédies présentées par J. Racine, la moitié du texte est occupé par des phrases de moins de 19 mots (deux vers environ) – ou encore : la moitié du temps, le spectateur entend des phrases courtes, dans une sorte d'accumulation haletante qui est assez caractéristique de son théâtre, dominé par les passions portées à leur paroxysme (voir plus bas). A l'opposé, cette valeur est de 32 mots dans les tragédies de T. Corneille, soit 68% de plus… Autrement dit, nous sommes en face de choix stylistiques différents.

Enfin, le tableau 4 met à nouveau en valeur la singularité des deux premières pièces présentées par J. Racine. Pour tous les paramètres, sauf les longueurs totales, les valeurs enregistrées sur *La Thébaïde* et *Alexandre* diffèrent grandement de celles observées sur les suivantes. De nouveau, la suite de cette analyse portera donc sur le noyau central allant d'*Andromaque* à *Iphigénie.*

La comparaison des différentes valeurs suggère donc des caractéristiques profondément différentes entre écrivains. Cependant, peut-on affirmer que les différences entre les longueurs des phrases dans les pièces publiées par J. Racine et par les frères Corneille ne sont pas dues au hasard, qu'elles sont vraiment différentes ?

II. COMPARAISON DES PHRASES

Les données présentées ci-dessus conduisent à se demander s'il est possible d'utiliser la phrase pour reconnaître un écrivain. La réponse est apportée en deux temps. Premièrement, il est proposé d'utiliser le raisonnement présenté en première partie de cette recherche. Deuxièmement, l'examen graphique complète cette analyse et permet de reconnaître le style propre de chaque écrivain.

Tests statistiques

Deux hypothèses sont examinées. La première est celle d'un écrivain unique avec des habitudes d'écriture stables (H_0). Naturellement, même dans ce cas, un écrivain ne produit jamais exactement les mêmes caractéristiques. Il existe une marge de fluctuation normale et il faut définir cette marge pour pouvoir décider si les valeurs observées chez un autre écrivain tombent dans cette marge (H_0) ou sont effectivement différentes (H_1).

Mais ici, les procédures présentées dans la première partie de cette note doivent être modifiées pour deux raisons. D'une part, les longueurs de phrases ne forment pas une population homogène, gouvernée par une loi de distribution unique. Deux caractéristiques le signalent : l'éloignement des valeurs centrales (mode, médiane et moyenne) et la dispersion considérable de ces longueurs (mesurée par leur écart type)[1]. En revanche, chacun des paramètres (mode, médiane, moyenne et médiale) semble relativement stable. Les tests devraient donc porter sur la stabilité de ces paramètres chez un écrivain donné et leurs différences – significatives ou non - entre écrivains différents. Il faut pour cela vérifier, au préalable, que chacun des paramètres est effectivement stable ("normal") entre les différentes pièces et dans l'affirmative, associer à chacun d'eux un intervalle de variation normale.

Mais, d'autre part, l'utilisation de la loi normale exige au moins 30 valeurs alors qu'il n'y a sous revue que 7 pièces de J. Racine, 9 et 10 par T. et P. Corneille. Pour surmonter cette difficulté, on utilise non plus les pièces mais les actes. En effet, toutes ces tragédies sont découpées en cinq actes – de longueurs très semblables (pour chacune des pièces). On obtient ainsi des effectifs de 35 textes pour J. Racine, 45 pour T. Corneille et 50 pour P. Corneille.

Le tableau 5 récapitule les résultats de cette expérience sur les sept pièces du noyau central des tragédies présentées par J. Racine.

[1] Voir également l'histogramme de ces longueurs présenté plus bas (tableau 6)

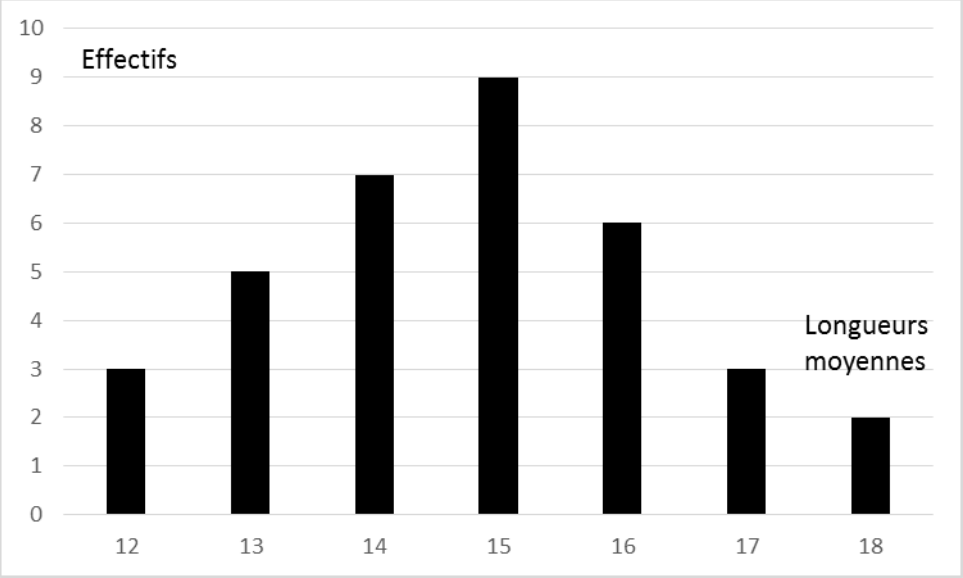
Tableau 5. Caractéristiques des phrases dans les tragédies publiées par J. Racine découpées en actes (en mots, classement par ordre chronologique).

Textes	Longueurs	Ponctuations internes	Mode	Médiane	Moyenne	Médiale	Ecart type
Andromaque 1	3 472	1,70	9	14,75	17,45	22,25	11,56
Andromaque 2	2 964	1,40	9	9,53	13,47	17,07	10,23
Andromaque 3	3 116	1,30	9	9,25	12,37	15,78	9,46
Andromaque 4	3 166	1,84	18	15,08	16,32	19,23	11,63
Andromaque 5	2 358	1,29	9	9,29	12,48	16,43	10,12
Britannicus 1	3 180	1,54	10	13,88	16,31	19,60	11,67
Britannicus 2	3 498	1,77	9	9,45	14,89	19,24	13,69
Britannicus 3	2 903	1,46	8	10,42	13,50	17,53	10,96
Britannicus 4	3 275	2,06	9	11,67	17,24	24,54	13,19
Britannicus 5	2 531	1,63	9	9,29	14,14	19,88	12,30
Bérénice 1	2 833	1,65	8	10,75	15,40	19,05	13,03
Bérénice 2	3 019	1,70	9	9,63	15,02	19,83	13,87
Bérénice 3	2 492	1,26	9	8,68	11,70	16,16	9,51
Bérénice 4	2 677	1,40	9	8,93	12,06	15,54	9,53
Bérénice 5	2 221	1,52	8	8,63	13,80	19,39	13,22
Bajazet 1	3 620	1,62	8	14,32	16,23	18,93	11,21
Bajazet 2	3 246	1,61	8	9,63	14,89	19,55	12,49
Bajazet 3	2 945	1,49	9	9,55	14,44	18,96	12,47
Bajazet 4	2 691	1,22	8	8,17	11,50	15,60	10,65
Bajazet 5	2 795	1,74	7	11,00	14,71	19,90	13,13
Mithridate 1	3 288	1,89	9	13,50	16,95	26,44	12,79
Mithridate 2	3 414	1,76	8	9,63	16,03	26,28	14,65
Mithridate 3	3 276	1,67	3	11,92	15,67	21,40	12,68
Mithridate 4	2 942	1,67	9	10,05	14,49	18,92	12,15
Mithridate 5	2 171	1,93	8	14,83	17,09	21,66	13,75
Iphigénie 1	3 406	1,98	17	15,86	17,56	22,09	12,60
Iphigénie 2	3 293	1,51	9	9,82	15,39	18,52	12,52
Iphigénie 3	2 821	1,37	8	9,43	13,63	16,88	11,67
Iphigénie 4	3 631	1,25	9	9,30	13,25	17,26	11,49
Iphigénie 5	2 631	1,45	1	9,63	12,71	16,84	10,35
Phèdre 1	3 164	1,46	9	9,72	15,07	18,96	15,14
Phèdre 2	3 247	1,37	8	10,18	13,64	17,09	10,46
Phèdre 3	2 299	1,09	8	9,62	12,84	15,53	8,29
Phèdre 4	2 874	1,23	8	9,27	13,43	16,90	10,27
Phèdre 5	2 810	1,63	7	13,38	15,36	18,11	10,32
Moyenne	2 979	1,56	9	10,80	14,60	19,07	11,80
Ecart type	391	0,22		2,18	1,70	2,78	
Borne -	2 228	1,18		6,65	11,36	13,79	
Borne +	3 728	1,90		14,94	17,84	24,35	

Les effectifs sont suffisants pour appliquer la procédure présentée dans la première partie (loi normale, test de l'écart réduit). En dernière ligne du tableau figurent les écarts types de chacune des variables avec les bornes inférieures et supérieures de l'intervalle à 95%.

Pour appliquer le raisonnement présenté dans la première partie (test de la loi normale centrée réduite), il faut auparavant s'assurer que les différentes valeurs prises par les 4 paramètres se répartissent à peu près de manière harmonieuse (population unique). C'est bien le cas ici. A titre d'exemple, le tableau 6 donne l'histogramme des fréquences des moyennes de longueurs des phrases – du tableau 5 - classées par ordre croissant dans des intervalles de classes égaux (à l'horizontale les longueurs, à la verticale, les effectifs des classes). La courbe en cloche atteste d'une répartition normale autour de la moyenne (14,60 mots par phrase) comprise dans la classe modale (15 mots par phrase).

Tableau 6. Histogramme des longueurs moyennes de phrase dans les 35 actes du noyau central des tragédies présentées par J. Racine.



Les mêmes profils de distribution sont obtenus – pour les trois corpus - sur les différentes variables (nombre moyen de ponctuations par phrases, longueurs modales, médianes, moyennes et médiales). Toutes ces populations sont "gaussiennes". Cela révèle des habitudes stables et permet d'utiliser les tests présentés dans le chapitre III.

On dispose de corpus dotés de valeurs moyennes pour les différents paramètres sélectionnés. Ces moyennes sont inscrites dans des intervalles de fluctuation contenant 95% des valeurs. Soit deux corpus (A, B), si une moyenne de la population A tombe en dehors de l'intervalle de fluctuation standard pour la variable correspondante dans la population B, on peut conclure avec moins de 5% de chances d'erreur que les corpus présentent une caractéristique

significativement différente. Plus grand est le nombre d'écarts non-significatifs, plus grande la probabilité d'un écrivain unique. Mais auparavant, il faut vérifier que la méthode a la capacité de discriminer des écrivains différents.

Le tableau 7 donne ces plages de fluctuation standard (α = 5%) pour les tragédies contemporaines présentées par J. Racine et par T. Corneille.

Tableau 7. Indices stylistiques moyens et plages de fluctuation normale dans les pièces présentées par T. Corneille et J. Racine (tragédies découpées en actes, en nombre de mots, α = 5%).

	Longueur	Ponctuation par phrase	Longueur médiane	Longueur moyenne	Longueur médiale
T. Corneille					
Moyenne	3 618	2,18	19,17	22,85	32,44
Borne-	3 016	1,60	13,06	16,82	23,30
Borne+	4 201	2,76	25,27	28,87	41,58
Racine (noyau)					
Moyenne	2 979	1,56	10,80	14,60	19,07
Borne-	2 228	1,11	6,65	11,36	13,79
Borne+	3 728	2,00	14,94	17,84	24,35

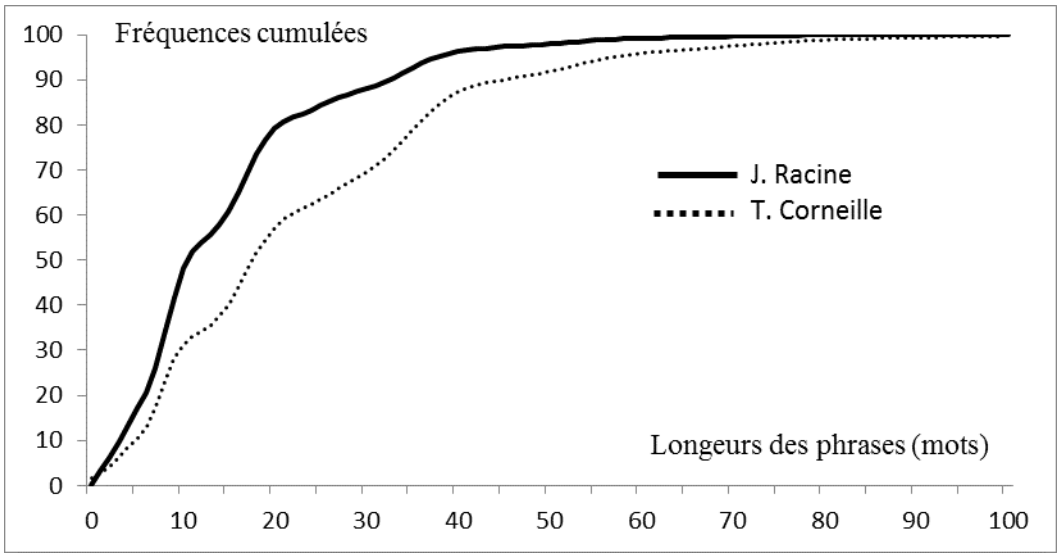
Entre les deux corpus toutes les moyennes diffèrent significativement sauf une. La longueur des actes de T. Corneille est en moyenne de 3 618 mots, légèrement inférieure à la borne haute de l'intervalle dans les pièces publiées par J. Racine (3 728). En revanche, la longueur moyenne des actes dans les tragédies présentées par J. Racine est de 2 979 mots, inférieure à la borne basse de l'intervalle standard chez T. Corneille (3 016). Il en est de même pour les quatre autres valeurs significatives (nombre moyen de ponctuations par phrase et les trois valeurs centrales). En utilisant les trois valeurs centrales et en admettant que ces variables sont indépendantes, la marge d'erreur est de 0.05 à la puissance trois, soit 0.000125. Autrement dit, on peut affirmer avec une chance sur dix mille de se tromper que les tragédies contemporaines présentées par J. Racine et T. Corneille ne sortent pas du même moule. En revanche, en se reportant aux premières lignes du tableau 4 ci-dessus, on voit que – sauf pour les longueurs - les caractéristiques des deux premières tragédies publiées par J. Racine (*La Thébaïde* et *Alexandre*) peuvent entrer soit dans l'un ou dans l'autre des deux cadres du tableau 5, ce qui confirme leur place à part et justifie que, là encore, elles soient placées en dehors de l'analyse.

La comparaison entre J. Racine et P. Corneille aboutit aux mêmes conclusions avec le même risque d'erreur très faible[1].

Plusieurs types de phrases

Cette analyse est complétée par un examen des séries entières grâce à une représentation graphique. Les phrases sont rangées par longueurs croissantes et l'on compte leur effectif relatif pour chaque longueur. Le diagramme des fréquences (relatives) cumulées est la figure habituellement utilisée pour représenter cette population et porter un premier jugement (tableau 8). Plus la courbe s'éloigne de la diagonale, plus la répartition du caractère (longueurs) est inégal entre les individus (phrases).

Tableau 8. Diagramme des longueurs cumulées des phrases dans les tragédies publiées par J. Racine et T. Corneille

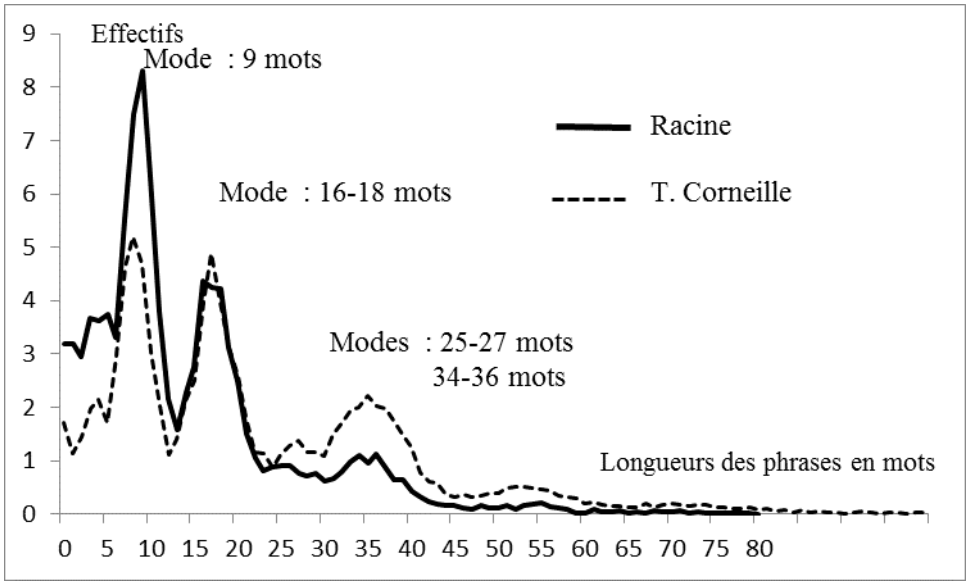


Pour les tragédies publiées par J. Racine et de T. Corneille, les deux courbes sont totalement distinctes. Autrement dit, dans ces œuvres, les phrases sont faites de manière très différente.

On remarque que les courbes ne sont pas régulières mais légèrement bossuées. L'histogramme des longueurs de phrases classées par longueurs croissantes permet d'observer ces irrégularités et de constater que le phénomène est loin d'être homogène (Tableau 9).

[1] Pour les phrases de P. Corneille, voir : Labbé Cyril & Labbé Dominique. Ce que disent…, *Art cit.*

Tableau 9. Histogramme des longueurs de phrases dans les tragédies publiées par J. Racine et T. Corneille



Même si l'influence de l'alexandrin se voit clairement chez les deux – avec les trois mêmes modes principaux -, J. Racine (trait gras) et T. Corneille (pointillés) sont bien distincts. Chez le premier, les phrases courtes prédominent largement ; les phrases moyennes et longues, voire très longues chez le second. Cela traduit des choix stylistiques profondément différents.

Que donnent ces outils appliqués aux pièces présentées par J. de la Chapelle et J. G. Campistron ?

III. STYLES DANS LES PIECES PUBLIEES PAR J. RACINE, J. DE LA CHAPELLE ET J.-G. CAMPISTRON

Les longueurs de phrases dans les pièces présentées par J. La Chapelle et J.-G. Campistron sont relevées selon les conventions exposées au début de ce chapitre. Un examen graphique est associé aux tests selon la double démarche présentée ci-dessus.

Tests statistiques

Pour le corpus La Chapelle, comme il n'y a que 15 observations (3 tragédies comportant chacune 5 actes), il est impossible d'associer des intervalles de variation normale aux valeurs moyennes (dernières lignes du tableau 10). On ne donne ici que les valeurs par pièce pour alléger les tableaux.

Tableau 10. Caractéristiques des phrases dans les pièces présentées par J. de La Chapelle (découpés en actes)

Textes	Longueur (mots)	Ponctuations par phrase	Mode	Médiane	Moyenne	Médiale
Cléopâtre	2 432	1,95	8	12,60	16,39	19,58
Téléphone	2 485	1,81	9	11,21	15,44	18,61
Zaïde	2 609	1,83	9	12,58	16,60	20,65
Moyennes	2 509	1,81	9	12,11	16,02	19,63
J. Racine						
Moyennes	2 979	1,56	9	10,80	14,60	19,07
Borne moins	2 228	1,11		6,65	11,36	13,79
Borne plus	3 728	2,00		14,94	17,84	24,35

Le test ne porte donc que sur les pièces présentées par J. de La Chapelle comparées à celles de J. Racine et J.-G. Campistron sans pouvoir contrôler le résultat avec la comparaison inverse. Le second cadre du tableau 10 donne cette comparaison les pièces publiées par J. Racine : pour tous les indicateurs, on peut conclure que les valeurs ne diffèrent pas significativement, ce qui permet de retenir l'hypothèse d'un auteur unique.

Le tableau 11 présente les mêmes résultats sur les tragédies du corpus Campistron.

Comme précédemment, on associe aux moyennes un intervalle de variation normale égal à cette moyenne plus ou moins deux écarts-types (calculés sur les actes). Si la moyenne de l'un des corpus tombe en dehors de l'intervalle d'un deuxième corpus, on moins de 5% de chances de se tromper en affirmant que le caractère ne suit pas les mêmes lois dans les deux corpus. Pour en être complètement assuré, il faut faire l'opération inverse (la moyenne du deuxième corpus tombe-t-elle hors de l'intervalle du premier ?), opération qui n'a pas été possible sur le corpus La Chapelle.

Les valeurs moyennes obtenues sur les corpus J. Racine, J. de La Chapelle et J.-G. Campistron tombent toutes dans les intervalles de variation normale calculés sur le noyau des tragédies publiées par J. Racine. Il est possible de conclure que les caractéristiques de longueur des phrases de ces trois corpus ne sont pas significativement différentes et qu'elles appartiennent donc toutes à une même population (écrivain unique).

Tableau 11. Caractéristiques des phrases dans le corpus Campistron (classement chronologique, pièces découpées en actes)

Texte	Longueur (mots)	Ponctuations par phrase	Mode	Médiane	Moyenne	Médiale
Virginie	2687	1,79	8	13.08	16.75	22,08
Arminius	2696	1,32	9	10,61	14,62	18,13
Andronic	2491	1,43	9	11,72	15,21	18,64
Alcibiade	2636	1,27	9	13,82	15,96	18,98
Phocion	2143	1,55	8	12,54	16,28	22,79
Adrien	2563	1,01	9	11,20	15,91	19,95
Tiridate	2298	1,59	8	10,47	15,20	20,47
Pompéia	2376	1,44	8	9,48	13,35	16,90
Aétius	2074	1,10	8	11,54	15,95	19,82
Juba	1951	1,34	9	10,21	14,85	18,00
Tachmas	-	1,80	8	13,60	16,54	19.30
Moyennes	2 392	1,38	9	11,29	15,26	19,57
Borne moins	1 856	1,00		8,86	13,47	16,14
Borne plus	2 927	1,81		13,72	17,05	23,00
J. Racine						
Moyennes	2 979	1,56	9	10,80	14,60	19,07
Borne moins	2 228	1,11		6,65	11,36	13,79
Borne plus	3 728	2,00		14,94	17,84	24,35

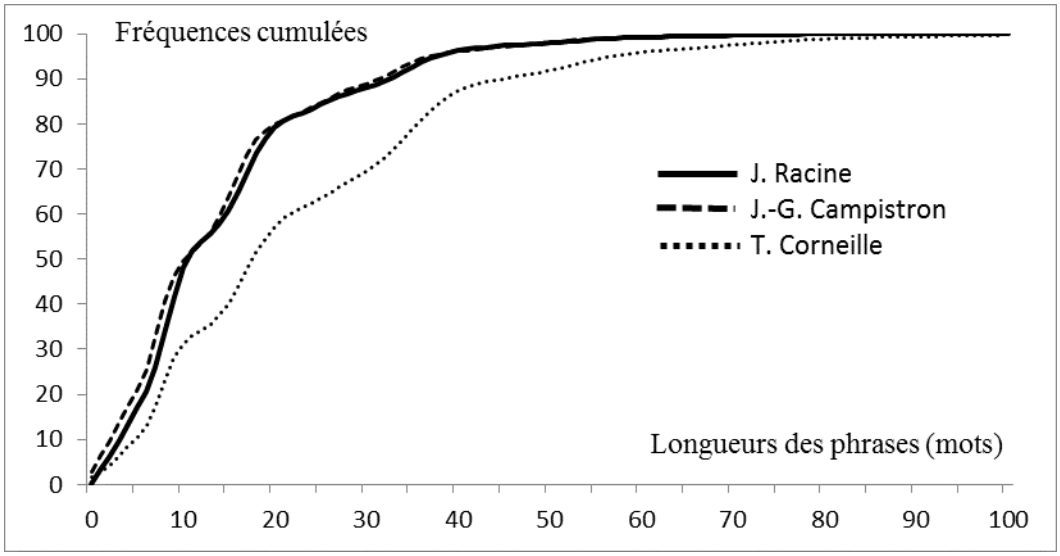
* Les décimales sont données pour indiquer le sens de l'arrondi.

Toutefois ce jugement est porté sur les valeurs centrales (médiane, moyenne, médiale) et sur la dispersion standard autour de ces valeurs. Dans les procédures standards, la convergence de ces indices est largement suffisante pour considérer que les mêmes lois de composition sont bien à l'œuvre dans les trois corpus. On peut souhaiter vérifier que cette identité se retrouve sur l'ensemble des phrases. Cette demande peut se justifier par le fait qu'il ne s'agit pas d'une population homogène mais du mélange de quatre populations différentes (voir plus bas).

Contrôles graphiques

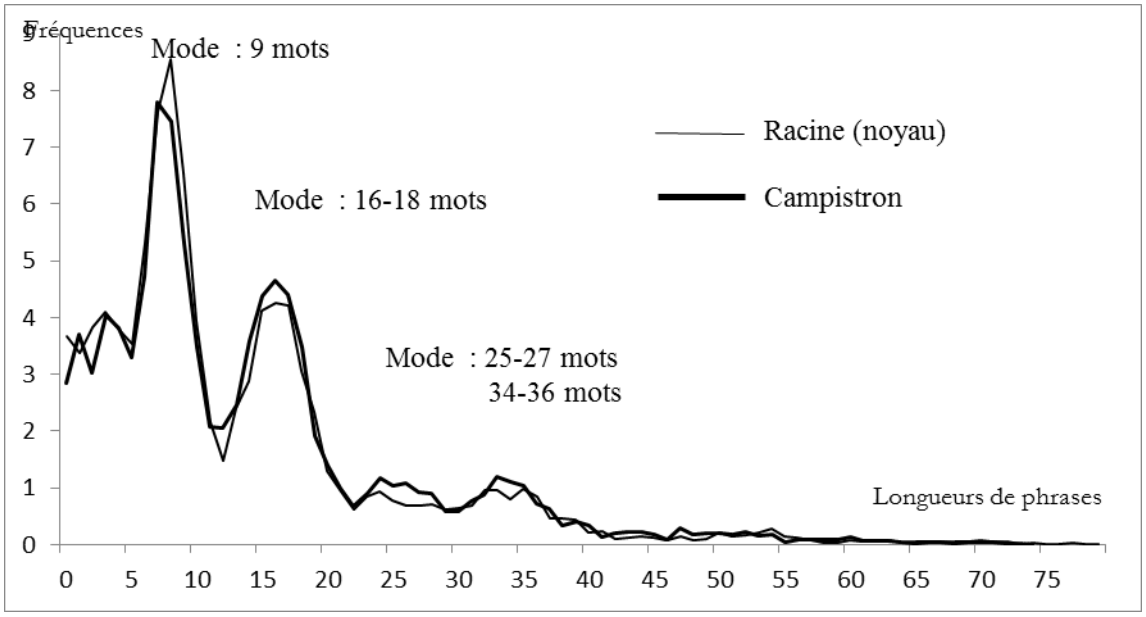
Sur le diagramme des fréquences cumulées (tableau 12), pour l'ensemble des tragédies de J. Racine et de J.-G. Campistron, les deux courbes sont confondues. Autrement dit, dans ces œuvres, les phrases sont faites de la même façon et celle-ci diffère certainement de celle observée chez P. Corneille ou chez T. Corneille.

Tableau 12. Diagramme des longueurs de phrases cumulées dans les tragédies publiées par J. Racine, J.-G. Campistron et T. Corneille.



Le diagramme ci-dessous (tableau 13) permet de vérifier que la partition en quatre sous-population est bien la même dans les deux corpus (à comparer avec le tableau 9).

Tableau 13. Histogramme des longueurs de phrases classées par ordre croissant (tragédies)



Il n'y a donc qu'un auteur avec un style stable au cours de plus d'un quart de siècle.

Comment expliquer ces profils de courbes singuliers ?

Des profils de courbes, comme ceux des tableaux 9 et 13, indiquent l'existence de plusieurs populations. Comme indiqué plus haut, le mode principal, dans les trois corpus est situé à 8-9 mots, c'est-à-dire la longueur moyenne de l'alexandrin. Les modes secondaires sont des multiples de 8-9 : 16-18 ; 24-27… ce qui correspond à des phrases de deux, trois, quatre vers, etc. Autrement dit, les différents écrivains sont contraints par la versification mais ils donnent aux différentes phrases des poids différents (signalés par la hauteur des modes).

On suppose que chacun des quatre types de phrase est homogène (la distribution du caractère épouse une forme en cloche). Cela permet d'utiliser les points d'inflexion des courbes (tableaux 9 et 13) pour délimiter ces populations, chacune étant centrée sur un mode : phrases courtes (de 1 à 13 mots) ; moyennes (de 14 à 22 mots) ; longues (23 à 46 mots) ; très longues (47 mots et plus).

Le vocabulaire de chacun de ces ensembles est comparé à celui des trois autres, grâce à la méthode du vocabulaire caractéristique[1].

Cette comparaison porte sur les catégories grammaticales et sur les vocables.

Dans les phrases courtes, le verbe domine et l'indicatif l'emporte sur les autres temps. Les vocables les plus caractéristiques de ces phrases sont les pronoms personnels : *je*, *tu* et *vous*. Ils sont suivis de : *seigneur*, *monsieur*, *adieu*, *madame* et des verbes *aller*, *dire*, *écouter*, *parler*… Ce sont les phrases de l'interpellation et de l'action. Car l'essentiel de l'action – sur une scène de théâtre – consiste à interpeller l'autre ou les autres, parler, entrer et sortir de scène ! Ces phrases commencent souvent par des interjections et se terminent par des points d'interrogation ou d'exclamation (c'est-à-dire que l'intonation monte jusqu'à la fin au lieu de redescendre comme lorsque la phrase se termine par un point), ce qui ajoute à la tension dramatique. Dans les tragédies présentées par J. Racine, comme dans celles de J. La Chapelle et de J.-G. Campistron, ces brefs échanges ont d'avantage de poids par rapport à celles des deux frères Corneille. Au moins autant que le contenu de ces confrontations verbales, c'est le choix de leur donner la première place qui caractérise le "style" des pièces présentées par J. Racine.

Les phrases de longueur moyenne sont celles de la conversation courante mais débarrassée de la tension présente dans le groupe précédent. Ici encore le groupe verbal l'emporte mais moins nettement que dans les phrases courtes.

Les phrases longues remplissent deux fonctions différentes. La majorité d'entre elles sont des phrases *d'exposition* : récits d'évènements qui se passent en dehors de la scène et sont

[1] Monière Denis, Labbé Cyril et Labbé Dominique. Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada. *Corpus*, 4, 2005, p. 79-104 (texte consultable en ligne sur les Archives ouvertes du CNRS). Nous présentons ici un simple résumé. Un article ultérieur reviendra en détail sur les caractéristiques du style dans les pièces présentées par J. Racine.

rapportés au spectateur ; retours en arrière, ou précisions historiques indispensables pour comprendre un incident ou le comportement d'un personnage. En effet, le théâtre classique était enserré dans un très grand nombre de règles. L'action devait se dérouler en moins de 24 heures, on ne devait pas voir sur scène de violences et encore moins de sang, les personnages devaient se comporter avec bienséance, etc. L'auteur ne peut donc pas montrer de combat, de duel ou de meurtre ; il doit les faire raconter par un témoin. On trouve également des phrases longues dans la bouche de quelques personnages – comme les rois, les empereurs, les grands capitaines – lorsqu'ils parlent aux autres dans le cadre de leurs fonctions. Lorsque de tels personnages parlent beaucoup – par exemple : *Cinna* de Corneille ou *Alexandre* de Racine – la moyenne s'en trouve nettement augmentée.

Les phrases très longues comportent 48 mots et plus. On y trouve quelques rappels historiques et, surtout, l'exposition de la pensée, de l'intériorité, des déchirements d'un personnage clef, soit qu'il parle à un confident, soit qu'il se livre à un monologue (stances).

Voici la phrase la plus longue des pièces publiées par J. Racine. Au premier acte de *Phèdre*, Hippolyte parle de son père (Thésée) :

> Tu sais combien mon âme, attentive à ta voix,
> S'échauffait au récit de ses nobles exploits,
> Quand tu me dépeignais ce héros intrépide
> Consolant les mortels de l'absence d'Alcide,
> Les monstres étouffés et les brigands punis,
> Procruste, Cercyon, et Scirron, et Sinnis,
> Et les os dispersés du géant d'Epidaure,
> Et la Crète fumant du sang du Minotaure :
> Mais quand tu récitais des faits moins glorieux,
> Sa foi partout offerte et reçue en cent lieux ;
> Hélène à ses parents dans Sparte dérobée ;
> Salamine témoin des pleurs de Péribée ;
> Tant d'autres, dont les noms lui sont même échappés,
> Trop crédules esprits que sa flamme a trompés :
> Ariane aux rochers contant ses injustices,
> Phèdre enlevée enfin sous de meilleurs auspices ;
> Tu sais comme à regret écoutant ce discours,
> Je te pressais souvent d'en abréger le cours,
> Heureux si j'avais pu ravir à la mémoire
> Cette indigne moitié d'une si belle histoire.
> (*Phèdre*, acte 1, scène 1, vers 75 à 94)

Cette phrase couvre 20 vers et comporte 163 mots. Elle illustre bien le style de ces pièces : une suite de petits segments épousant le vers (ou le demi vers), juxtaposés plus que coordonnés ou subordonnés. Elle se situe dans la première scène où le fils, discutant avec son confident expose la situation et l'ambivalence de sa relation à son père. Il se parle surtout à lui-même, mais la présence du confident évite l'aspect artificiel des stances.

Conclusions du chapitre

Par rapport aux frères Corneille, les pièces présentées par J. Racine entre 1667 et 1677, se caractérisent par un poids prépondérant donné aux phrases courtes et très courtes. Elles surviennent dans les scènes de confrontation marquées par de fortes densités des interpellations, interjections et exclamations. Au moins autant que le contenu de ces confrontations, c'est le choix de leur donner une place prépondérante qui caractérise le "style" des pièces présentées sous les noms de J. Racine, J. de La Chapelle et J.-G. Campistron.

Pour les pièces présentées par J. Racine, le style n'est stable qu'entre la troisième tragédie (*Andromaque*) et la neuvième (*Phèdre*). La longueur des phrases, leur spécialisation lexicale et leur construction ne seraient donc pas des caractéristiques intrinsèques à l'écrivain mais plutôt le résultat de choix dramaturgiques qui peuvent changer au cours du temps. Dès lors, on ne peut utiliser ces indices pour une attribution d'auteur qu'après s'être assuré, comme nous l'avons fait ici, qu'ils sont stables chez la plume présumée, du moins pour la période considérée.

Sans doute, cette spécialisation des phrases en fonction de leur longueur et de leur construction a-t-elle parue évidente au lecteur. Pourtant, à notre connaissance personne n'avait relevé ni défini précisément ces caractéristiques. Comme pour l'identification de l'écrivain, cette étude est hors de portée d'une simple lecture érudite et les données à manipuler sont trop volumineuses pour un recensement manuel.

Pour l'instant, dans nos dépouillements, nous n'avons pas rencontré deux écrivains présentant de telles ressemblances dans deux de leurs textes et *a fortiori dans toutes leurs œuvres*. La probabilité d'occurrence d'un tel événement paraît donc extrêmement faible. On peut conclure que les caractéristiques des phrases dans les trois corpus J. Racine, J. La Chapelle et J.-G. Campistron désignent un écrivain unique.

Cela renforce les conclusions tirées grâce aux distances intertextuelles et aux classifications automatiques.

CHAPITRE IX
UNE ECHELLE DE LA DISTANCE

Comment éviter de refaire à chaque fois toutes ces opérations complexes. Est-il possible de porter un jugement en considérant seulement quelques couples ?

De multiples expériences – comme celle présentée dans les deux premières parties - ont permis de construire une échelle d'attribution selon les procédures utilisées dans les sciences de l'ingénieur pour calibrer des jauges ou des capteurs. Certaines de ces expériences ont été publiées (voir annexe 1).

Cette échelle peut être utilisée dans certaines conditions strictement définies. Dans ces limites, elle permet de résoudre de nombreuses énigmes.

I. L'ECHELLE

Cette échelle s'applique aux textes en français contemporains – sans fautes d'orthographe et étiquetés - dont les longueurs sont comprises entre 5 000 et 25 000 mots. Pour les textes plus longs ou plus courts, il faut auparavant les traiter avec la technique de la fenêtre glissante appliquée pour *Tachmas*. Une fois les distances acquises, on utilise les repères suivants.

• La distance est inférieure ou égale à 0.20.

Les deux textes ont été écrits par le même écrivain, dans le même genre, à la même époque et les thèmes sont proches. L'attribution des deux textes à un auteur unique peut être faite sans risque d'erreur.

• Entre 0.20 et 0.25, l'écrivain est probablement le même. Sinon, les deux textes ont été écrits à la même époque, dans un même genre, sur un sujet identique et avec des arguments semblables. Ce cas se rencontre dans les articles de presse, à propos d'un même événement, parce que les journalistes travaillent à partir des mêmes sources et citent les mêmes noms de lieux et de personnes... Dans le cas d'œuvres littéraires présentées par deux auteurs différents, le second s'est "inspiré" du premier (dans l'ordre chronologique) ou l'un a aidé l'autre. En tous cas, ce genre de "collision" peut difficilement se produire plusieurs fois entre deux écrivains distincts. Pour une attribution, le risque d'erreur évolue entre 1% pour les distances égales à 0,23 jusqu'à 5% pour les distances égales à 0.25. Une autre manière d'énoncer la même proposition : dans une comparaison entre deux corpus par des écrivains différents, contemporains, travaillant dans un

même genre et sur des thèmes proches, il n'est pas anormal de rencontrer quelques distances comprises entre 0.23 et 0.25 (au maximum 5%). Effectivement, nous avons rencontré quelques distances légèrement inférieures à 0.25 entre certaines tragédies présentées par les frères Corneille et J. Racine.

• A partir de 0.25, on entre dans une zone "grise" où deux hypothèses sont envisageables. Premièrement, les deux textes ont été écrits par un même écrivain dans un genre unique mais à une époque et/ou sur des thèmes éloignés. Deuxièmement, deux écrivains contemporains traitent, dans un même genre, un thème semblable… Plus la distance s'élève, plus la seconde hypothèse est envisageable. En tous cas, l'attribution n'est plus possible du fait d'un risque d'erreur trop élevé, même avec des classifications convergentes.

• Au-dessus de 0.30, pour un même écrivain, le genre est différent ou les dates de composition et les thèmes sont très éloignés. Nous n'avons rencontré ce cas qu'avec les deux dernières tragédies présentées par J. Racine plus de vingt-cinq ans après *La Thébaïde*, sur des thèmes complètement différents.

• Au-dessus de 0.45 les écrivains sont différents ou bien, pour un même auteur, les textes sont de genres très éloignés, par exemple : oral et écrit.

II. CONDITIONS DE VALIDITE ET PRECAUTION D'UTILISATION

Les considérations suivantes sont classiques en sciences de l'ingénieur, mais elles doivent être rappelées pour éviter tout malentendu.

La qualité de la mesure

La qualité de la mesure dépend de celle des observations. Toute recherche qui n'indique pas précisément les conventions et les méthodes d'observation utilisées est sans valeur.

Cela pose notamment le problème de la standardisation orthographique. Comme le lecteur l'aura noté en lisant les extraits de livres de l'époque, ce n'était pas le souci des typographes du XVIIe siècle. Il faut donc transcrire ces pièces en français contemporain et corriger soigneusement ces transcriptions pour s'assurer que chaque mot est toujours écrit de la même façon afin que, en quelque sorte, ces textes passent bien sous la même toise.

Ici la norme consiste à utiliser les transcriptions réalisées au XIXe siècle parues dans la collection « grands écrivains de la France » (Hachette). NB : pour P. Corneille et J. Racine, il s'agit

de la de la dernière édition parue de leur vivant, certaines pièces ayant subi des modifications importantes par rapport à leur première parution…

Les textes sont lemmatisés et le calcul porte les vocables. Cette opération présente plusieurs intérêts (outre la possibilité d'une étude scientifique du vocabulaire). D'une part, on ne prend pas pour des mots différents de simples variations graphiques. Par exemple, Le, La, Les, L', l', le, la, les" sont un seul et même vocable : "le, article" ou "le, pronom" – distingués de la note de musique et des noms propres composés ("La Fontaine"). Ces tâches sont confiées à des automates soigneusement programmés.

D'autre part, cette lemmatisation réduit drastiquement les effectifs des mots à basses fréquences qui sont les principaux éléments perturbateurs en lexicométrie.

La précision de la mesure

D'abord, cette précision ne peut excéder celle des données à partir desquelles la distance est calculée. Cela signifie que pour des textes de moins de 10 000 mots, on ne peut considérer que les trois premières décimales et quatre au-dessus de 10 000 (la décimale suivante servant à indiquer le sens de l'arrondi).

De plus, quand les textes n'ont pas la même longueur, l'indice de la distance est une approximation. Pour les textes dont l'échelle des longueurs est supérieure à 1:2, seules les deux premières décimales de la distance sont significatives, la troisième sera conservée pour indiquer le sens des arrondis et pouvoir prendre des décisions. Lorsque les longueurs s'étalent sur toute la plage 5 000 à 25 000 mots, il faut être attentif aux "effets de bords" qui peuvent se produire pour les textes situés à ces extrémités.

III. ATTRIBUTION D'AUTEUR

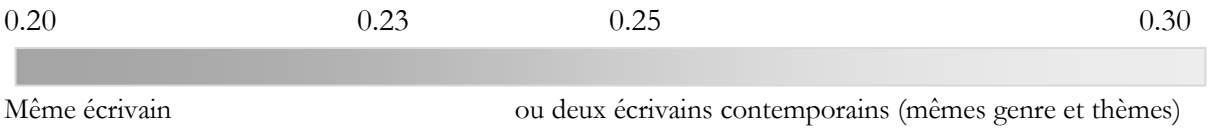
L'échelle présentée ci-dessus peut donc servir à attribuer un texte à un auteur sans avoir à suivre l'ensemble des procédures présentées dans ce rapport. Outre les précautions déjà mentionnées, on mentionnera la question des zones d'acceptation et de rejet ainsi que la nécessité d'indices convergents.

Zones d'acceptation et de rejet

L'échelle est un continuum avec des bornes et non des seuils (schéma ci-dessous). Pour les distances supérieures à 0.22, il existe un risque d'erreur. Ce risque atteint 1% à 0.23 et 5% à 0,25.

Le lecteur peut vérifier la validité de ces bornes sur les tableaux présentés en première partie de ce rapport et dans nos publications sur d'autres corpus (annexe 1)

Tableau 1. Schéma de principe de l'acceptation ou du rejet de l'hypothèse de l'écrivain unique pour deux textes en fonction de leur distance.



Pour une attribution d'auteur, plusieurs hypothèses concurrentes sont à envisager : un auteur unique, une coopération entre deux écrivains ou des écrivains contemporains dans un même genre, sur des thèmes proches et en concurrence pour conquérir un même public. Plus la distance est faible, plus la première hypothèse doit être considérée (zone gris foncé sur la figure ci-dessus). Mais, la troisième hypothèse – deux écrivains contemporains en concurrence - ne peut être écartée en cas d'un petit nombre de valeurs anormales, surtout à proximité de 0,25. En revanche, plusieurs distances proches de 0.20 et une proportion nettement supérieure à 5% de distances inférieures à 0.25 conduisent au rejet de l'hypothèse "deux plumes différentes" et à l'acceptation de l'hypothèse d'un écrivain unique ou d'une collaboration.

Influence, collaboration ou plume de l'ombre ?

Il existe de nombreux cas d'influence (ou de collaboration) dans l'histoire littéraire. Les sœurs Brontë ou les frères Corneille sont emblématiques (nous reviendrons sur ce dernier cas dans une prochaine publication). Dans de précédents travaux, nous avons mentionné : Flaubert (*Madame Bovary*) et le premier roman de Maupassant (*Une vie*) ; Vigny (*Cinq-Mars*) et Dumas (*Les trois mousquetaires*) ou Hugo (*Notre-Dame de Paris*) ; Vigny (*Grandeur et servitude militaires*) sur Musset (*Confession d'un enfant du siècle*). Mais dans tous les cas, ces influences ne sont visibles que dans certains passages et aucune ne concerne toute une œuvre.

Autrement dit, l'attribution d'un texte ou d'une œuvre à une plume de l'ombre demande plusieurs textes de l'écrivain potentiel et l'*on ne retient l'hypothèse d'un écrivain unique, ou d'une coopération entre deux écrivains, qu'avec un nombre anormal de distances anormalement faibles*. Ou encore : il ne s'agit pas d'attribuer tout texte, mais de rejeter l'hypothèse contraire (deux écrivains différents sans collaboration entre eux) avec un risque d'erreur aussi faible que possible. En contrepartie, on accepte de ne pas conclure dans un certain nombre de cas.

Ce raisonnement vaut pour tout texte en français moderne, y compris le XVIIe siècle. C'est lui qui a permis de reconnaître P. Corneille dans l'ombre de Molière, T. Corneille dans l'ombre de Hauteroche et Montfleury – confrères et contemporains de Molière – ou un écrivain unique pour les pièces présentées par J. Racine, J. de La Chapelle et J.-G. Campistron.

Dans ce cas, l'attribution sera confortée en considérant d'autres indices statistiques – examen des plus proches voisins, longueurs des phrases, combinaisons des mots les plus fréquents, etc. – et grâce l'existence d'indices historiques. Pour le cas Racine-La Chapelle-Campistron, ce rapport en reproduit un : le témoignage du Père Colonia. Les autres indices historiques sont examinés dans la postface aux pièces inédites publiées par J.-C. Basson et D. Labbé.

Rappelons enfin que l'attribution d'auteur n'est pas une fin en soi mais qu'elle peut amener un regard neuf sur des œuvres méconnues ou mal comprises parce qu'il y avait eu méprise sur leur véritable auteur.

Conclusion générale

Notre recherche débouche d'abord sur des réponses aux questions posées au début de ce rapport à propos du témoignage du Père Colonia sur la paternité des pièces de théâtre parues sous le nom de J.-G. Campistron.

Au-delà cette question précise, notre recherche a mis en valeur l'efficacité et la fiabilité de l'attribution d'auteur par ordinateur.

Paternité des œuvres parues sous les noms de J. Racine, J. de La Chapelle et J.-G. Campistron

En ce qui concerne les manuscrits de Toulouse et plus largement les œuvres parues sous les noms de J. Racine, J.-G. Campistron et J. de La Chapelle, tous les indices lexicométriques et stylistiques (mais aussi historiques) convergent et permettent de conclure que :

1. Sept tragédies présentées par J. Racine (d'*Andromaque* à *Phèdre*) ont été composées par la même main, avec une certaine hétérogénéité qui ne remet pas en cause l'existence d'une plume unique. En revanche, les deux premières et les deux dernières pièces présentées par J. Racine - *La Thébaïde* et *Alexandre* d'une part, *Esther* et *Athalie* d'autre part – sont nettement différentes des sept composant le noyau central. Il existe donc une énigme J. Racine sur laquelle nous reviendrons ultérieurement ;

2. Le même écrivain a écrit les neuf tragédies présentées par J.-G. Campistron - de *Virginie* (1683) à *Aétius* (1693) - ainsi que les trois présentées et publiées par J. de La Chapelle (*Zaïde*, *Téléphonte* et *Cléopâtre*) ;

3. Il a également composé les deux manuscrits inédits conservés aux archives départementales de Toulouse : *Juba* et *Tachmas*.

Naturellement, les tests statistiques ne "reconnaissent" pas l'écrivain à proprement parler. Ils indiquent simplement quels textes ont été écrits par une même main. L'identification de l'écrivain passe par l'examen d'autres indices, notamment historiques. Par exemple, étant donné leurs dates de naissance, J.-G. Campistron (1656-1723) et J. de La Chapelle (1651-1723) ne peuvent être cet écrivain puisqu'ils étaient trop jeunes au moment où parurent les pièces présentées par J. Racine (1666 à 1677) et qu'ils n'étaient pas présents à Paris. Sous les réserves indiquées au § 1 ci-dessus, il faut donc attribuer à J. Racine les 12 pièces présentées par J. de La Chapelle puis J.-G. Campistron, ainsi que *Juba* et *Tachmas*, comme on lui attribue les pièces parues sous son nom entre 1667 (*Andromaque*) et 1677 (*Phèdre*).

Outre l'énigme J. Racine, d'autres questions restent pendantes. Par exemple, des pièces lyriques et deux comédies sont également parues sous le nom de J.-G. Campistron et une comédie sous le nom de J. de La Chapelle (voir corpus en annexe 1). Sont-elles aussi de la même main ? Ou bien leurs "producteurs" ont-ils fait appel à d'autres plumes de l'ombre pour les composer ? Ce sera l'objet d'une étude à paraître prochainement.

Enfin, les documents conservés aux archives départementales de Toulouse présentent un autre intérêt : ils montrent les différents stades d'élaboration des pièces de théâtre, notamment celles parues sous les noms de J. Racine, J.-G. Campistron et J. de La Chapelle.

Sur l'attribution d'auteur

Nous avons présenté en détail la méthode afin de satisfaire les curiosités légitimes et de prévenir certaines objections. Cette méthode a été mise au point avec le plus grand soin selon les méthodes statistiques usuelles. Les mesures présentées sont vérifiables et reproductibles : toute personne, appliquant cette méthode aux textes présentés, trouvera les mêmes résultats et aboutira aux mêmes conclusions. Les fichiers électroniques ainsi que les programmes sont à la disposition des chercheurs qui souhaitent refaire ces calculs ou approfondir tel ou tel point (prendre contact avec l'auteur).

La démarche consiste à enregistrer les données lexicales, stylistiques (mais aussi historiques) - comme le font les recenseurs avec la population ou les météorologues avec les températures, les pressions, l'hygrométrie - puis à traiter ces observations avec les méthodes éprouvées dans les sciences de l'ingénieur. Outre la précision et le respect des procédures, les principales règles sont : ne rien imaginer, ne pas chercher à combler les imprécisions ou les observations manquantes par supposition, ne porter aucune appréciation qualitative.

Nous avons particulièrement insisté sur la confiance que l'on peut accorder aux conclusions. En effet, l'existence d'un "risque d'erreur" – spécialement dans les études sur échantillons - sert souvent, dans les milieux non-scientifiques, pour rejeter *a priori* toute étude statistique (du moins quand les résultats vont à l'encontre de la doxa). En fait, l'incertitude (plutôt que l'erreur) réside essentiellement dans les échantillonnages. Or, nous ne travaillons pas sur des échantillons mais sur des corpus complets (recensement exhaustif à l'unité près), ce qui élimine la plupart des incertitudes. De plus, *l'hypothèse d'un auteur unique, ou d'une coopération entre deux auteurs, n'est retenue qu'avec un nombre anormal de distances anormalement faibles et avec des indices stylistiques et historiques concordants.*

Dans le cas Racine – La Chapelle – Campistron, plus de 90% des distances entre les trois œuvres ("inter") sont anormalement faibles alors qu'il devrait y en avoir moins de 5%. De plus, il n'est jamais arrivé de rencontrer, chez deux auteurs différents, des nombres de ponctuations

internes à la phrase, des longueurs de phrases médianes, moyennes, médiales non significativement différentes. Quant au cumul de ces quatre événements avec 90% de distances anormalement faibles… De plus, la classification automatique indique que tous les textes présentés par J. Racine, J. de La Chapelle et J.-G. Campistron ont des indices d'appartenance à une œuvre unique proches de 99%. Enfin, rien dans l'histoire n'indique que J. de La Chapelle et J.-G. Campistron se sont comportés comme des écrivains. En revanche, il n'y a pas que le Père Colonia pour affirmer que J. Racine leur était lié et qu'il a continué à produire pour le théâtre après sa retraite officielle de 1677[1].

C'est la convergence de tous ces indices qui conduit à conclure à l'écrivain unique.

Cependant, il reste un "argument" que nous avons déjà signalé dans notre article de 2001 à propos de Molière-Corneille[2]. Dans les années 1680-90, les œuvres de J. Racine était appréciées du roi et de la cour. Elles auraient donc servi de modèle à J. de La Chapelle et à J.-G. Campistron qui les auraient imitées dans leur vocabulaire, leur thématique, leur prosodie et leur style[3].

Cet argument "ad hoc", indémontrable, va encore beaucoup servir car, pour le XVIIe siècle, plus de la moitié des pièces sont concernées (voir la liste des comédiens poètes et de leur "œuvres" en annexe 3). Rappelons que :

- dans l'histoire, on ne connaît aucun cas d'un tel mimétisme ininterrompu entre deux écrivains. Nous avons donné l'exemple des deux frères Corneille dont les vies et les œuvres fournissent un cas d'école, comme celles des sœurs Brontë : bien que les influences mutuelles et des collaborations localisées soient repérables, ces écrivains sont aisément départagés et leurs œuvres sont attribuées sans erreur ;

- pour le théâtre du XVIIe siècle, le nom sur la couverture du livre, sur les affiches des troupes et dans les gazettes donne l'identité de celui qui a négocié le texte avec les comédiens et avec l'éditeur mais pas celle de l'écrivain qui a composé ce texte ;

- avec les méthodes de la critique littéraire traditionnelle, il est impossible de reconnaître l'écrivain qui a composé un texte.

---

[1] Le dossier historique est discuté dans : Basson Jean-Charles & Labbé Dominique *Op. cit.*

[2] "Molière usually directed and played Corneille's works and he could have been " immersed " in Corneille's language and ready to write in the same way as the author he prefered and of whom he knew thousands of verses" (Labbé Cyril & Labbé Dominique. Inter-Textual Distance and Authorship Attribution. Corneille and Molière. *Journal of Quantitative Linguistics*. 8-3, december 2001, p. 229).

[3] L'argument a servi à R. Gary et au pseudo E. Ajar quand leur parenté avait été révélée. Rappelons que Paul Pavlovitch et R. Gary étaient cousins. Le second avait demandé au premier d'incarner E. Ajar pour dissimuler la supercherie que personne n'a été capable d'éventer.

L'intérêt de la statistique appliquée aux textes ne s'arrête pas à l'identification de l'auteur. Elle permet de jeter un regard neuf sur ces textes, sur leur vocabulaire et leur style.

Pour le théâtre des XVIIe et XVIIIe siècles, ce travail de dépouillement n'en est qu'à ses débuts. La principale difficulté réside dans la graphie des mots qui est très changeante, de telle sorte qu'aucune statistique ne peut être établie sur les textes originaux. La transcription en français contemporain, la standardisation de l'orthographe, le balisage des pièces et leur traitement informatique sont des opérations longues mais cruciales. De leur qualité dépend la solidité des conclusions. C'est grâce à ces traitements que l'on peut identifier les écrivains qui ont composé les pièces et mettre au jour leurs vocabulaires, leurs singularités, leurs caractéristiques stylistiques et thématiques.

Ce rapport n'a pu qu'effleurer ce sujet sur lequel reviendront de futures publications.

Enfin, ces outils s'appliquent non seulement à la littérature moderne (du XVIIe à nos jours), mais aussi au discours politique, au vocabulaire des sciences et des techniques, au journalisme, au français oral, etc. Ces outils sont également très utiles dans de multiples domaines, comme la lexicographie, l'enseignement des langues ou la gestion des grandes bases de données documentaires.

ANNEXE I.

L'attribution d'auteur assistée par ordinateur

Love 2002 présente une synthèse des problèmes d'attribution d'auteur[1].
Pour l'utilisation de la statistique et de l'ordinateur, beaucoup de méthodes et d'indices ont été proposés. On trouvera une présentation dans : Stamatatos, 2009 ; Jokers 2013 ; Koppel & Al., 2009 ; en français : Savoy, 2014[2].

Une méthode originale a été mise au point à la fin des années 1990 par Cyril Labbé et Dominique Labbé, grâce à un réseau de chercheurs comprenant notamment : Jean-Guy Bergeron, Mathieu Brugidou, Pierre Hubert, Nelly & Jean Leselbaum, Xuan Luong, Thomas Merriam, Denis Monière, Mathieu Ruhlman. Cette mise au point a suivi les procédures habituelles en sciences de l'ingénieur : tests sur de larges échantillons, expériences en aveugles, présentations préalables dans des séminaires et congrès, et enfin, publications dans des revues internationales à comité de lecture : Labbé & Labbé 2001 ; Monière & Labbé 2002 ; Labbé & Labbé 2003 ; Labbé & Labbé 2004b ; Labbé & Labbé 2006 ; Labbé 2007 ; Labbé & Labbé 2011, 2012, 2013, 2014.

Certaines des communications préalables ont donné lieu à publication et sont consultables en ligne :
- les discours des Premiers ministres de la Ve République : Université de Montpellier (3 décembre 1998) et Ecole Normale de Fontenay Saint-Cloud (12 février 1999) : Labbé 1998.
- les discours des Premiers ministres québécois à Lausanne en mars 2000 (Monière & Labbé 2000).
- des entretiens portant sur les relations industrielles au Québec (Congrès mondial des sociologues de langue française : Bergeron & Labbé 2000).
- des entretiens sur les usages de l'électricité : Grenoble (mars 2001) : Labbé & Labbé 2001a.
- les romans de R. Gary et E. Ajar : Labbé 2004d.

Autres présentations devant des séminaires et conférences :
- Université de Paris-Orsay, janvier 2004 (Labbé 2004a)
- Table ronde de Louvain, mars 2004 (Labbé 2004b)
- Trinity College (Dublin) : Labbé 2004c

[1] Voir bibliographie à la fin de cette note.

[2] A la date de rédaction de cette note, ces articles étaient librement consultables en ligne.

- Séminaire Mathématiques et société, Université de Neuchâtel (Labbé, 2009)
- Société jurassienne d'émulation, Porrentruy (Labbé 2010)
- Cercle philologique, Université de Padoue (Labbé 2011a)
- Ecole supérieure d'interprétation. Trieste (en italien : Labbé 2011b)
- Séminaire de stylistique française. Université de Cologne (Labbé 2011c)
- Université Inter-Ages du Dauphiné. Grenoble 2014 (Labbé 2014a)
- Séminaire *L'œuvre et son auteur : problèmes d'attribution*. Lille 2014 (Labbé 2014b).

Expériences sous contrôle scientifique (outre les expériences préalables déjà citées ci-dessus).
- en 2001, sur des textes choisis par E. Brunet (Labbé 2002)
- en 2004, sur des textes choisis par G. Ledger et T. Merriam (Labbé 2007)
- en 2007, sur des poètes français (Labbé & Labbé 2007)
- en 2010-2011 sur des textes anonymés : Labbé & Labbé 2011
- en 2014 sur quatre romanciers français : Labbé 2014b
- en 2012-2014, sur un corpus de plagiats d'articles scientifiques : Labbé & Labbé 2012b, 2013, 2014.

Récemment, cette méthode a permis de détecter plus d'une centaine de fausses publications scientifiques dans les catalogues de deux des éditeurs internationaux les plus prestigieux (comptant au total plus de 11 millions de références) : Van Noorden 2014, Labbé & Labbé 2012a.

Beaucoup de chercheurs utilisent la distance intertextuelle dans leurs travaux, par exemple :
- Cortelazzo, Nadalutti & Tuzzi 2013
- T. Merriam 2002, 2003a, 2003b
- Pauli & Tuzzi 2009
- Savoy 2012, 2014.

Références

Par ordre alphabétique d'auteur et, pour chaque auteur, par ordre chronologique.
Tous nos articles et communications sont consultables en ligne sur le site HAL (archives ouvertes du CNRS).

Bergeron Jean-Guy & Labbé Dominique (2000). L'évaluation de la négociation raisonnée par les acteurs. Une analyse lexicométrique. *Communication au XVIe Congrès international de l'Association internationale des sociologues de langue française*. Québec : juillet 2000. Reproduit dans Bernier Colette et Al. *Formation, relations professionnelles à l'heure de la société-monde*. Paris-Québec : L'Harmattan - Les Presses de l'Université Laval, 2002, p. 239-252.

Burrows John (2003). "Questions of Authorship: Attribution and Beyond". *Computers and the Humanities*. 37-1, 1-32.

Cortelazzo Michele A., Nadalutti Paolo & Tuzzi Arjuna (2013). Improving Labbé's Intertextual Distance: Testing a Revised Version on a Large Corpus of Italian Literature. *Journal of Quantitative Linguistics*. 20-2, June 2013, p. 125-152.

Garette Robert (1995). *La phrase de Racine. Etude stylistique et stylométrique*. Toulouse : Presses universitaires du Mirail.

Jockers Matthew L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: Univ. of Illinois Press, 2013.

Koppel Moshe, Schler Jonathan & Argamon Shlomo (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*. 2009, 60-1, p. 9-26.

Labbé Cyril & Labbé Dominique (2001a). Discrimination et classement au sein d'un groupe d'entretiens. Le cas du confort électrique. *Communication aux journées d'études du CIDSP*. Grenoble : 9 mars 2001.

Labbé Cyril & Labbé Dominique (2001b). Inter-Textual Distance and Authorship Attribution Corneille and Molière. *Journal of Quantitative Linguistics*. 8-3, december 2001, p. 213-231.

Labbé Cyril & Labbé Dominique (2003). La distance intertextuelle. *Corpus*, 2, 2003, p. 95-118.

Labbé Cyril & Labbé Dominique (2006). A Tool for Literary Studies: Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*. 21-3, 2006, p. 311-326.

Labbé Cyril & Labbé Dominique (2007). Baudelaire, Rimbaud et Verlaine. Communication aux VIIIe Journées de l'ERLA. Brest : 16-17 novembre 2007. Publié dans BANKS David (Ed). *Aspects linguistiques du texte poétique*. Paris, l'Harmattan, 2011, p. 17-45.

Labbé Cyril & Labbé Dominique (2009). Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. Communication aux Xe Journées de l'ERLA. Brest : 20-21 novembre 2009. Publié dans : BANKS David. Le texte épistolaire du XVIIe siècle à nos jours. Paris : L'Harmattan, 2013, p. 53-85.

Labbé Cyril & Labbé Dominique (2011). 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 images*. (http://images.math.cnrs.fr/La-classification-des-textes.html). 28 mars 2011.

Labbé Cyril & Labbé Dominique (2012a). Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*. Published on line: 22 june 2012.

Labbé Cyril & Labbé Dominique (2012b). Detection of Hidden Intertextuality in the Scientific Publications. In Dister Anne, Longrée Dominique, Purnelle Gérald (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège : LASLA - SESLA, p.537-551.

Labbé Cyril & Labbé Dominique (2013). *L'intertextualité dans les publications scientifiques*. Conférence invitée. Séminaire du Laboratoire de Linguistique et Didactique des langues. Grenoble, 28 juin 2013.

Labbé Cyril & Labbé Dominique (2014). *Who wrote this scientific text?* Technical report. Grenoble : Laboratoire d'Informatique de Grenoble (LIG). September 2014.

Labbé Dominique (1998). Les déclarations gouvernementales sous la Ve République (1959-1997). In Autin Jean-Louis et Weill Laurence (Eds). *Le Droit figure du politique. Etudes offertes au professeur Michel Miaille*. Montpellier : Université de Montpellier I, 2008, tome I, p. 843-865.

Labbé Dominique (2002). *Qui a écrit quoi ? L'attribution d'auteur et la distance intertextuelle*. Grenoble : CERAT, juillet 2002.

Labbé Dominique (2004a). Corneille et Molière. *Séminaire du Groupe Langues Informations Représentations*. Université de Paris XI-Orsay : Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), 13 janvier 2004.

Labbé Dominique (2004b). *Corneille et Molière. Table ronde 7e Journées d'Analyse des Données Textuelles*. Louvain-la-Neuve 11 mars 2004. Grenoble : CERAT-IEP, 2004.

Labbé Dominique (2004c). *Corneille in the shadow of Molière*. French Department Research Seminar. Dublin : University of Dublin (Trinity College), April 6 2004.

Labbé Dominique (2004d). *Romain Gary et Emile Ajar*. Grenoble : Cerat-IEP, mai 2004.

Labbé Dominique (2007). Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*. 14-1, 1, April 2007, p. 33-80.

Labbé Dominique (2009). Qui a écrit Dom Juan ? Molière est-il l'auteur des pièces parues sous son nom ? *Communication devant le séminaire Mathématiques et société*. Université de Neuchatel : 9 décembre 2009.

Labbé Dominique (2010). Molière est-il l'auteur des pièces parues sous son nom ? *Communication devant la Société jurassienne d'émulation*. Porrentruy, 2 novembre 2010.

Labbé Dominique (2011a). *Corneille dans l'ombre de Molière. Comment identifier un auteur ?* Conférence invitée. Cercle philologique. Université de Padoue (Italie), 19 janvier 2011.

Labbé Dominique (2011b). *Corneille nell'ombra di Molière*. Conférence invitée. Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, Università di Trieste (Italia), 21 gennaio 2011 (Traduzione : Irene Borsato). Publié dans : *Rivista Internazionale di Tecnica della Traduzione*. 12-2010, p. 117-138.

Labbé Dominique (2011c). *Comédiens et écrivains au XVIIe siècle. A la redécouverte des frères Corneille*. Séminaire de stylistique française. Université de Cologne. Jeudi 9 juin 2011.

Labbé Dominique (2014a). *Les plumes de l'ombre. Molière a-t-il écrit ses pièces ?* Conférence invitée. Université Inter-Ages du Dauphiné. Grenoble : 18 février 2014.

Labbé Dominique (2014b). Identification de l'auteur d'un texte (Hugo, Lamartine, Musset et Vigny). *Conférence invitée au séminaire L'œuvre et son auteur : problèmes d'attribution.* Lille : Université de Lille-Nord de la France, 21 mai 2014.

Love Harold (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press, 2002.

Mikhail Marusenko & Elena Rodionova. Mathematical Methods for Attributing Literary Works when Solving the "Corneille–Molière" Problem. *Journal of Quantitative Linguistics*. 17-1, January 2010, p. 30-54.

Merriam Thomas (2002). "Intertextual Distances between Shakespeare Plays, with Special Reference to Henry V (verse)". *Journal of Quantitative Linguistics*. 9-3, December 2002, p. 260-273.

Merriam Thomas (2003a). "An Application of Authorship Attribution by Intertextual Distance in English". *Corpus*. 2, p 167-182.

Merriam Thomas (2003b). "Intertextual Distances, Three Authors". *Literary and Linguistic Computing*, 18-4, p. 379-388.

Monière Denis & Labbé Dominique (2000). La connexion intertextuelle. Application au discours gouvernemental québécois. In Rajman Martin et Chappelier Jean-Cédric (Eds). *Actes des 5e journées internationales d'analyse des données textuelles*. Lausanne : Ecole polytechnique fédérale, 2000, vol 1, p 85-94.

Monière Denis & Labbé Dominique (2002). Le vocabulaire gouvernemental en France, au Canada et au Québec : 1944-2000. *Etudes canadiennes*. 52, 2002, p. 103-116.

Monière Denis & Labbé Dominique (2006). "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, II, p. 687-696.

Monière Denis & Labbé Dominique (2008). *Les mots qui nous gouvernent. Le discours des premiers ministres québécois : 1960-2005*. Montréal : Monière-Wollank Editeurs, 2008.

Pauli Francesco & Tuzzi Arjuna (2009). The end of year addresses of the Presidents of the Italian Republic (1948–2006): discoursal similarities and differences. *Glottometrics*, 18, p. 40–51.

Pibarot André, Picard Jacques & Labbé Dominique (1998). Les syntagmes répétés dans l'analyse des commentaires libres. In Mellet Sylvie (ed). *4e Journées d'analyse des données textuelles*. Nice, 1998, p. 507-516.

Richaudeau François (1988). *Ce que révèlent leurs phrases*. Paris : Retz.

Savoy Jacques (2012). Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. *Journal of Quantitative Linguistics*. 19(2), 2012, p. 132-161.

Savoy Jacques (2014). La voix du président américain (1934-2014). Née Emilie, Daube Jean-Michel, Valette Mathieu, Fleury Serge (dir.). *Proceedings of the 12th International Conference on Textual Data Statistical Analysis*. Paris: June 3-6 2014, p. 593-604.

Stamatatos Efstathios (2008). Author Identification: Using Text Sampling to Handle the Class Imbalance Problem. *Information Processing & Management*. Volume 44-2. March 2008, Pages 790–799.

Stamatatos Efstathios (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*. 60-3, p. 538-556.

Van Noorden Richard (2014). Publishers withdraw more than 120 gibberish papers. *Nature*. 24 February 2014.

ANNEXE II.

Les corpus

1. Jean Racine

Titre	Genre	Création	Longueur (mots)	Vocabulaire
La Thébaïde ou les frères ennemis	Tragédie	20 juin 1664	13 813	2 256
Alexandre le Grand	Tragédie	4 décembre 1665	13 864	2 383
Andromaque	Tragédie	17 novembre 1667	15 076	2 441
Les Plaideurs	comédie	novembre 1668	8 041	1 839
Britannicus	Tragédie	13 décembre 1669	15 387	2 748
Bérénice	Tragédie	21 novembre 1670	13 242	2 254
Bajazet	Tragédie	1er janvier 1672	15 297	2 598
Mithridate	Tragédie	23 décembre 1672	15 091	2 650
Iphigénie	Tragédie	18 août 1674	15 782	2 753
Phèdre	Tragédie	1er janvier 1977	14 394	2 918
Esther	Tragédie	26 janvier 1689	11 147	2 518
Athalie	Tragédie	5 janvier 1691	15 492	2 989
			166 626	10 145

Source : Paul Mesnard. *Oeuvres de Jean Racine*. Paris : Hachette, 1885 (Les Grands écrivains de la France)

2. Jean de la Chapelle

Titre	Genre	Création	Longueur (mots)	Vocabulaire
Zaïde	Tragédie	26 janvier 1681	13 048	2 314
Cléopâtre	Tragédie	12 décembre 1681	12 158	1 342
Téléphonte	Tragédie	26 décembre 1682	12 425	1 342
			37 631	2 138

Ces pièces ont été transcrites en français contemporain par D. Labbé d'après :
Cléopâtre : La Haye : A. Moetjens, 1683.
Téléphonte : Paris : Thomas Guillain, 1683.
Zaïde : Paris : Jean Ribou, 1681.
(Déchargés sur Google books)

3. Jean-Galbert Campistron

Titre	Genre	Création	Longueur (mots)	Vocabulaire
Virginie	Tragédie	12 février 1683	13 430	1 349
Arminius	Tragédie	19 février 1684	13 478	1 455
Amante amant (l')	Comédie	2 août 1684	16 331	1 621
Andronic	Tragédie	8 février 1685	12 457	1 400
Alcibiade	Tragédie	28 décembre 1685	13 182	1 389
Acis et Galatée	Pastorale	septembre 1686	4 369	799
Achille et Polixène	Tragédie musique	7 novembre 1688	5 990	924
Phocion	Tragédie	16 décembre 1688	10 713	1 315
Adrien	Tragédie	11 janvier 1690	12 835	1 499
Tiridate	Tragédie	12 décembre 1691	11 488	1 330
Aétius	Tragédie	28 janvier 1693	10 368	1 290
Alcide	Opéra	3 février 1693	5 563	926
Pompeia	Tragédie	1693	11 881	1 469
Le jaloux désabusé	Comédie	1709	11 882	1 548
Juba	Tragédie inédite	1695	9 757	1 367
César	Tragédie inédite	-	15 322	1 716
Tachmas	Tragédie inédite inachevée	-	4 039	756
Total			183 085	4 233

Toutes les pièces de ce corpus sont en vers sauf *l'Amante amant* (prose)
Transcrites d'après les éditions originales des œuvres (voir bibliographie) et les éditions de J. Truchet (*Tiridate*) et de J.-P. Grosperrin (*Arminius*, *Andronic* et *Alcibiade*).

ANNEXE III

Les principaux comédiens poètes du XVIIe siècle et leurs "œuvres"
(classement par ordre alphabétique et chronologique)

Auteurs et titres	Date de création
Abeille Nicolas	?
Crispin jaloux	1713
Fausse alarme de l'Opéra (La)	1708
Fille valet (La)	1712
Beaumont (Claude Louvart dit)	?
Mort d'Alexandre (La)	1684
Beauregard (Aymé de)	?
Le docteur extravagant (Le)	1684
Baron (Michel Boiron)	1653-1729
Ecole des pères (L')	168?
Rendez-vous des Tuileries (Le) ou le coquet	1685
Enlèvements (Les)	1685
Homme à bonne fortune (L')	1686
Coquette et la fausse prude (La)	1686
Jaloux (Le)	1687
Fontanges maltraitées, ou les vapeurs	1689
Répétition (La)	1689
Débauché (Le)	1689
Andrienne (L')	1703
Adelphes (Les) ou l' Ecole des pères	1705
Belleroche (Raymond Poisson)	1633-1690
Lubin ou le sot vengé	1661
Baron de la Crasse (Le)	1662
Zig-zag (Le)	1662
Fou de qualité (Le)	1664
Fou raisonnable (Le)	1664
Après-souper des auberges (L')	1665
Poète basque (Le)	1668
Faux moscovites (Les)	1668
Pipeurs (Les) ou les femmes coquettes	1671
Hollande malade (La)	1672
Cocu battu et content (Le)	1672
Fous divertissants (Les)	1680
Comédie sans titre (La)	1683
Brécourt (Guillaume Marcoureau)	1638-1685
Noce au village (La)	165?
Feinte mort de Jodelet (La)	1659
Grand benêt de fils (Le)	1664
Jaloux invisible (Le)	1666
Infante salicoque ou le héros de roman (L')	1667
Les Régals des cousins et des cousines	1674
Ombre de Molière (L')	1674
Appartements	1683
Cassette (La)	1683
Timon, les Flatteurs trompés, ou l'ennemi des	1684
Champmeslé (Charles Chevillet)	1642-1701
Délie	1667
Grisettes ou Crispin chevalier (Les)	1671
Heure du berger (L')	1672
Fragments de Molière (Les)	1674
Parisien (Le)	1682
Rue Saint-Denis (La)	1682
Divorce (Le)	1683
Ragotin ou le roman comique	1684
Florentin (Le)	1685
Coupe enchantée (La)	1688
Veau perdu (Le)	1689
Je vous prends sans vert	1693
Veuve (La)	1699
Chevalier (Jean Simonin)	16??-1674
Pédagogue amoureux (Le)	165?
Cartel de Guillot (Le)	1660
Désolation des filous (La)	1661
Galants ridicules (Les)	1661
Intrigue des carrosses (L')	1662
Disgrâce des domestiques (La)	1662
Barbons amoureux (Les)	1662
Soldat poltron (Le)	1668
Amours de Calotin (Les)	166?
Aventures de nuit (Les)	1680
Crosnier Jacques	?
Ambassadeur d'Afrique (L')	1666
Dancourt (Florent Carton)	1661-1725
Mort d'Hercule (La)	1683
Nouvellistes de Lille	1683
Notaire obligeant (Le)	1685
Angélique et Médor	1685
Ballet de la jeunesse	1686
Fonds perdus (Les)	1686
Renaud et Armide	1686
Désolation des joueuses (La)	1687
Chevalier à la mode (Le)	1687
Maison de campagne (La)	1688
Dame à la mode ou Suite de la Coquette (La)	1689
Folle Enchère (La)	1690
Eté des coquettes (L')	1690
Merlin Déserteur	1690
Carnaval de Venise (Le)	1690
Parisienne (La)	1691
Bon soldat	1691
Femme d'intrigues (La)	1692
Gazette d'Hollande (La)	1692
Opéra de village (L')	1692
Impromptu de garnison (L')	1692

Bourgeoises à la mode (Les)	1692
Femmes à la mode (Les)	1692
Baguette (La)	1693
Vendanges (Les)	1694
Tuteur amoureux (Le)	1695
Foire de Bezons (La)	1695
Vendanges de Suresnes (Les)	1695
Foire Saint-Germain (La)	1696
Moulin de Javelle (Le)	1696
Eaux de Bourbon (Les)	1696
Vacances (Les)	1696
Loterie (La)	1697
Charivari (Le)	1697
Retour des officiers (Le)	1697
Curieux de Compiègne (Les)	1698
Mari retrouvé (Le)	1698
Fées (Les)	1699
Famille à la mode	1699
Opérateur Barry (L')	1700
Fête de village (La)	1700
Trois cousines	1700
Colin-Maillard	1701
Enfants de Paris (Les) ou la Famille à la mode	1704
Mort d'Alcide (La)	1704

Desfontaines (Nicolas Marie)	16??-1652
Eurimédon ou l' illustre pirate	1635
Vraie suite du Cid (La)	1637
Orphise ou la beauté persécutée	1637
Hermogène	1638
Bélisaire	1640
Galantes vertueuses (Les)	1641
Alcidiane ou les quatre rivaux	1642
Perside ou La Suite d'Ibrahim Bassa	1642
Saint Eustache ou Le Martyre de Saint Eustach	1642
Illustre Olympie ou le Saint Alexis (L')	1643
Bellissante ou la fidélité reconnue	1646
Véritable Sémiramis (La)	1646
Illustre comédien (L') ou le martyre de Saint-	164?

Dorimond (Nicolas Drouin)	1628-1673
Festin de pierre (Le) ou L' Athée foudroyé,	1658
Ecole (L') des cocus ou la précaution inutile	1659
Inconstance punie (L')	1659
Amant de sa femme (L')	1660
Comédie de la comédie et les Amours de (La)	1660
Femme industrieuse (La)	1661
Rosélie, ou le Dom Guillot (La)	1661

Du Perche (Jean Crosnier dit)	1643-1709
Epouse fugitive (L')	?
Ombre de son rival (L')	1681
Frayeurs de Cripin (Les)	1682

Dufresny Charles	1657-1724
Epreuve (L')	?
Portrait (Le)	?
Superstitieux (Le)	?
Vapeurs (Les)	?
Deux veuves ou le faux Damis (Les)	?
Négligent (Le)	1692
Opéra de campagne (L')	1692
Union des deux opéras (L')	1692
Chinois (Les)	1692
Baguette de Vulcain (La)	1693
Adieux des officiers ou Vénus justifiée (Les)	1693
Mal-assortis (Les)	1693
Sancho Pança	1694
Attendez-moi sous l'orme	1694
Départ des comédiens (Le)	1694
Foire Saint-Germain (La)	1695
Suite de la Foire Saint-Germain	1696
Momies d'Egypte (Les)	1696
Pasquin et Marforio médecins des moeurs	1697
Chevalier joueur (Le)	1697
Fées (Les) ou les contes de ma mère l'Oie	1697
Malade sans maladie (La)	1699
Noce interrompue (La)	1699
Esprit de contradiction (L')	1700
Double veuvage (Le)	1702
Faux honnête homme (Le)	1703
Bailli marquis (Le)	1703
Faux instinct (Le)	1707
Jaloux honteux de l'être (Le)	1708
Amant masqué (L')	1709
Joueuse (La)	1709
Coquette de village (La) ou le lot supposé	1715
Nouveautés de la foire Saint Germain (La)	1716
Réconciliation normande (La)	1719
Dédit (Le)	1719
Mariage fait et rompu (Le)	1721
Dominos	1722
Faux sincère (Le)	1731

Grandval (Daniel Racot dit)	?
Quartier d'hiver (Le)	1696

Hauteroche (Noel Lebreton)	1617-1707
Feint Polonais (Le) ou la veuve impertinente	166?
Amant qui ne flatte point (L')	1668
Souper mal apprêté (Le)	1669
Crispin médecin	1670
Deuil (Le)	1672
Apparences trompeuses (Les) ou les maris infi	1673
Crispin musicien	1674
Nouvellistes (Les)	1678
Nobles de province (Les)	1678
Bassette	1680
Esprit follet (L') ou La Dame invisible	1684

Cocher supposé (Le)	1684
Bourgeoises de qualité (Les)	1690

La Thorillière (François Le Noir)	1626-1680
Cléopâtre	1667

La Thuillerie (Jean-François Juvenon)	1650-1688
Crispin précepteur	1679
Soliman	1680
Crispin bel esprit	1681
Hercule	1681
Nitocris	1683
Merlin peintre	1687

Legrand Marc-Antoine	1673-1728
Fourberies de Cartouche (Les)	sd
Libertin puni (Le)	sd
Chute de Phaéton (La)	sd
Cafetier (Le)	sd
Fille précepteur (La)	sd
Rue Mercière ou les maris dupés (La)	1694
Carnaval de Lyon (Le)	1699
Comédiens de campagne (Les)	1699
Divertissement pour le retour du roi à Vars.	1701
Femme fille et veuve (La)	1707
Amour diable (L')	1708
Famille extravagante (La)	1709
Foire Saint-Laurent (La)	1709
Amants ridicules (Les)	1711
Epreuve réciproque (L')	1711
Métamorphose amoureuse (La)	1712
Usurier gentilhomme (L')	1713
Aveugle clairvoyant (L')	1716
Triomphe du temps (Le)	1716
Animaux raisonnables (Les)	1718
Roi de Cocagne (Le)	1718
Oedipe travesti	1719
Momus fabuliste ou les noces de Vulcain	1719
Plutus	1720
Belphégor ou la Descente d'Arlequin aux e	1721
Fleuve d'oubli (Le)	1721
Terres australes (Les)	1721
Amours aquatiques (Les)	1721
Cartouche ou les voleurs	1721
Galant coureur ou l'Ouvrage d'un mome (Le)	1722
Polyphème	1722
Ballet des vingt-quatre heures (Le)	1722
Paniers ou La vieille prétieuse (Les)	1723
Triomphe de la folie (Le)	1723
Agnès de Chaillot	1723
Bois de Boulogne (Le)	1723
Départ des comédiens italiens pour l'A (Le)	1723
Philanthrope ou l' Ami de tout le monde (Le)	1724
Mauvais ménage (Le)	1725
Chaos (Le)	1725
Temps passé (Le)	1725
Temps présent (Le)	1725
Nouveaux débarqués (Les)	1725
Impromptu de la folie (L')	1725
Française italienne (La)	1725
Chevalier errant (Le)	1726
Chasse du cerf (La)	1726
Nouveauté (La)	1727
Amazones modernes (Les)	1727
Luxurieux ou le libertin puni (Le)	1731
Brouilleries ou le rendez-vous (Les)	1753
Poupées (Les)	1777

Marcel	?
Mariage sans mariage (Le)	1671

Maréchal André	?
Force du sang (La)	?
Inconstance (L')	1630
Généreuse Allemande (La)	1631
Sœur valeureuse (La)	1633
Railleur ou la satire du temps (Le)	1636
Véritable capitan Matamore (Le)	1637
Cour bergère (La)	1638
Mausolée (Le)	1639
Jugement équitable de Charles le Hardi (Le)	1643
Papyre ou le Dictateur romain (Le)	1645

Molière (Jean-Baptiste Poquelin) 1622-1673	
Jalousie du barbouillé (La)	?
Médecin volant (Le)	?
Etourdi (L')	1659
Dépit amoureux (Le)	1659
Précieuses ridicules (Les)	1660
Sganarelle ou le cocu imaginaire	1660
Dom Garcie de Navarre	1661
Ecole des maris (L')	1661
Fâcheux (Les)	1661
Ecole des femmes (L')	1662
Critique de l'Ecole des femmes (La)	1663
Impromptu de Versailles (L')	1663
Mariage forcé (Le)	1664
Princesse d'Elide (La)	1664
Tartuffe (Le)	1664
Dom Juan	1665
Amour médecin (L')	1665
Misanthrope (Le)	1666
Médecin malgré lui (Le)	1666
Mélicerte	1666
Comédie pastorale (La)	1667
Sicilien ou l'Amour peintre (Le)	1667
Amphytrion	1668
Georges Dandin	1668
Avare (L')	1668
M. de Pourceaugnac	1669

Amants magnifiques (Les)	1670
Bourgeois gentilhomme (Le)	1670
Fourberies de Scapin (Les)	1671
Comtesse d'Escarbagnac (La)	1671
Femmes savantes (Les)	1672
Malade imaginaire (Le)	1673

Montfleury (Zacharie Jacob)	16??-1667
Mort d'Asdrubal (La)	1647

Montfleury (Antoine Jacob)	1640-1685
Garçon sans conduite (Le)	166?
Mariage de rien (Le)	1660
Bêtes raisonnables (Les)	1661
Ecole des jaloux (L') ou le cocu volontaire	1662
Mari sans femme (Le)	1663
Impromptu de l'Hôtel de Condé (L')	1663
Trasibule	1664
Ecole des filles (L')	1666
Fille capitaine (La)	1669
Femme juge et partie (La)	1669
Procès de la Femme juge et partie (Le)	1669
Gentilhomme de Beauce (Le)	1670
Ambigu-comique (L'), ou les amours de Didon	1673
Semblable à soi-même (Le)	1673
Comédien poète (Le)	1673
Trigaudin ou Martin Braillart	1674
Crispin gentilhomme	1677
Dame médecin (La)	1678
Dupe de soi-même (La)	1679

Nanteuil (Denis Clerselier)	1650-17??
Brouilleries nocturnes (Les)	1669
Campagnard dupé (Le)	1671
Comte de Rocquefeuilles (Le)	1672
Amour sentinelle (L')	1672
Fille vice-roi (La)	1672
Amante invisible (L')	1673
Héritier imaginaire	1674

Raisin Jacques	1653-1702
Niais de Sologne (Le)	1686
Petit homme de la Foire (Le)	1687
Faux Gascon (Le)	1688
Merlin Gascon	1690
Baguette (La)	1693

Rosidor (Jean Guillemay du Chesnay)	?
Mort du grand Cyrus (La)	1661

Rosidor (Claude-Ferdinand Guillemay du Chesnay)	?
Amours de Merlin (Les)	1691
Divertissements du temps (Les)	1691

Rosimond (Jean-Baptiste du Mesnil)	1640-1686
Grand festin de pierre (Le) ou l' Athée foudroyé	166?
Duel fantasque (Le) ou les valets rivaux	1668
Nouveau festin de pierre (Le) ou l'Athée foud.	1669
Dupe amoureuse (La)	1670
Trompeurs trompés (Les) ou Les Femmes vert	1670
Savetier avocat (Le) ou L' Avocat sans étude	1670
Quiproquo (Le) ou Le Valet étourdi	1671
Volontaire (Le)	1676

Subligny (Adrien-Thomas Perdou de)	1640-17??
Folle querelle (La) ou la critique d'Andromaque	1668
Désespoir extravagant (Le)	1670

Vauselle (Jean-Baptiste L'Hermite de Souliers dit)	?
Chute de Phaéton (La)	1639

Villiers (Claude Deschamps de)	1601-1681
Apothicaire dévalisé (L')	1658
Côteaux (Les) ou Les Marquis friands	1665
Festin de pierre (Le) ou Le fils criminel	1659
Ramoneurs (Les)	1662
Trois visages (Les)	166 ?

ANNEXE IV

Corpus électronique des pièces de théâtre du XVIIe siècle (décembre 2014)
(classement alphabétique par ordre d'auteur et par ordre des titres de pièces – les auteurs sont ceux qui ont présenté les pièces)

Premier entête	Longueur (mots)	Vocabulaire (vocables)
Boisrobert (Le Metel de - Abbé de Châtillon) - Les deux Alcandres ou les deux semblables - Tragi-comédie représentée pour la première fois en 1640	14 360	1 484
Boisrobert (Le Metel de - Abbé de Châtillon) - La belle Lisimène – Tragi-comédie représentée pour la première fois en 1633	13 722	1 730
Boisrobert (Le Metel de - Abbé de Châtillon) - Les rivaux amis - Tragi-comédie représentée pour la première fois en 1638	15 960	1 714
Total Boisrobert	44 042	2 676
Boursault Edmé - La comédie sans titre - Représentée pour la première fois - le 5 mars 1683 au Théâtre de l'Hôtel Guénégaud.	15 560	2004
Boursault Edmé - Ésope à la cour - Comédie héroïque représentée pour la première fois le 16 décembre 1701 à la Comédie française	22 124	2279
Boursault Edmé – Germanicus – Tragédie représentée - pour la première fois - à Paris - par les Comédiens du Roi - le 25 mai 1673 au Théâtre du Marais.	17 524	1 469
Boursault Edmé - Marie Stuard Reine d'Ecosse - Tragédie représentée pour la première fois le 17 décembre 1683 au Théâtre de l'Hôtel Guénégaud	15 187	1 550
Total Boursault	32 711	1 993
Boyer Claude (sous le nom de Pader d'Assezan) – Agamemnon – Tragédie représentée pour la première fois le 12 mars 1680	14 570	1 481
Boyer (Claude) – Artaxerce – Tragédie représentée pour la première fois au Théâtre Guénégaud le 22 novembre 1682	14 441	1 303
Boyer Claude – Fédéric – Tragédie représentée pour la première fois le 14 novembre 1659 à l'Hôtel de Bourgogne	16 252	1 296
Boyer (Claude) – Judith – Tragédie représentée pour la première fois au Théâtre de la rue des Fossés Saint-Germain le 4 mars 1695	12 896	1 497
Boyer Claude – La mort des enfants de Brute – Tragédie représentée pour la première fois à l'Hôtel de Bourgogne en 1847	12 831	1 376

Boyer (Claude) – Oropaste – Tragédie représentée pour la première fois le 17 novembre 1662 au Théâtre du Palais-Royal	19 140	1 478
Boyer Claude – La Porcie romaine – Tragédie représentée pour la première fois en 1645	14 045	1 480
Boyer (Claude) - Tyridate – Tragédie – 1648 – Théâtre du Marais	14 973	1 384
Total Boyer	119 148	3 161
Campistron Jean-Galbert - Virginie - Tragédie représentée pour la première fois le 12 février 1683 par la Troupe des Comédiens français au Théâtre Guénégaud	13 430	1 349
Campistron Jean-Galbert - Arminius - Tragédie représentée pour le première fois 19 février 1684 au Théâtre de l'Hôtel Guénégaud par la Comédie française.	13 478	1 455
Campistron Jean-Galbert – l'Amante amant – Comédie en prose présentée pour la première fois le 2 août 1684 à la comédie française	16 331	1 621
Campistron Jean-Galbert - Andronic - Tragédie représentée pour le première fois le 8 février 1685 au Théâtre de l'Hôtel Guénégaud par la Comédie française.	12 457	1 400
Campistron Jean-Galbert - Alcibiade - Tragédie représentée pour la première fois le 28 décembre 1685 au Théâtre de l'Hôtel Guénégaud.	13 182	1 389
Campistron Jean-Galbert - Acis et Galatée - pastorale héroïque sur une musique de Lully - représentée pour la première fois au château d'Anet par l'Académie royale de musique en septembre 1686	4 369	799
Campistron Jean-Galbert - Achille et Polixène – Tragédie en musique créé le 7 novembre 1688 par l'Académie royale de musique - Musique de Lully achevée par Colasse	5 990	924
Campistron Jean-Galbert - Phocion - Tragédie représentée pour le première fois 16 décembre 1688 au Théâtre de l'Hôtel Guénégaud par la Comédie française.	10 713	1 315
Campistron Jean-Galbert – Adrien – Tragédie en 5 actes représentée pour la première fois le 11 janvier 1690 par la Comédie française	12 834	1 499
Campistron Jean-Galbert - Tiridate – Tragédie en 5 actes représentée pour la première fois le 12 décembre 1691 à la Comédie française	11 488	1 330
Campistron Jean-Galbert - Pompeia – Tragédie en 5 actes jamais représentée - probablement composée en 1693	11 881	1 469
Campistron Jean-Galbert – Aétius - Tragédie représentée pour la première fois le 28 janvier 1693 par la Comédie française	10 368	1 289
Campistron Jean-Galbert - Alcide - Opéra (paroles de Campistron - musique de Lulli et Marais) - représentée par l'Académie royale de musique le 3 février 1693	5 563	926

Campistron Jean-Galbert – Le jaloux désabusé – Comédie représentée pour la première fois le 13 décembre 1709 à la comédie française	11 882	1 548
Juba. Tragédie inédite	9 757	1 367
Tachmas - Tragédie inédite et inachevée	4 039	756
César – Tragédie en cinq actes	15 322	1 716
Total Campistron	183 084	4 232
Champmeslé (Charles Chevillet dit) – La coupe enchantée – Comédie en prose en un acte représentée pour la première fois le 16 juillet 1688 à la comédie française	6 673	924
Champmeslé (Charles Chevillet dit) - Le Florentin – Comédie en un acte et en vers représentée pour la première fois le 23 juillet 1685 à la Comédie française	4 864	1 070
Champmeslé – Le Parisien- Comédie en vers et en 5 actes représentée pour la première fois le 7 février 1682 à la Comédie française	14 889	2 017
Champmeslé (Charles Chevillet dit) - Ragotin ou le roman comique – Comédie en actes et en vers représentée pour la première fois le 21 avril 1684 à la Comédie française	12 768	1 954
Champmeslé (Charles Chevillet dit) – Je vous prends sans vert – comédie en vers et un acte – Représentée pour la première fois le 1er mai 1693 à la Comédie française	3 315	830
Total Champmeslé	42 509	3 637
Les Cinq Auteurs – L'Aveugle de Smyrne - Comédie écrite en collaboration par F. de Boisrobert G. Colletet P. Corneille C. de l'Estoile J. Rotrou – Représentée en février 1637	16 531	1 643
Les Cinq Auteurs – La comédie des Tuileries - Comédie écrite en collaboration par F. de Boisrobert G. Colletet P. Corneille C. de l'Estoile J. Rotrou – Représentée en février 1635	18 652	1 918
Total Cinq auteurs	35 183	2 432
Corneille (Pierre).- Mélite : comédie - Première représentation : décembre 1629	16 690	1 915
Corneille (Pierre).- Clitandre : tragédie - Première représentation : 1631	14 402	1 811
Corneille (Pierre).- La Veuve : comédie - Première représentation : 1632	17 661	1 845
Corneille (Pierre).- La Galerie du Palais : comédie - Première représentation : 1633	16 140	1 709
Corneille (Pierre).- La Suivante : comédie - Première représentation : 1634	15 160	1 586
Corneille (Pierre).- La Comédie des Tuileries. acte III - Première représentation : février 1635	3 627	813
Corneille Pierre).- Médée : tragédie - Première représentation : 1635	14 269	1 830

Corneille (Pierre).- La Place Royale : comédie - Première représentation : 1634	13 801	1 466
Corneille (Pierre).- L'Illusion comique : comédie - Première représentation : 1635	15 428	1 947
Corneille (Pierre).- Le Cid : tragédie - Première représentation : 5 janvier 1637	16 677	1 626
Corneille (Pierre).- Cinna ou la Clémence d'Auguste : tragédie - Représentée pour la première fois en 1639	16 126	1 691
Corneille (Pierre).- Horace : tragédie - représentée pour la première fois en 1639	16 482	1 586
Corneille (Pierre).- Polyeucte martyr : tragédie - représentée pour la première fois en 1643	16 472	1 708
Corneille (Pierre).- La Mort de Pompée : tragédie - représentée pour la première fois en novembre 1643	16 492	1 743
Corneille (Pierre).- Le Menteur : comédie - Première représentation en 1644	16 653	1 745
Corneille (Pierre).- La Suite du Menteur : comédie - Représentée pour la première fois en 1644	17 675	1 774
Corneille (Pierre).- Rodogune princesse des Parthes : tragédie - Représentée pour la première fois en 1644	16 842	1 625
Corneille (Pierre).- Théodore vierge et martyre : tragédie chrétienne - Représentée pour la première fois en 1645	17 121	1 592
Corneille (Pierre).- Héraclius empereur d'Orient : tragédie - Représentée pour la première fois en 1647	17 433	1 595
Corneille (Pierre).- Andromède : tragédie - Représentée pour la première fois en janvier 1650	15 514	1 561
Corneille (Pierre).- Don Sanche d'Aragon : comédie héroïque - Représentée pour la première fois en 1650	16 947	1 518
Corneille (Pierre).- Nicomède : tragédie - Représentée pour la première fois en février 1651	16 923	1 613
Corneille (Pierre).- Pertharite roi des Lombards : tragédie - Représentée pour la première fois en 1651	17 121	1 528
Corneille (Pierre).- OEdipe : tragédie - Représentée pour la première fois le 24 janvier 1659	18 618	1 686
Corneille (Pierre).- La Toison d'or : tragédie - Représentée pour la première fois en novembre 1660	20 343	1 849
Corneille (Pierre).- Sertorius : tragédie - Représentée pour la première fois le 25 février 1662	17 675	1 651
Corneille (Pierre).- Sophonisbe : tragédie - Représentée pour la première fois le 18 janvier 1663	16 858	1 535
Corneille (Pierre).- Othon : tragédie - Représentée pour la première fois le 1 août 1664	16 971	1 613
Corneille (Pierre).- Agésilas : tragédie - Représentée pour la première fois le 25 février 1666	18 227	1 492
Corneille (Pierre).- Attila roi des Huns : tragédie - Représentée pour la première fois le 4 mars 1667	16 789	1 576
Corneille (Pierre).- Tite et Bérénice : comédie héroïque - Représentée pour la première fois le 28 novembre 1670	16 697	1 472
Corneille (Pierre) Molière (Jean-Baptiste Poquelin dit).- Psyché : tragédie-ballet en alexandrins - Musique de Lully - Représentée pour la première fois le 17 janvier 1671	10 067	1 273
Corneille (Pierre).- Pulchérie : comédie héroïque - Représentée pour la première fois le 25 novembre 1672	16 630	1 433

Corneille (Pierre).- Suréna général des Parthes : tragédie - Représentée pour la première fois : 14 décembre 1674	16 545	1 471
Total Corneille Pierre	547 076	6 152
Corneille Thomas – La mort d'Achille – Tragédie représentée pour la première fois le 29 décembre 1673 au Théâtre de l'Hôtel Guénégaud.	16 014	1391
Corneille Thomas – Le Baron d'Albikrac – Comédie en 5 actes représentée pour la première fois en 1667	17 205	1747
Corneille Thomas – L'Amour à la mode – Comédie représentée pour la première fois en 1651 ou 1653 au Théâtre de l'Hôtel de Bourgogne.	17 625	1741
Corneille Thomas – La Mort d'Annibal – Tragédie en 5 actes et en alexandrins - représentée pour la première fois le 25 novembre 1669 à l'Hôtel de Bourgogne	17492	1465
Corneille Thomas – Ariane – Tragédie en cinq actes et en alexandrins - créée le 26 février 1672 à l'Hôtel de Bourgogne	16 320	1354
Corneille Thomas – Le Feint astrologue – comédie représentée pour la première fois en 1648 à l'Hôtel de Bourgogne	17 732	1765
Corneille Thomas – Bellérophon – Livret d'opéra en cinq actes avec un prologue - Représentée pour la première fois le 31 janvier 1679	5 643	892
Corneille Thomas – Bérénice – Tragédie représentée pour la première fois en 1657 au théâtre du Marais	18 519	1488
Corneille Thomas – Le Berger extravagant – Comédie mêlée d'ornements et de musique - Représenté pour la première fois en 1652	17 284	2073
Corneille Thomas – Bradamante – Tragédie représentée pour la première fois le 18 novembre 1695 au Théâtre de la rue des Fossés Saint-Germain	12 080	1192
Corneille Thomas – Camma Reine de Galatie – Tragédie représentée pour la première fois le 28 janvier 1661 au Théâtre de l'Hôtel de Bourgogne	18 403	1404
Corneille Thomas – Le Charme de la voix – Comédie en vers et en 5 actes représentée pour la première fois en 1656	17 314	1528
Corneille Thomas - Circé - Tragédie ornée de machines - de Changements de Théâtre - et de Musique - représentée pour la première fois le 1er février 1675 au Théâtre Guénégaud.	22 799	1862
Corneille Thomas – La Mort de l'empereur Commode - Tragédie en cinq actes et en alexandrins - Représentée pour la première fois en 1657 au Théâtre du Marais	17 846	1487
Corneille Thomas – La Comtesse d'Orgueil – Comédie en 5 actes et en alexandrins - représentée pour la première fois en 1670 à l'Hôtel de Bourgogne	17 653	1836
Corneille Thomas – Darius – Tragédie en 5 actes en alexandrins - représentée pour la première fois en 1659 à l'Hôtel de Bourgogne	18 320	1424

Corneille (Thomas) et Donneau de Visé (Jean) – La Devineresse ou les faux enchantements – Comédie représentée pour la première fois 19 novembre 1679 au théâtre Guénégaud	27 421	1775
Corneille Thomas – Dom Bertran de Cigarral – Comédie en 5 actes et en alexandrins - représentée pour la première fois en mai 1651 au Théâtre du Marais	17 711	1975
Corneille Thomas – Dom César d'Avalos – Comédie en 5 actes et en alexandrins - représentée pour la première fois le 21 décembre 1674 à l'Hôtel Guénégaud	16 521	1669
Corneille Thomas – Les Engagements du hasard – comédie en 5 actes et en alexandrins - représentée pour la première fois en 1649 à l'Hôtel de Bourgogne	16 213	1567
Corneille Thomas – Les Illustres ennemis – Comédie en 5 actes et en alexandrins - représentée pour la première fois en 1655 à l'Hôtel de Bourgogne	18 513	1500
Corneille Thomas – Le Comte d'Essex – Tragédie en 5 actes et en alexandrins – Représentée pour la première fois le 7 janvier 1678 à l'Hôtel de Bourgogne	14 838	1332
Corneille Thomas – Le Galant doublé – Comédie en 5 actes et en alexandrins représentée pour la première fois en 1659 à l'Hôtel de Bourgogne	18 056	1749
Corneille Thomas – Le Geôlier de soi-même – Comédie en 5 actes et en alexandrins représentée pour la première fois en 1655 au Théâtre du Marais	16 738	1852
Corneille Thomas - L'Inconnu – Comédie mêlée d'ornements et de musique - Représenté pour la première fois le 17 novembre 1675 au Théâtre Guénégaud	17 756	1683
Corneille Thomas – Maximian – Tragédie en 5 actes et en alexandrins représentée pour la première fois en février 1662 au Théâtre de l'Hôtel de Bourgogne	18 238	1360
Corneille Thomas – Médée – Tragédie en musique représentée par l'Académie royale de musique pour la première fois le 4 décembre 1693 au Théâtre du Palais-Royal	8 401	1042
Corneille Thomas – Persée et Démétrius – Tragédie en 5 actes et en alexandrins représentée pour la première fois en décembre 1662 à l'Hôtel de Bourgogne	19 345	1512
Corneille Thomas – Pyrrhus roi d'Epire – Tragédie en 5 actes et en alexandrins - représentée pour la première fois en 1661 à l'Hôtel de Bourgogne	18 913	1351
Corneille Thomas – Stilicon – Tragédie en 5 actes et en alexandrins représentée pour la première fois le 27 janvier 1660 à l'Hôtel de Bourgogne	18 903	1477
Corneille Thomas – Théodat – Tragédie en 5 actes et en alexandrins représentée pour la première fois le 18 novembre 1672 à l'Hôtel de Bourgogne	16 470	1389

Corneille Thomas – Timocrate – Tragédie en 5 actes et en alexandrins représentée pour la première fois en novembre 1656 à l'Hôtel du Marais	18 144	1537
Total Corneille Thomas	550 430	6692
Dancourt (Florent Carton) – Madame Artus – Comédie en vers représentée pour la première fois - le 8 Mai 1708 au Théâtre de la rue des Fossés Saint-Germain.	16 090	1 644
Dancourt (Florent Carton) – Céphale et Procris – Comédie en vers représentée pour la première fois - le 17 Octobre 1711 au Théâtre de la rue des Fossés Saint-Germain.	19 442	2 007
Dancourt (Florent Carton) – Le chevalier à la mode – Comédie en prose représentée pour au mois d'Octobre 1687	20 994	1 640
Dancourt (Florent Carton) – La comédie des comédiens – comédie en prose représentée pour la première fois - le 5 août septembre 1710 au Théâtre de la rue des Fossés Saint-Germain.	15 709	1 745
Dancourt (Florent Carton) – Les enfants de Paris – Comédie en vers représentée pour la première fois - le 18 Décembre 1699 au Théâtre de la rue des Fossés Saint-Germain	15 529	1 703
Dancourt (Florent Carton) – La trahison punie – Comédie en vers représentée pour la première fois le Novembre au Théâtre de la rue des Fossés Saint-Germain.	16 107	1 554
Total Dancourt	103 871	4 177
Desfontaines – Eurimédon ou l'illustre pirate – Tragi-comédie représentée pour la première fois en 1635	15 903	1 741
Desfontaines – L'illustre comédien ou le martyre de Saint-Genest – Tragédie représentée pour la première fois en 1635	14 324	1 686
Total Desfontaines	30 227	2 339
Desmarets de Saint-Sorlin (Jean). Aspasie. Comédie créée en 1636 au Palais Cardinal	13 369	1 422
Desmarets de Saint-Sorlain (Jean) - Mirame - Tragicomédie (représentée pour la première fois le 14 janvier 1641 pour l'inauguration de la grande salle du Palais Cardinal)	16 752	1 506
Desmarests de Saint-Sorlain Jean - Roxane - Tragicomédie (1639)	16 405	1 496
Desmarets de Saint-Sorlain (Jean) - Scipion - Tragicomédie (1638)	13 982	1 589
Desmarets de Saint-Sorlain (Jean) - Les visionnaires - Comédie (1637)	17 973	2 336
Total Desmarets	78 481	3 638
Donneau de Visé Jean - La mère coquette ou les amants brouillés - comédie représentée pour la première fois le 23 octobre 1665	9 631	1 144
Estoile (Claude de l') – La Belle esclave – Tragi-comédie - Représentée pour la première fois en 1643	14 711	1 728

Estoile (Claude de l') – L'Intrigue des filous – Comédie en cinq actes en alexandrins - 1646	15 191	2 077
Total Estoile	29 902	2 804
Genest Charles-Claude (Abbé de) – Pénélope – Tragédie représentée pour la première fois le 22 janvier 1684	10 638	1 388
Hauteroche (Noël Lebreton de). Crispin Musicien. Comédie en 5 actes en vers alexandrins. Représentée pour la première fois en 1671	18 079	1 856
Hauteroche (Noël Lebreton de). L'esprit follet - ou La dame invisible. Comédie en 5 actes en vers alexandrins. Représentée pour la première fois en 1684	14 395	1 722
Hauteroche (Noël Lebreton de) – Le Deuil - Comédie en un acte - Représentée pour la première fois en 1672.	6 159	988
Total Hauteroche	38 633	2 927
La Calprenède (Gautier Costes de) - Le comte d'Essex - Tragédie représentée pour la première fois à Paris en 1636.	14 913	2 511
La Calprenède (Gautier Costes de) - La mort de Mithridate - Tragédie représentée la première fois en 1635 à l'Hôtel de Bourgogne	15 246	2 635
Total La Calprenède	30 159	3 925
La Chapelle (Jean de) – Cléopâtre – Tragédie représentée pour la première fois le 12 décembre 1681	12158	1342
La Chapelle (Jean de) – Téléphonte – Tragédie représentée pour la première fois le 26 décembre 1682	12425	1343
La Chapelle (Jean de) – Zaïde – Tragédie représentée pour la première fois le 26 Janvier 1681 au Théâtre Guénégaud	13048	1415
Total La Chapelle	37631	2138
La Fontaine Jean de - Achille – Tragédie – Manuscrit des deux premiers actes d'une tragédie en vers inédite	5 423	964
La Fontaine (Jean de) – Daphné – Opéra -1674	6 397	1 081
La Fontaine Jean de - L'Eunuque - Comédie en vers imitée de Térence - 1654	17 441	1 813
Total La Fontaine	29 261	2 387
La Fosse – Manlius Capitolinus – Tragédie représentée pour la première fois le 8 janvier 1698 au Théâtre de la rue des Fossés Saint-Germain.	12 892	1 467
Mairet Jean - La Sylvie - Tragédie en 5 actes et en alexandrins - 1621?	19 813	2 347
Mairet Jean - La Sophonisbe - Tragédie en 5 actes et en alexandrins - 1629?	16 166	1 890
Total Mairet	35 979	2 979
Molière (Jean-Baptiste Poquelin dit).- La Jalousie du barbouillé : comédie - Représentée pour la première fois en 1660	3 501	698

Molière (Jean-Baptiste Poquelin dit).- Le Médecin volant - Représentée pour la première fois en 1659	3 876	662
Molière (Jean-Baptiste Poquelin dit).- L'Etourdi ou Les Contre-temps : comédie - Première représentation en 1658	18 671	2 172
Molière (Jean-Baptiste Poquelin dit).- Dépit amoureux : comédie - Représentée pour la première fois en 1659	16 242	1 898
Molière (Jean-Baptiste Poquelin dit).- Les Précieuses ridicules : comédie - Représentée pour la première fois le 18 novembre 1659	6 648	1 115
Molière (Jean-Baptiste Poquelin dit).- Sganarelle ou Le Cocu imaginaire : comédie - Représentée pour la première fois le 28 mai 1660	6 042	1 132
Molière (Jean-Baptiste Poquelin dit).- Dom Garcie de Navarre ou Le Prince jaloux : comédie - Représentée pour la première fois le 4 février 1661	17 049	1 615
Molière (Jean-Baptiste Poquelin dit).- L'Ecole des maris : comédie - Représentée pour la première fois le 24 juin 1661	10 536	1 401
Molière (Jean-Baptiste Poquelin dit).- Les Fâcheux : comédie - Représentée pour la première fois le 17 août 1661	7 922	1 378
Molière (Jean-Baptiste Poquelin dit).- L'Ecole des femmes : comédie - Représentée pour la première fois le 26 décembre 1662	16 625	1 955
Molière (Jean-Baptiste Poquelin dit).- La Critique de l'Ecole des femmes : comédie - Représentée pour la première fois le 1er juin 1663	8 610	1 113
Molière (Jean-Baptiste Poquelin dit).- L'Impromptu de Versailles : comédie - Représentée pour la première fois le 20 novembre 1663	7 168	987
Molière (Jean-Baptiste Poquelin dit).- Le Mariage forcé : comédie - Représentée pour la première fois le 29 janvier 1664	6 058	949
Molière (Jean-Baptiste Poquelin dit).- La Princesse d'Elide : comédie galante - Représentée pour la première fois le 8 mai 1664	11 333	1 393
Molière (Jean-Baptiste Poquelin dit).- Le Tartuffe ou L'Imposteur : comédie - Représentée pour la première fois le 12 mai 1664	18 271	1 951
Molière (Jean-Baptiste Poquelin dit).- Dom Juan ou Le Festin de Pierre : comédie - Représentée pour la première fois le 15 février 1665	17 452	1 757
Molière (Jean-Baptiste Poquelin dit).- L'Amour médecin : comédie - Représentée pour la première fois le 14 septembre 1665	6 147	989
Molière (Jean-Baptiste Poquelin dit).- Le Misanthrope : comédie - Représentée pour la première fois le 4 juin 1666	17 180	1 807
Molière (Jean-Baptiste Poquelin - dit).- Le Médecin malgré lui : comédie - Représentée pour la première fois le 6 août	9 317	1 206

1666		
Molière (Jean-Baptiste Poquelin dit).- Mélicerte : comédie pastorale héroïque - Représentée pour la première fois le 2 décembre 1666	5 540	922
Molière (Jean-Baptiste Poquelin dit).- Pastorale comique - Représentée pour la première fois le 5 janvier 1667	732	273
Molière (Jean-Baptiste Poquelin - dit).- Le Sicilien ou L'Amour peintre : comédie - Représentée pour la première fois le 5 janvier 1667	5 375	890
Molière (Jean-Baptiste Poquelin - dit).- Amphitryon : comédie - Représentée pour la première fois le 13 janvier 1668	15 117	1 757
Molière (Jean-Baptiste Poquelin dit).- George Dandin ou Le Mari confondu : comédie - Représentée pour la première fois le 18 juillet 1668	11 009	1 223
Molière (Jean-Baptiste Poquelin dit).- L'Avare : comédie - Représentée pour la première fois le 9 septembre 1668	21 033	1 981
Molière (Jean-Baptiste Poquelin dit).- Monsieur de Pourceaugnac : comédie-ballet - Représentée pour la première fois le 6 octobre 1669	11 803	1 613
Molière (Jean-Baptiste Poquelin dit).- Les Amants magnifiques : comédie - Représentée pour la première fois le 4 février 1670	11 983	1 422
Molière (Jean-Baptiste Poquelin - dit).- Le Bourgeois gentilhomme : comédie-ballet - Représentée pour la première fois le 14 octobre 1670	17 132	1 737
Molière (Jean-Baptiste Poquelin - dit).- Les Fourberies de Scapin : comédie - Représentée pour la première fois le 24 mai 1671	14 245	1 449
Molière (Jean-Baptiste Poquelin - dit).- La Comtesse d'Escarbagnas : comédie - Représentée la première fois le 2 décembre 1671	5 564	918
Molière (Jean-Baptiste Poquelin - dit).- Les Femmes savantes : comédie - Représentée la première fois le 11 mars 1672	16 863	1 901
Molière (Jean-Baptiste Poquelin - dit).- Le Malade imaginaire : comédie - Représentée la première fois le 10 février 1673	19 919	2 057
Corneille (Pierre) Molière (Jean-Baptiste Poquelin dit).- Psyché : tragédie-ballet en alexandrins - Musique de Lully - Représentée pour la première fois le 17 janvier 1671	4 816	839
Total Molière	369 779	8 096
Montfleury (Zacharie Jacob dit) - La mort d'Asdrubal - Tragédie – Représentée en 1647 sur le théâtre de l'Hôtel de Bougogne	14 016	1538
Montfleury (Antoine Jacob dit) - Le Comédien poète - Comédie jouée pour la première fois le 12 novembre 1673 à	13 899	1726

l'hôtel Guénégaud		
Montfleury (Antoine Jacob dit) - L'École des jaloux ou le Cocu volontaire - comédie en vers et en 3 actes (1664)	7 801	1188
Montfleury (Antoine Jacob dit) – La femme juge et partie – Comédie en 5 actes représentée pour la première fois le 2 mars 1669 à l'Hôtel de Bourgogne	14 515	1590
Montfleury (Antoine Jacob dit). La fille capitaine : comédie en 5 actes représentée pour la première fois à Paris en 1669	15 272	1626
Montfleury (Antoine Jacob dit) - Trasibule – Tragi-comédie représentée sur le Théâtre Royal de l'Hôtel de Bourgogne en 1664.	14 906	1184
Total Montfleury	80 409	3776
Pradon (Jacques) – Phèdre et Hippolyte – Tragédie représentée pour la première fois le 3 janvier 1677 à l'Hôtel Guénégaud.	15 392	1 406
Pradon (Jacques) - Scipion - Tragédie représentée pour la première fois le 22 février 1697 au Théâtre de la rue des Fossés Saint-Germain par la Comédie française	12 390	1 313
Total Pradon	27 782	1 901
Quinault Philippe – Alceste ou le triomphe d'Alcide - 19 janvier 1674 – Livret d'opéra sur une musique de J.-B. Lully	6 816	985
Quinault Philippe - Le feint Alcibiade - Tragi-comédie en cinq actes et en vers représentée pour la première fois en 1658	15 578	1389
Quinault Philippe - Amadis - Tragédie en musique représentée pour la première fois 15 janvier 1684	4 505	723
Quinault Philippe - Amalasonte - Tragi-comédies - représentée pour la première fois le 9 novembre 1657 à l'Hôtel de Bourgogne	15 009	1257
Quinault Philippe – L'amant indiscret ou le maître étourdi - Comédie en vers - 1654	14 986	1788
Quinault Philippe - Armide - Tragédie en musique sur un livret de Lully - Représentée pour la première fois le 15 février 1686	5 154	785
Quinault Philippe - Astrate Roi de Tyr - Tragédie présentée pour la première fois à l'Hôtel de Bourgogne dans les derniers jours de 1664 ou le début de 1665	15 833	1331
Quinault Philippe - Atys - Opéra sur une musique de J.-B. Lully - Représentée pour la première fois le 10 janvier 1676	7 959	971
Quinault Philippe – Belléphoron - Tragédie représentée pour la première fois le 1 janvier 1671 à l'Hôtel de Bourgogne	13 768	1324
Quinault Philippe – Cadmus et Hermione – Livret d'Opéra sur une musique de J.-B. Lully - Représentée pour la	5 881	931

première fois le 1er février 1673		
Quinault Philippe - Les coups de l'amour et de la fortune - Tragi-comédie en cinq actes et en vers représentée pour la première fois en 1655	14 536	1628
Quinault Philippe - Le fantôme amoureux - Tragi-comédie en cinq actes et en vers représentée pour la première fois en 1656	16 015	1494
Quinault Philippe – Isis – Livret d'Opéra sur une musique de J.-B. Lully - Représentée pour la premiere fois le 5 janvier 1677	6 343	926
Quinault Philippe – La mère coquette ou les amants brouillés – Comédie en vers - Représentée pour la première fois le 10 janvier 1665	16 130	1573
Quinault Philippe - Persée - Tragédie en musique représentée pour la première fois le 17 avril 1782	7 014	933
Quinault Philippe - Phaéton - Tragédie en musique représentée pour la première fois le 6 janvier 1683	6 685	916
Quinault Philippe - Proserpine - tragédie en musique représentée pour la première fois le 3 février 1680	7 402	887
Quinault Philippe - Psyché (Prologue à) : Tragédie ballet en alexandrins de Molière et Corneille - Musique de Lully - Représentée pour la première fois le 17 janvier 1671	1 184	369
Quinault Philippe - Les Rivales - Comédie en cinq actes et en vers - représentée pour la première fois à l'Hôtel de Bourgogne en 1653	15 483	1683
Quinault Philippe – Roland - Représentée pour la première fois le 8 janvier 1685 – Livret d'opéra sur une musique de J.-B. Lully	7 572	951
Quinault Philippe – Stratonice – Tragi-comédie représentée pour la première fois le 2 janvier 1660 au théâtre de l'Hôtel de Bourgogne.	16 566	1 271
Quinault Philippe - Thésée - Tragédie en musique ornée d'entrées de ballets - de machines et de changements de théâtre - représentée devant S. M. - à Saint-Germain en Laye - le onzième jour de janvier 1675	8 284	950
Total Quinault	228 691	4 599
Racine (Jean).- La Thébaïde ou Les Frères ennemis : tragédie - Représentée pour la première fois le 20 juin 1664	13 813	1 314
Racine (Jean).- Alexandre le Grand : tragédie - Représentée pour la première fois le 4 décembre 1665	13 864	1 372
Racine (Jean).- Andromaque : tragédie - Représentée pour la première fois le 17 novembre 1667	15 076	1 392
Racine (Jean).- Les Plaideurs : comédie - Représentée pour la première fois en novembre 1668	8 041	1 310
Racine (Jean).- Britannicus : tragédie - Représentée pour la première fois le 13 décembre 1669	15 387	1 637
Racine (Jean).- Bérénice : tragédie - Représentée pour la première fois le 21 novembre 1670	13 242	1 346

Racine (Jean).- Bajazet : tragédie - Représentée pour la première fois le 1er janvier 1672	15 297	1 507
Racine (Jean).- Mithridate : tragédie - Représentée pour la première fois le 23 décembre 1672	15 091	1 549
Racine (Jean).- Iphigénie : tragédie - Représentée pour la première fois le 18 août 1674	15 782	1 602
Racine (Jean).- Phèdre : tragédie - Représentée pour la première fois le 1er janvier 1677	14 394	1 775
Racine (Jean).- Esther : tragédie tirée de l'Ecriture Sainte - Représentée pour la première fois le 26 janvier 1689	11 147	1 656
Racine (Jean).- Athalie : tragédie tirée de l'Ecriture Sainte - Représentée pour la première fois le 5 janvier 1691	15 492	1 875
Total Racine	166 626	4 319
Régnard (Jean-François) – Démocrite – Comédie en cinq actes et en vers - 11 janvier 1700	15 249	1881
Régnard (Jean-François) - Le Distrait – comédie représentée pour la première fois le 2 décembre 1697	15 723	1900
Régnard (Jean-François) – Le Joueur – Comédie représentée pour la première fois le mercredi 19 décembre 1996	16 577	2084
Régnard (Jean-François) – Le légataire universel – Comédie en cinq actes et en vers - 9 janvier 1708	16 831	2010
Régnard (Jean-François) - Les Ménechmes ou les jumeaux. Comédies en cinq actes et en vers - précédée d'un prologue en vers libres - représentée pour la première fois le vendredi 4 décembre 1705.	17 553	2048
Total Régnard	81 933	4536
Rotrou (Jean de) - La belle Adelphe – Comédie représentée pour la première fois en 1635 ou 1636	17 556	1 881
Rotrou (Jean de) - La bague de l'oubli - comédie représentée pour la première fois en 1629 ?	11 267	1 503
Rotrou (Jean de) – Cléagénor et Doristée – Tragi-comédie représentée pour la première fois en 1634	15 409	1 798
Rotrou (Jean de) – Les Sosies – Comédie représentée pour la première fois en 1636	16 123	1 993
Rotrou (Jean de) – Venceslas – Tragi-comédie représentée pour la première fois en 1647	16 828	1 805
Total Rotrou	77 183	3 634
Ryer Pierre de – Alcionée Ou le combat de l'honneur et de l'amour - Représentée pour la première fois en 1639 au Théâtre du Marais.	14 914	1 399
Ryer Pierre de – Esther – Tragédie représentée pour la première fois en 1640	16 555	1 519
Total Ryer	31 469	1 923
Scarron – Dom Japhet d'Arménie – Comédie représentée pour la première fois en 1652 au Théâtre de l'Hôtel de Bourgogne.	14 238	2 014
Scarron – Le Jodelet duelliste – Comédie représentée pour la première fois en 1646 au théâtre de l'Hôtel de Bourgogne	16 869	1 835

Scarron – Le Jodelet ou le Maître valet – Comédie représentée pour la première fois en 1648	16 703	1 860
Total Scarron	47 810	3 449
Scudéry (Georges de) - L'Amour tyrannique – Tragi-comédie représentée pour la première fois en 1638 au Jeu de Paume du Marais	17 069	1658
Scudéry (Georges de) – Arminius ou les frères ennemis – Tragédie représentée pour la première fois en 1642.	15 631	1501
Scudéry (Georges de) – Eudoxe – Tragi-comédie représentée pour la première fois en 1633 au Jeu de Paume de La Fontaine	19 442	1668
Scudery (Georges de) – Ligdamon et Lidias ou la ressemblance – Tragi-comédie représentée pour la première fois en 1630 à l'Hôtel de Bourgogne	20 511	2532
Scudery (Georges de) – Le Prince déguisé – Tragi-comédie représentée pour la première fois en 1635 à l'Hôtel de Bourgogne	13 986	1748
Total Scudéry	86 639	3706
Villiers (Claude Deschamps de) – Le festin de pierre ou le fils criminel – Tragi-comédie traduite de l'italien par Le Sieur de Villiers - Représentée pour la première fois en août 1659 au théâtre de l'Hôtel de Bourgogne	15 571	1 981
Total 235 pièces	3 253 063	