

Université de Paris X-Orsay
Laboratoire d'Informatique pour la Mécanique et les Sciences de
l'Ingénieur, LIMSI
Séminaire du Groupe Langues Information Représentations
(13 janvier 2004)

Corneille et Molière

Dominique Labbé
dominique.labbe@iep.upmf-grenoble.fr
<http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/Labbe>
(Institut d'Etudes Politiques - BP 48 - F 38040 Grenoble Cedex)

Où trouvera-t-on un poète qui ait possédé à la fois tant de grands talents (...) capable néanmoins de s'abaisser, quand il le veut, et de descendre jusqu'aux plus simples naïvetés du comique, où il est encore inimitable.
(Racine, *Eloge de Corneille*)

En décembre 2001, le *Journal of Quantitative Linguistics* a publié un article de C. Labbé et moi-même (voir références complètes dans la bibliographie à la fin de ce texte). Nous y présentons un outil d'attribution d'auteur (la distance intertextuelle) avec comme exemple les principales pièces de Molière écrites par Corneille (la liste des pièces attribuées à Corneille est donnée en annexe VI). Un essai, destiné au grand public et paru en juin 2003 (*Corneille dans l'ombre de Molière*), raconte l'histoire de cette recherche et répond aux principales questions. Mon exposé ne reviendra pas sur le cheminement ; il sera consacré pour l'essentiel à la méthode, c'est-à-dire à la distance intertextuelle et aux techniques de classification automatique. Ces méthodes et les résultats obtenus soulèvent évidemment de nombreuses questions. Je vous propose d'examiner celles-ci : comment vérifier leur valeur ? Que prouvent nos résultats ? Enfin, comment notre recherche a-t-elle été accueillie et quelles objections nous a-t-on opposées ? Naturellement, au cours du débat, vous pourrez y ajouter toutes les interrogations qui vous viendront à l'esprit.

I. Distance intertextuelle et attribution d'auteur

L'ensemble du raisonnement est présenté dans les articles parus en décembre 2001 dans le *Journal of Quantitative Linguistics* et, en français, dans le numéro 2 de la revue *Corpus* (décembre 2003). Il s'agit de répondre, le plus exactement possible, à la question de savoir comment mesurer la proximité ou l'éloignement de deux ou plusieurs textes, les uns par rapport aux autres, en considérant tous leurs mots.

Le calcul

Soit deux textes A et B avec :

— N_a et N_b longueurs de A et B en mots ou "occurrences" (les anglo-saxons disent "tokens") ;

— V_a et V_b le nombre de "mots différents" ou "vocables" (les Anglo-saxons disent "types" : c'est l'étendue du vocabulaire) ;

— F_{ia} et F_{ib} la fréquence absolue d'un vocable i dans les textes A et B.

Si les deux textes ont la même taille, la réponse à la question de leur distance sera donnée par :

$$(1) D_{(A,B)} = \sum_{i \in (A,B)} |F_{ia} - F_{ib}| \text{ avec } N_a = N_b$$

Pour tous les vocables appartenant à A et/ou B, on calcule la différence de leurs fréquences et on fait la somme de ces différences. Cette somme possède toutes les propriétés d'une distance euclidienne (voir le début de notre article dans *Corpus*). Comment procéder quand les deux textes ont des tailles différentes, tout en conservant autant que possible ces propriétés ?

Soit $N_a < N_b$. Nous proposons de simuler le tirage aléatoire (sans remise), dans une urne contenant les N_b mots de B, d'un très grand nombre d'échantillons de taille N_a . Considérons un vocable i de fréquence F_{ib} dans B. Combien de fois figurera-t-il dans un quelconque de ces échantillons ? Cette "espérance mathématique" est :

$$E_{ia(u)} = F_{ib} * U \text{ avec } U = \frac{N_a}{N_b}$$

Dans la formule (1), en remplaçant la fréquence de chacun des vocables de B par cette espérance mathématique, on obtient la distance intertextuelle absolue :

$$(2) D_{(A,B)} = \sum_{i \in (A,B)} |F_{ia} - E_{ia(u)}|$$

On en tire la définition de la distance entre deux textes : *Somme des différences entre les fréquences de tous les vocables du plus petit texte comparés à ceux tous les échantillons aléatoires possibles de la taille du plus petit que l'on peut extraire du plus grand*. De manière plus littéraire, on peut dire que l'on projette sur le petit texte une réduction du plus grand, ce qui permet de les comparer sans biais.

La distance relative est :

$$(3) D_{rel(A,B)} = \frac{\sum_{i \in (A,B)} |F_{ia} - E_{ia(u)}|}{\sum_{i \in A} F_{ia} + \sum_{i \in B} E_{ia(u)}}$$

Cette dernière mesure varie uniformément entre un minimum de zéro — distance nulle : mêmes vocables et mêmes fréquences relatives — et un maximum de 1 : tous les mots sont différents. On pourra ainsi comparer les résultats obtenus sur un grand nombre de textes.

Remarques et limites du calcul

L'idée n'est pas extraordinaire. Pourtant, dans le passé, elle ne semble pas avoir été explorée, même si plusieurs chercheurs ont étudié des choses assez voisines bien que toujours appliquées à des sélections de mots et non à la totalité des textes (par exemple, l'article récent de l'Australien Burrows qui est un des spécialistes de l'attribution d'auteur).

A ce stade de l'exposé, trois remarques s'imposent.

— ce calcul exige une stricte normalisation des graphies entre les textes. En effet, un traitement sur les formes graphiques "brutes" classera les textes selon les conventions graphiques propres à leur genre (par exemple, les majuscules initiales de vers différencieront tous les poèmes de tous les textes en prose). Il faut également lemmatiser les textes car le français est une langue fortement flexionnelle, avec de très nombreuses homographies (dans tout texte en français, plus du tiers des mots peuvent être rattachés à plus d'une entrée de dictionnaire et ce sont les plus fréquents : le, être, avoir...) De plus, la lemmatisation représente le seul moyen de diminuer les effets parasites des basses fréquences (voir sur ce point notre article paru dans *Corpus*) ;

— en théorie, les différences de taille sont totalement neutralisées puisque, aux arrondis près, la somme des E_{ia} sera égale à N_a . En fait, les nombreuses expériences que nous avons conduites montrent que ce n'est pas totalement le cas. Lorsque les différences de taille sont trop grandes, les plus longs textes ont un certain avantage par rapport aux autres, avantage qui tient à l'étendue de leur vocabulaire, à la "spécialisation lexicale" — qui s'élève généralement avec la longueur des textes — et à la déformation de la distribution des fréquences avec l'augmentation de cette longueur. Dans une collection de textes de tailles très différentes, ce sont pratiquement toujours les textes les plus longs qui se trouvent au centre de la distribution et les plus courts qui sont les plus décalés. Pour compenser partiellement ce biais, il est recommandé de :

— ne pas appliquer le calcul sur des textes trop courts (en tous cas, pas moins de 1.000 mots) et de s'en tenir à une échelle inférieure à 1:10 dans les différences de taille ;

— éliminer du calcul les vocables de B dont l'espérance mathématique dans A n'est pas au moins égale à 1 et de ne pas faire entrer dans la distance les écarts inférieurs à .5.

Ces considérations ne seront pas d'avantage développées car elles ne concernent pas l'expérience sur Corneille et Molière, comme on pourra s'en assurer en consultant les deux premières annexes. Le plus petit texte attribué à Corneille (*Mélicerte*) dépasse 5 000 mots (à ce niveau, les perturbations dues aux basses fréquences dans les textes courts sont négligeables) et la plus longue pièce (*l'Avare*) en compte 21 000. Ce sont les dimensions idéales pour l'expérience.

Ces logiciels ont été appliqués à de grandes collections de textes français, du moins ce qui existe sur support électronique : collections de journaux, discours politiques, brochures techniques aussi bien qu'œuvres littéraires. En ce qui concerne ces dernières, nous utilisons les textes accessibles sur le web, soit près d'un millier d'œuvres, dans une trentaine de bases comme Gallica ou Abu, pour ne citer que les plus célèbres, auxquelles s'ajoutent des textes qui nous ont été fournis par différents chercheurs. Il se trouve que, parmi toutes ces œuvres, Corneille et Molière illustrent parfaitement l'intérêt de la distance intertextuelle...

Que mesure la distance intertextuelle ?

Sous réserve des précautions mentionnées ci-dessus, la distance intertextuelle mesure donc exactement la plus ou moins grande ressemblance entre deux textes. Quatre facteurs expliquent cette ressemblance :

— le **genre**. On ne parle pas comme on écrit ; la fiction romanesque a ses codes qui ne sont pas ceux du théâtre, etc. Le carcan imposé par le genre est plus ou moins rigide. Evidemment, on écarte ici les genres "techniques" où l'auteur est contraint d'adopter une expression

impersonnelle et de suivre des canevas préétablis... C'est pour cela que nos contradicteurs ont prétendu que le genre comédie — spécialement celui de la comédie en alexandrins — était, au XVIIIe, tellement contraignant qu'il est impossible de reconnaître l'auteur. C'est évidemment ridicule : si le *Tartuffe* avait été rédigé comme un compte rendu de l'Académie des Sciences, le succès était peu probable ! Mais enfin, l'argument sera examiné plus bas en détail ;

— le vocabulaire de **l'époque**. L'œuvre de Corneille s'étend sur plus de 40 ans. Celle de Molière sur plus de 15 ans. C'est d'ailleurs le laps de temps qui sépare la *Suite du Menteur* (1643), dernière comédie connue de Corneille, de l'*Etourdi* première comédie en vers de Molière présentée en 1658. Sur de telles étendues de temps, le style et le vocabulaire d'un auteur évoluent nécessairement. D'ailleurs, il est lui-même pris dans le flux qui change lentement la langue, au moins dans son lexique.

— le **thème** traité. Chaque thème a un vocabulaire propre. Ce sont d'abord des noms de lieux, de personnes et une série de substantifs particuliers. Par exemple, la tragédie romaine verra nécessairement apparaître l'empereur, le sénat, le forum, les légions et une foule de choses qu'on ne trouvera pas dans une pièce inspirée de la mythologie grecque...

— enfin et surtout : **l'auteur**...

Pour rechercher la paternité d'un texte anonyme ou dont l'auteur est contesté, il faut donc le comparer à d'autres dont la signature n'est pas contestée, ayant été écrits à la même époque et traitant de thèmes voisins, dans un même genre (poésie, prose, roman, théâtre...). Ce point est important. En l'état actuel de la technique, on ne peut comparer que ce qui est comparable : le théâtre avec le théâtre, une lettre avec de la correspondance...

Nous avons appliqué ce calcul à plusieurs milliers de textes de toutes origines (romans, pièces de théâtre, poésie, articles de presse, discours politiques, entretiens...). Cela a permis de confirmer la validité du raisonnement et d'étalonner une échelle de la distance.

Une échelle de la distance

Pour une taille d'échantillons comprise entre 5 000 et 20 000 mots :

— une valeur inférieure ou égale à 0.20 ne se rencontre jamais chez des auteurs différents ;
— entre 0.20 et 0.25, il est pratiquement certain que l'auteur est le même. Sinon, les deux textes ont été écrits à la même époque, dans un même genre, sur un sujet et avec des arguments semblables. Ce cas se rencontre parfois dans les articles de presse, à propos d'un 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 appartenant à deux auteurs différents, il est très probable que le second s'est "inspiré" du premier (dans l'ordre chronologique)...

— au-dessus 0.25, on entre dans une zone "grise" où deux hypothèses sont envisageables : un même auteur et des thèmes différents ou deux auteurs contemporains traitant, dans un même genre, un thème proche avec leur style propre... De telle sorte que, plus on s'élève au-dessus de ce seuil, plus il sera difficile d'attribuer la paternité d'un texte anonyme à l'auteur considéré sans que, pour autant, cette paternité puisse être rejetée ;

— au-dessus de 0.40 les auteurs sont certainement différents ou bien, pour un même auteur, les textes sont de genres très éloignés, par exemple : oral et écrit.

Naturellement,

- ces chiffres ne valent que pour des textes dont les graphies ont été rigoureusement normalisées et dont chaque mot a été lemmatisé ;
- le calcul doit suivre exactement l'algorithme présenté et respecter l'échelle des tailles signalée ci-dessus ;
- il ne faut pas considérer ces chiffres comme des seuils mais comme des bornes balisant un continuum ...
- quand on analyse de vastes collections de textes, l'interprétation de ces chiffres doit être assistée par des automates de classification.

Les classifications

La superposition des pièces de Corneille et de Molière, prises deux à deux, donne un tableau de plus de 2 200 cases... quantité qu'il est impossible de maîtriser d'un coup d'œil de manière synthétique. Et l'obstacle devient inextricable quand on examine plusieurs centaines de textes. Au passage, rappelons que nous mettons au point des outils capables de traiter les grandes collections de textes. Corneille, Molière, Racine étaient des fichiers d'essai parmi beaucoup d'autres...

Il existe des techniques de classification assez sophistiquées qui permettent de dire quel est le "meilleur classement possible" dans ces grandes collections de textes. Les guillemets sont ici nécessaires car, malgré la puissance des ordinateurs, aucune de ces techniques n'est parfaite. Nous en avons utilisé deux : la plus classique, la classification automatique ("cluster analysis"), et la plus moderne : l'analyse arborée ("tree-analysis"). Pour cette dernière, il s'agit plus précisément de la technique des arbres "valués" issue des travaux de X. Luong (voir aussi : Barthélémy et Guénoche). Dans le cas de Corneille et Molière, ces deux techniques aboutissent exactement aux mêmes conclusions (voir notre article de décembre 2001). L'annexe III reproduit l'arbre issu de la seconde technique.

A ce propos, une remarque : un graphe n'apporte pas de "preuve" en lui-même. Et ceci d'autant plus que le danger de manipulation est réel. Par exemple, vous verrez sur internet et dans certaines revues, qu'un grand nombre de figures ont été mises en circulation contre nous. Est-il nécessaire de préciser qu'on ne peut pas avoir accès aux données, aux matrices des distances (et à la manière dont elles sont calculées) ni aux algorithmes de classification dont ces figures sont issues. Il faut "faire confiance" à ces gens... A l'inverse, notre travail est entièrement vérifiable et transparent : nous avons mis dans le domaine public nos fichiers, nos données et nos algorithmes. Et à l'été 2000, quand nous préparions notre article pour le *Journal of Quantitative Linguistics*, les données sur Corneille et Molière ont été envoyées à X. Luong après avoir été anonymées. Sans savoir sur quoi portait ces chiffres, X. Luong a réalisé, avec son propre programme, l'arbre qui figure en annexe IV (c'est la raison pour laquelle, sur cet arbre, les titres des pièces sont remplacés par des numéros).

Dans cette figure, les pièces forment les feuilles terminales. Leur proximité est mesurable par la longueur du chemin qu'il faut parcourir, le long des branches et du tronc de l'arbre, pour aller de l'une à l'autre. Ainsi les deux textes les plus éloignés sont le prologue de *Psyché* écrit par Quinault pour Molière (n° 36) et la première pièce de Molière (37).

L'analyse arborée isole très clairement trois ensembles : en bas, les pièces de Corneille, au milieu les pièces en vers de Molière et en haut, avec une dispersion nettement plus grande, les pièces en prose.

Mais il y a quelques "anomalies", notamment :

— en bas, les contributions respectives de Corneille et de Molière à une même pièce, *Psyché*, se rejoignent avec *Dom Garcie de Navarre* (traditionnellement attribuée à Molière) et avec le troisième acte de la *Comédie des Tuileries*, écrite par Corneille pour le Cardinal en... 1634 (soit 37 ans avant *Psyché* !) !

— au milieu des pièces en vers de Molière, figurent les deux *Menteurs* de Corneille. La chose est d'autant plus surprenante que les *Menteurs* ont été créés 15 ans avant *l'Etourdi* (première pièce en vers de Molière) et 30 ans avant les *Femmes savantes*.

L'attribution des pièces de Molière

Une fois constatées ces "anomalies", il faut revenir à la matrice des distances. Autrement dit, la classification automatique ou l'analyse arborée — comme toutes les techniques "exploratoires" — ne sont là que pour assister le décryptage des grandes masses de données. Elles permettent de poser des questions, de formuler des hypothèses. Ceci fait, il faut retourner aux données pour répondre à ces questions, vérifier ces hypothèses.

Examinons donc en détail les deux principales questions soulevées par l'annexe III.

— Premièrement, les *Menteurs* et les pièces en vers de Molière.

On trouvera en annexe II, les distances séparant ces deux *Menteurs* de toutes les pièces de Molière. Etant donné le laps de temps considérable séparant ces œuvres, les distances obtenues sont les plus faibles que l'on puisse constater *chez un auteur unique*. Ainsi, chez Corneille, toutes les pièces éloignées de 15 ans sont séparées par des distances comprises entre 0.20 et 0.25. Il en est de même entre les premières et les dernières comédies en vers attribuées traditionnellement à Molière. Ce sont des distances faibles que l'on rencontre rarement chez un même auteur quand l'œuvre s'étend sur de longues périodes... D'ailleurs, les dernières lignes du tableau II indiquent que les deux *Menteurs* sont plus proches de l'œuvre entière de Molière que de celle de Corneille. Ils le sont même si l'on ne considère que les comédies en vers de Corneille qui ont précédé ces deux *Menteurs* (de *Mélite* à *l'Illusion comique*).

Les deux *Menteurs* sont donc les sœurs aînées de toutes les pièces en vers de Molière et, très probablement de *Dom Juan* et *l'Avare*. Pour ces deux dernières s'ajoute en effet une différence de genre (vers et prose) qui devrait engendrer des distances nettement plus élevées encore.

— Deuxièmement, la position curieuse de *Dom Garcie de Navarre* et de *Psyché*. On a souvent souligné l'étrangeté de *Dom Garcie* dans l'œuvre de Molière. Effectivement, la distance la plus faible avec une autre comédie signée par Molière est de 0.243 (le *Dépit amoureux* qui est contemporaine de *Dom Garcie*). Pour toutes les autres pièces de Molière, la distance avec *Dom Garcie* (comme avec *Psyché*) est toujours supérieure à 0.25. En revanche, l'annexe IV montre clairement que *Dom Garcie* et *Psyché* sont sœurs et que leurs autres sœurs sont toutes les œuvres ultimes de Corneille *contemporaines de (ou postérieure à) leur création*. Le calcul ne laisse aucun doute : toutes ces pièces ont un seul père (Corneille). Pour *Psyché*, on en est sûr grâce à l'indiscrétion du premier éditeur...

On en conclut que :

— *Dom Garcie* (1661) et *Psyché* (1671) descendent d'*Andromède* (1650) et de la *Toison d'or* (1661) :

— les autres pièces en vers de Molière, de même que *Dom Juan* et *l'Avare*, descendent des *Menteurs* (1642-1643).

Et il faut souligner encore une fois qu'il n'existe, dans la littérature française, aucun autre exemple de tels groupements croisés entre deux auteurs différents. Des distances aussi faibles se rencontrent même rarement dans une œuvre unique quand elle comporte autant de textes et que sa création s'étale sur une aussi longue période. Tout indique donc un auteur unique pour ces pièces : Corneille (pour le détail, voir annexe VI).

Excusez-moi d'avoir été un peu long, mais ces calculs et ce raisonnement forment l'élément essentiel du débat.

2. Tests de la méthode et portée des résultats

Puisque ces méthodes sont relativement nouvelles — seule la puissance des calculateurs modernes les rend possibles —, il est légitime de discuter des moyens de les valider.

En premier lieu, les sceptiques peuvent essayer de démontrer que notre formule et nos raisonnements sont faux. Nos critiques se sont effectivement tournés vers un certain nombre de statisticiens qui ne nous veulent pas du bien. Aucun n'a été capable d'identifier une erreur. D'ailleurs c'est à cela que sert la publication dans une revue à comité de lecture : en plus du rédacteur en chef, deux relecteurs anonymes passent à la moulinette les formules et les raisonnements et ne font généralement pas de cadeau ! Notre travail a franchi avec succès ce barrage, ce qui nous donne une certaine tranquillité d'esprit sur ce point.

Ceci acquis, il faut mettre expérimentalement la méthode à l'épreuve de trois manières

Une expérience cruciale

En premier lieu, on met sur pied des expériences "cruciales". Par exemple, on donne à deux auteurs un même thème à traiter dans un genre identique : théâtre, poésie, roman... On leur impose de rendre leur copie à la même date et on les isole pour qu'ils ne copient pas l'un sur l'autre. Ainsi se trouvent neutralisés trois des quatre facteurs qui influencent la distance (le genre, le thème et l'époque), ce qui permet d'isoler l'action du quatrième facteur : l'auteur.

Cela paraît difficile à organiser ? Eh bien ! deux des plus grands écrivains français se sont prêtés à l'expérience : Corneille et Racine ont écrit chacun une tragédie sur l'amour impossible entre un empereur romain (Titus) et une reine orientale (Bérénice). Ils l'ont fait en même temps, en alexandrins et en respectant les fameuses "règles" qui enserraient la création théâtrale, spécialement la tragédie. Et tous deux, dit-on, avaient en tête la même référence : Louis XIV et Henriette d'Angleterre, sa belle sœur. Le lieu de l'action et les personnages étaient donc identiques ainsi que les obstacles à cet amour partagé (d'où un vocabulaire commun important...) Superposons ces deux pièces : leur distance (0,256) est plus élevée que toutes celles constatées entre les pièces en vers de Molière et les deux Menteurs de Corneille alors que ces dernières sont séparées par un laps de temps important, que les thèmes sont toujours un peu différents et que le genre "comédie" en vers est nettement moins contraignant que celui de la grande tragédie en alexandrins.

Pour parfaire l'expérience, on peut comparer ces deux pièces avec leurs sœurs dans l'ordre chronologique (comme nous venons de le faire avec *Dom Garcie* et *Psyché*) : Annexe IV. Au passage, ce tableau permet de répondre à la question de savoir qui, de Corneille ou de Racine, a pu influencer l'autre. Mais c'est un autre sujet.

Une proposition simple se déduit de cette annexe IV : les distances entre des textes appartenant à un même genre et écrits à une même époque, *sur des thèmes différents*, par un

même auteur sont systématiquement et nettement inférieures à celles séparant deux auteurs différents, *même quand ils écrivent au même moment, sur un thème identique...* Pour l'instant, nous n'avons trouvé aucune exception à cette loi dans la littérature française accessible sur support électronique.

N'est-ce pas la plus belle contre-épreuve que nous puissions fournir à nos détracteurs ?

Quel est le poids du genre ?

Deuxièmement, on peut généraliser l'expérience précédente et voir comment se comporte la distance quand on neutralise un, deux ou trois des facteurs explicatifs. Par exemple, selon nos détracteurs, la proximité, entre les deux *Menteurs* et les principales pièces de Molière, prouverait simplement la profonde unité du genre "comédie en vers". A cela, il est facile de répondre qu'il faudrait expliquer en quoi ce genre "comédie en vers" serait un moule plus contraignant que la tragédie en alexandrins. Mais, là aussi, il y a un contre-exemple : les *Plaideurs* de Jean Racine (1668). Cette comédie en alexandrins est contemporaine des œuvres de Molière. En comparant ces pièces aux *Plaideurs*, on mesure donc l'influence de deux facteurs (le thème et l'auteur) à l'exclusion des deux autres (genre et époque). Si nos détracteurs ont raison, les distances entre les *Plaideurs* et les pièces en vers de Molière doivent être inférieures à celles mesurées entre les *Menteurs* et ces mêmes pièces (puisque dans ce cas, trois facteurs agissent : auteur, thème et époque). Dans la dernière colonne de l'annexe II, vous verrez qu'aucune distance entre les *Plaideurs* et Molière n'est inférieure à 0.25 (comme permettait de le prédire l'échelle présentée ci-dessus). La plus faible est obtenue avec *l'Ecole des femmes* (0.26) puis *l'Etourdi*, le *Dépit amoureux*, *l'Avare* (0.27). Cette dernière pièce est en prose, mais elle est produite la même année que les *Plaideurs*. Il est intéressant de constater que, avec un goût assez sûr, Racine s'inspire des pièces écrites par Corneille — en vers comme en prose — et qu'il choisit les plus inoffensives...

Naturellement, si certaines personnes estiment que ces exemples ne prouvent rien : qu'elles indiquent les pièces qui, selon elles, mettront ce raisonnement en échec. Nous les traiterons volontiers et de manière totalement transparente : elles auront tous les fichiers correspondants et des experts neutres pourront examiner toute la chaîne de traitement. Cette proposition a été faite publiquement il y a plus d'un an. Nous attendons toujours. Voilà encore une preuve expérimentale : si ces critiques étaient si sûrs d'eux, je devrais crouler sous les pièces !!!

Une expérience en "double aveugle"

Troisièmement, pour valider expérimentalement une théorie scientifique, il y a des procédures plus élaborées comme celle du "double aveugle" bien connue en médecine : on mélange plusieurs populations, on glisse des placebos au milieu du principe actif et on traite le tout de manière anonyme... En 2001, un professeur de lettres (E. Brunet) a accepté de réaliser cette expérience. Son hostilité à notre thèse élimine le soupçon d'une complicité. Dans les 50 textes anonymés qu'il a tiré de la littérature française, nos algorithmes ont repéré, sans aucune erreur, tous ceux qui appartenaient à un même livre et ils ont isolé toutes les "chimères" imaginées par notre critique à titre de "placebos". Une seule limite, d'ailleurs prévue par le modèle : on échoue parfois à marier les œuvres d'un même auteur quand elles sont séparées par un laps de temps important. En effet, E. Brunet connaissait l'importance du

temps : il a donc logiquement choisi les première et dernière œuvres des écrivains qu'il avait sélectionnés.

Il avait été convenu que nos deux compte rendus seraient publiés conjointement dans une revue. E. Brunet ayant renié sa parole et fait pression sur la revue pour empêcher également la publication de mon texte, je l'ai placé sur ma page personnelle (références en tête de ce document). Personne ne l'a contesté. Nous avons proposé à tous nos critiques de réaliser cette expérience avec les textes de leur choix... Nous attendons toujours !

Si ces critiques étaient de bonne foi se conduiraient-ils ainsi ?

Finalement, leur attitude, depuis plus d'un an, est la meilleure preuve que nous puissions administrer : ils *savent* que notre méthode est solide et que nous avons travaillé sérieusement.

Que prouvent ces résultats ?

Permettez-moi une comparaison avec la police scientifique. Si les empreintes digitales ou l'ADN relevés sur le lieux du crime ne sont pas celles du suspect, alors il est innocenté sans discussion. Mais si ces empreintes ou cet ADN sont semblables à ceux de ce suspect ? En l'absence de tout autre indice, je doute que l'accusé puisse être condamné. Tout tribunal, agissant rationnellement, acquittera au "bénéfice du doute" (et ceci, malgré la très grande précision de la technique de l'ADN, précision difficilement concevable en matière textuelle).

Alors : avons-nous d'autres indices pour conforter notre raisonnement ?

Bien sûr ! Et ces indices sont nombreux et concordants. S'il en avait été autrement, nous n'aurions pas publié nos résultats, même à titre de curiosité.

Distinguons ceux qui se rattachent aux textes et ceux qui sont tirés de l'histoire des deux hommes.

En ce qui concerne les textes, est-il nécessaire de rappeler qu'avant nous, P. Louÿs, H. Poulaille ou H. Wouters avaient relevé de nombreuses ressemblances troublantes entre les deux œuvres. Ces trois personnes sont les véritables "découvreurs". Notre travail n'est qu'une modeste contribution à leur démonstration, du même ordre que celle que la "police scientifique" peut apporter aux enquêteurs.

Nos conclusions sont renforcées par deux indices statistiques qui ont une force probablement plus grande encore que la distance intertextuelle, car ces calculs reposent non plus sur le comptage des mots isolés mais sur leurs combinaisons.

En premier lieu, nos expériences ont mis en lumière le caractère très personnel des combinaisons "verbe + verbe" (sur le modèle "vouloir dire", "savoir faire"...) qui révèlent un certain rapport au monde et à la création. Pour réaliser l'expérience, il faut simplement embrasser une grande quantité de textes — pas question ici d'examiner chaque pièce individuellement — et admettre que, parfois, certains auteurs changent de "vision du monde" au cours de leur vie. Donc, les mêmes combinaisons signalent pratiquement à coup sûr un auteur unique ; des combinaisons différentes ne permettent pas de conclure à deux auteurs différents, sauf en cas de corpus homogènes, contemporains et pas trop étirés dans le temps. Cette limite admise, nous n'avons, pour l'instant, jamais rencontré deux auteurs différents présentant les mêmes préférences avec des densités d'emploi voisines... sauf Corneille et Molière (annexe VII). Le tableau donne aussi les résultats obtenus sur Racine : à ce niveau de diversité, une "coincidence" est absolument invraisemblable. Au passage, la lemmatisation est ici indispensable sinon les flexions du "pseudo-auxiliaire" et les mots outils insérés entre les

deux verbes — par exemple : "(ne) vouloir (rien) dire" — empêcheront de retrouver ces constructions...

Mieux encore : l'analyse des réseaux sémantiques permet d'accéder au sens spécifique que chaque auteur donne aux principaux mots qu'il emploie. Grâce à cette étude, nous pouvons affirmer que, chez Corneille et Molière, les principaux vocables ont le même sens, ou plutôt, que ceux de Molière s'inscrivent comme un sous-ensemble dans ceux de Corneille. Mon livre développe le cas du mot "amour" qui est particulièrement évocateur (voir également notre conférence publique à Lausanne en mars 2000 : <http://www.cavi.univ-paris3.fr/lexicometrica>). Là encore, il n'existe aucun autre cas semblable dans les œuvres accessibles sur support électronique...

Quand à l'histoire, H. Poulaille et H. Wouters ont présenté des dossiers solides concernant la vie des deux hommes et notamment ce fait troublant, également souligné par Louÿs : pour la période antérieure à 1658, il ne reste de Molière que trois très mauvaises farces (*La jalousie du Barbouillé*, *Gorgibus dans le sac* et *Le Médecin volant*). En 1658, Molière s'installe pendant 6 mois à Rouen (où habitent les deux frères Corneille). Après ce séjour, c'est un nouvel auteur complètement différent qui émerge. A partir de 1662, Corneille rejoint Paris et, alors, les chefs d'œuvre se succèdent (voir les dates des pièces de Molière en annexe 2)...

Ajoutons simplement à cette troublante chronologie, et parmi beaucoup d'autres "coïncidences", que :

— trois "collaborations" sont indiscutables. Deux éditeurs ont désigné Corneille comme étant l'auteur du *Dépit amoureux* et des trois quarts de *Psyché* (1671) (voir annexe VIII et IX). Après la mort de Molière, Thomas Corneille a mis *Dom Juan* en vers ;

— l'auteur le plus joué par Molière (après Molière !) est... Corneille. De son côté, celui-ci a confié la création de plusieurs de ses tragédies à Molière. Il est donc peu probable que les deux hommes aient été fâchés comme le prétendent nos critiques sans apporter la moindre preuve de cette brouille ;

— enfin, pour mémoire, le témoignage d'un contemporain :

*Dans ce sac ridicule où Scapin s'enveloppe
Je ne reconnais plus l'auteur du Misanthrope
(Boileau, Art poétique, chant III, 1674)*

Le dilemme de Boileau a une solution simple : Molière, qui jouait le personnage de Scapin, n'est pas l'auteur du *Misanthrope*. La statistique fait plus que suggérer cette évidence !

Que prouve ce faisceau concordant d'indices ?

Cela prouve simplement que Corneille a tenu la plume. Or, ce fait n'intéresse guère le public qui se pose d'autres questions. De Molière ou de Corneille, qui a eu l'idée du Tartuffe ? Corneille peut-il avoir été une simple plume de l'ombre mettant en forme les idées d'un autre ? Cette hypothèse est d'ailleurs suggérée par l'avertissement placé en tête de *Psyché*.

La statistique ne peut trancher de telles questions. Tout au plus, peut-on indiquer que les *Menteurs*, qui sont incontestablement de Corneille, sont bien les sœurs aînées des grandes comédies de Molière. Toute personne de bonne foi peut en déduire le nom du concepteur !

Quel accueil ?

La distance intertextuelle — présentée dans de nombreux séminaires à partir de 1998 et, pour la première fois en congrès à Lausanne en mars 2000 avec D. Monière, puis à Québec en juillet 2000 avec J.-G. Bergeron — a suscité un intérêt à travers le monde. A l'étranger, plusieurs chercheurs l'utilisent. Pour l'instant, deux articles sur Shakespeare ont été publiés (Merriam, 2002 et 2003). D'autres travaux sont annoncés.

Pour la France, ce calcul aurait été implanté dans un logiciel commercial (Hyperbase). Cependant, le manuel de ce logiciel donne une formule erronée. Il ne mentionne aucune des précautions que nous avons évoquées au début de cette conférence. Le logiciel opère sur les formes graphiques brutes sans aucune normalisation. Enfin, personne n'a accès au code source de ce programme : il est impossible de vérifier que le calcul est correctement fait. Nous récusons donc l'utilisation de notre nom pour une promotion commerciale.

De façon générale, il faut accorder aucune confiance aux logiciels d'«analyse textuelle» commercialisés en France, du moins tant que les principaux algorithmes implantés dans ces programmes ne seront pas accessibles librement à la communauté scientifique.

A ce propos, je tiens à souligner que les textes sur lesquels nous avons travaillé (bruts et lemmatisés) ainsi que nos programmes sont dans le domaine public — accessibles en ligne sur le site de l'Université d'Oxford (<http://ota.ahds.ac.uk/2466>) — et que les codes sources sont remis à tous les chercheurs qui en font la demande. Nous ne tirons aucun profit de ce travail. Nos critiques sont-ils aussi désintéressés ?

Mon essai explique que nous n'avons pas attaché une grande importance à notre trouvaille. D'une part, les Français ne se sont jamais intéressés à l'attribution d'auteur. Ils ont d'ailleurs toujours été d'une assez grande désinvolture concernant la propriété intellectuelle. Quant à Corneille et Molière : P. Louÿs, H. Poulaille et H. Wouters avaient apporté suffisamment d'éléments pour que toute personne de bonne foi admette comme hypothèse plausible la contribution décisive de Corneille dans l'œuvre de Molière.

Au début, une petite troupe de gens curieux et sans préjugés ont pris contact pour comprendre notre méthode et juger de notre sérieux. Parmi eux, quelques chercheurs : pas beaucoup de littéraires ; surtout des biologistes, des médecins qui savent que l'information est à la base du vivant et qu'il n'y a pas de science expérimentale sans modèles statistiques et sans informatique.

Il y a eu aussi quelques journalistes qui ont présenté à leurs lecteurs des dossiers équilibrés avec nos résultats et les arguments de nos critiques. Evoquons au moins : *Science et Vie* (décembre 2002), *Science et Vie Junior* (mars 2003) et un quotidien belge flamand du matin *die Morgen* (avril 2003).

A l'opposé, une campagne d'insultes a été orchestrée par un petit groupe d'universitaires et de critiques littéraires. Ils ont placé sur internet des pamphlets de bas niveau et fait circuler un certain nombre de ragots à travers les listes de discussion et par le courrier électronique. Ils ont même organisé une supercherie, dans une annexe de la Sorbonne, pour faire croire à la galerie que nos résultats étaient faux (voir sur notre page personnelle : *Baudelaire et Rimbaud victimes d'une rumeur malveillante*). La grande presse leur a donné un écho complaisant, tout en ignorant systématiquement les réponses placées sur mon site personnel. Aucun des spécialistes, qui participent à nos travaux et qui utilisent nos méthodes, n'a été interrogé. En

réponse à ces articles malveillants, nous avons envoyé quelques lettres aux rédacteurs en chef de ces journaux. Aucune n'a été publiée.

Au milieu des insultes et des quolibets, un certain nombre d'objections plus ou moins sérieuses ont été présentées.

Quelles objections ?

Nous avons déjà réfuté les deux principales : la prétendue impossibilité de distinguer les auteurs de pièces écrites à la même époque dans un même genre ; la prétendue "brouille" entre Molière et Corneille qui n'est étayée par aucun fait.

Dans notre article du *Journal of Quantitative Linguistics*, nous avons nous-mêmes suggéré une troisième objection. Corneille était manifestement l'auteur favori de Molière. Il en connaissait donc des milliers de vers et il aurait pu être, en quelque sorte, "imprégné" par le style, le vocabulaire et la technique de Corneille. Cette idée n'est pas très flatteuse pour Molière, mais plusieurs de nos critiques, dont l'académicien M. Fumaroli, l'ont reprise à leur compte. Nous avons répondu qu'il est curieux qu'une telle influence puisse se manifester par éclipse : profonde dans les *Fâcheux*, bien peu visible dans les *Précieuses*, profonde à nouveau dans *l'Ecole des femmes* et absente dans *la Critique de l'école des femmes...* etc. De deux choses l'une : ou bien ces critiques apportent des preuves qu'une telle influence intermittente peut s'exercer pendant 15 ans entre deux auteurs — y a-t-il d'autres exemples dans l'histoire d'un mimétisme d'une telle puissance et d'une telle durée ? — ou bien c'est une "théorie ad hoc", valable seulement pour Corneille et Molière, et l'on sait ce que cela vaut du point de vue scientifique...

Deux autres objections ont encore été présentées.

Premièrement, notre calcul ne tiendrait pas compte de la prosodie. Cela a été imprimé dans d'innombrables articles de presse et répété dans plusieurs pamphlets sur internet. Naturellement, c'est faux : les règles prosodiques appartiennent au genre et leur influence entre donc dans notre calcul. Si cette idiotie a été crue, c'est qu'il y a un préjugé courant selon lequel chaque auteur aurait sa "manière propre" de faire des vers (rythme et rime). On s'attend à ce que Corneille, Molière ou Racine aient un certain nombre de "structures favorites" ou des "rimes préférées" et l'on pense qu'elles distingueront aisément chacun des auteurs... Bien sûr, on peut repérer des "tours de main", chez certains auteurs à un moment donné, mais ce sont des techniques dont ils sont très conscients et sur lesquelles ils ont plein contrôle : ils peuvent les utiliser dans un genre particulier, à un moment donné, et les délaïsser lorsqu'ils passent à autre chose ou que la mode en est passée. La chose peut être vérifiée grâce au "métromètre" qui permet une analyse précise de la structure des vers. V. Beaudouin l'a appliqué à Corneille et Racine. Ses résultats sont clairs : pour des pièces contemporaines, dans un même genre, cet instrument discrimine mal les auteurs (naturellement, l'outil présente d'autres intérêts pour l'analyse littéraire). Si l'outil est appliqué à Molière, on retrouvera certainement, dans les pièces en vers, une proportion importante des techniques prosodiques propres aux *Menteurs*. Mais, étant donné l'évolution de ces techniques avec le temps, ce sont les *Plaideurs de Jean Racine — contemporains des comédies en vers écrites par Corneille pour Molière —*, qui comporteront la plus forte proportion de techniques "moliéresques"...

Deuxièmement, on nous oppose cinq vers, effectivement troublants, de *l'Ecole des femmes* et quelques lignes de l'Abbé d'Aubignac accusant Corneille d'ourdir une "cabale" contre cette

pièce. Je ne cherche nullement à escamoter ces deux textes. Bien au contraire, je les cite intégralement dans mon livre (voir annexes VIII et IX). P. Louÿs et H. Poulaille pensaient que les vers en question n'ont pas été écrits par Corneille — ils sont effectivement malhabiles et inférieurs au reste de la pièce — mais qu'ils ont été collés là par une autre main. C'est un argument qui mériterait d'être approfondi. Par exemple, nos calculs montrent que le *Bourgeois gentilhomme* et le *Malade imaginaire* sont une "collaboration" comme *Psyché* (l'essentiel étant là encore de la main de Corneille). Pourquoi n'en serait-il pas de même dans certaines autres pièces ?

Comme on le voit, nous ne prétendons pas avoir le dernier mot...

Il ne faudrait pas exagérer la portée de la campagne menée contre nous. Nous avons reçu un abondant courrier favorable. Plusieurs collègues littéraires nous ont écrit que ces résultats confortaient des doutes anciens ; d'autres nous ont dit leurs interrogations ou ont avoué qu'ils étaient ébranlés. Cela a été aussi l'occasion de nouer quelques collaborations nouvelles. A toutes ces personnes, nous avons dit notre souhait de ne pas organiser de "contre-cabale" et de ne pas répondre sur le même ton insultant que nos critiques. J'ai simplement placé, sur mon site internet, le maximum de documents, afin de permettre aux gens de bonne foi de se faire une opinion dans le calme et la sérénité.

L'acharnement de nos critiques s'explique aussi par des raisons plus profondes. Outre le rejet de la démarche scientifique, si fréquent dans les milieux littéraires, mon livre signale deux de ces raisons :

— il faudrait présenter aux jeunes les grandes œuvres littéraires, d'une façon plus vivante et moins scolastique qu'on le fait actuellement. C'est possible avec l'aide de l'ordinateur et nos travaux peuvent y aider. Naturellement, cela n'enchant pas ceux qui vivent du marché lucratif des "petits classiques" et qui ont bien conscience que, tôt ou tard, le "multi-média" mettra en péril leur trafic ;

— le français est la seule grande langue pour laquelle on ne connaît pas, de façon scientifique, la manière dont elle est parlée et écrite par ses usagers. En effet, pour les principales langues européennes, on dispose de vastes échantillons représentatifs oraux et écrits appartenant à de multiples genres et provenant de toutes les classes sociales. Chacun des mots est pourvu d'une étiquette indiquant sa catégorie grammaticale, son genre, etc. Cela permet une étude scientifique de la grammaire, des structures de phrase, du lexique... Ce sont des outils indispensables pour l'enseignement — surtout l'enseignement aux étrangers qui ne sont pas immergés dans une communauté francophone —, les correcteurs orthographiques, la documentation et l'archivage, les traducteurs automatiques. Sans ces outils, une langue est condamnée à plus ou moins court terme à la disparition dans un siècle dominé par la communication. Il y en a pour l'anglais, l'allemand, l'espagnol... Le tchèque est la dernière en date (cette langue est parlée par moins de 10 millions de personnes dans le monde). L'écossais est annoncé pour bientôt... Il n'y a rien en vue pour le Français. Au moins vingt-cinq ans ont été perdus sans espoir.

Tous les Français de mon âge, s'ils ont un peu voyagé, peuvent mesurer le désolant recul de notre langue, à travers le monde, sur le dernier demi-siècle. Ce recul tient à de nombreuses causes : culturelles, politiques, économiques, démographiques... Mais il ne faut pas négliger non plus les explications techniques comme celles que je viens de vous signaler. Mes travaux

montrent qu'une autre voie aurait été possible. J'ai la faiblesse de penser que c'est aussi pour cela qu'on veut me faire taire.

Bibliographie

- Barthélémy Jean-Pierre, Guénoche Alain (1988). *Les arbres et les représentations de proximité*. Paris. Dunod.
- Beaudouin Valérie (2002). *Mètre et rythmes du vers classique : Corneille et Racine*. Paris. Champion.
- Bergeron Jean-Guy et Labbé Dominique (2000). "L'évaluation de la négociation raisonnée par les acteurs: une analyse lexicométrique" - XVI^e congrès de l'Association Internationale des Sociologues de Langue Française. Québec. 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 (2002). "“Delta” : a Measure of Stylistic Difference and a Guide to Likely Authorship". *Literary and Linguistic Computing*. 17-3. September 2002. p. 267-287.
- Labbé Cyril, Labbé Dominique (2001). "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 Dominique, Monière Denis (2000), "La connexion intertextuelle. Application au discours gouvernemental québécois", Martin RAJMAN et Jean-Cédric CHAPPELIER (eds), *Actes des 5^e journées internationales d'analyse des données textuelles*, Lausanne, Ecole polytechnique fédérale, vol 1, p 85-94.
- Labbé Dominique (2003). *Corneille dans l'ombre de Molière. Histoire d'une recherche*. Bruxelles, Les impressions nouvelles.
- Luong Xuan (1988). *Méthodes d'analyse arborée. Algorithmes, applications*. Thèse pour le doctorat es sciences. Université de Paris V.
- Merriam Thomas (2002a). "Linguistic Computing in the Shadow of Postmodernism". *Literary and Linguistic Computing*. 17-2. June 2002. p 181-192.
- Merriam Thomas (2002b). "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 (2003). "An Application of Authorship Attribution by Intertextual Distance in English". *Corpus*. 2. 2003. p 167-182.
- Poulaille Henry (1957). *Corneille sous le masque de Molière*. Paris. Grasset.
- Wouters Hippolyte, Ville de Goyer, Christine de (1990). *Molière ou l'auteur imaginaire ?* Bruxelles. Eds Complexe.

Annexe I.
Les pièces de Corneille.

Corneille	Année de création	Genre	Taille en mots	
1	Mélite	1630	Comédie en vers	16 690
2	Clitandre	1631	Comédie en vers	14 402
3	La Veuve	1631	Comédie en vers	17 661
4	La Galerie du Palais	1632	Comédie en vers	16 140
5	La Suivante	1633	Comédie en vers	15 160
6	Comédie des Tuileries	1634	Comédie en vers	3 627
7	Médée	1635	Tragédie en vers	14 269
8	La Place Royale	1634	Comédie en vers	13 801
9	L'illusion comique	1636	Comédie en vers	15 428
10	Le Cid	1636	Tragédie en vers	16 677
11	Cinna	1641	Tragédie en vers	16 126
12	Horace	1640	Tragédie en vers	16 482
13	Polyeucte	1641	Tragédie en vers	16 472
14	Pompée	1642	Tragédie en vers	16 492
15	Le menteur 1	1642	Comédie en vers	16 653
16	Le menteur 2	1643	Comédie en vers	17 675
17	Rodogune	1644	Tragédie en vers	16 842
18	Théodore	1645	Tragédie en vers	17 121
19	Héraclius	1647	Tragédie en vers	17 433
20	Andromède	1650	Tragédie en vers	15 514
21	Don Sanche	1650	Tragédie en vers	16 947
22	Nicomède	1651	Tragédie en vers	16 923
23	Pertharite	1651	Tragédie en vers	17 121
24	Oedipe	1659	Tragédie en vers	18 618
25	Toison d'Or	1661	Tragédie en vers	20 343
26	Sertorius	1662	Tragédie en vers	17 675
27	Sophonisbe	1663	Tragédie en vers	16 858
28	Othon	1664	Tragédie en vers	16 971
29	Agésilas	1666	Tragédie en vers	18 227
30	Atilla	1667	Tragédie en vers	16 788
31	Tite et Bérénice	1670	Tragédie en vers	16 697
32	Pulchérie	1672	Tragédie en vers	16 630
33	Suréna	1674	Tragédie en vers	16 545
Psyché				
34	Psyché Corneille	1671	Comédie en vers	10 067
35	Psyché Molière	1671	Comédie en vers	4 816
36	Psyché Quinault	1671	Comédie en vers	1 399

Le corpus Corneille compte 34 pièces, dont 32 complètes. Il est long de 553 190 "occurrences". Son vocabulaire comporte : 15 535 formes graphiques normalisées 6 258 vocables.

Annexe II

Distances séparant les deux *Menteurs* (Corneille) et les *Plaideurs* (Racine)
de toutes les pièces de Molière

N°	Pièces	Genre	Le Menteur (Corneille 1642)	Suite du Menteur (Corneille 1643)	Les plaideurs (Racine : 1668)
15	Le Menteur (1642)	Vers	0,000	0,180	0,296
16	La suite du Menteur (1643)	Vers	0,180	0,000	0,293
34	Psyché Corneille (1671)	Vers	0,288	0,273	0,348
36	Psyché Molière (1671)	Vers	0,329	0,325	0,354
37	La jalousie du barbouillé (avant 1660)	Prose	0,341	0,331	0,327
38	Médecin volant (avant 1660)	Prose	0,310	0,293	0,302
39	L'étourdi (1658)	Vers	0,205	0,206	0,269
40	Dépit amoureux (1658)	Vers	0,215	0,212	0,270
41	Précieuses ridicules (1660)	Prose	0,315	0,314	0,314
42	Sganarelle ou le cocu imagin. (1660)	Vers	0,259	0,253	0,293
43	Dom Garcie de Navarre (1661)	Vers	0,280	0,273	0,359
44	L'école des maris (1661)	Vers	0,223	0,217	0,279
45	Les fâcheux (1661)	Vers	0,248	0,248	0,306
46	L'école des femmes (1662)	Vers	0,226	0,217	0,261
47	Critique de l'école des femmes (1663)	Prose	0,323	0,319	0,340
48	L'impromptu de Versailles (1663)	Prose	0,321	0,316	0,323
49	Mariage forcé (1664)	Prose	0,322	0,302	0,320
50	Princesse d'Elide (1664)	Vers Prose	0,251	0,243	0,314
51	Le Tartuffe (1664)	Vers	0,242	0,228	0,275
52	Dom Juan (1665)	Prose	0,259	0,248	0,281
53	L'amour médecin (1665)	Prose	0,292	0,289	0,287
54	Le Misanthrope (1666)	Vers	0,252	0,234	0,283
55	Médecin malgré lui (1666)	Prose	0,298	0,289	0,296
56	Mélicerte (1666)	Vers	0,257	0,250	0,322
57	Le sicilien ou l'amour peintre (1667)	Prose	0,277	0,260	0,301
58	Amphytrion (1668)	Prose	0,253	0,256	0,297
59	Georges Dandin (1668)	Prose	0,292	0,279	0,292
60	L'Avare (1668)	Prose	0,256	0,244	0,270
61	M. de Pourceaugnac (1669)	Prose	0,292	0,283	0,285
62	Amants magnifiques (1670)	Prose	0,282	0,279	0,329
63	Bourgeois gentilhomme (1670)	Prose	0,294	0,280	0,286
64	Fourberies de Scapin (1671)	Prose	0,269	0,263	0,281
65	Comtesse d'Escarbagnas (1671)	Prose	0,311	0,300	0,305
66	Femmes savantes (1672)	Vers	0,260	0,248	0,283
67	Malade imaginaire (1672)	Prose	0,282	0,270	0,278
<i>Moyenne oeuvre de Molière</i>			<i>0,275</i>	<i>0,266</i>	<i>0,299</i>
<i>Moyenne pièces en vers de Molière</i>			<i>0,241</i>	<i>0,234</i>	<i>0,290</i>
<i>Moyenne oeuvre de Corneille</i>			<i>0,252</i>	<i>0,249</i>	<i>0,347</i>
<i>Moyenne oeuvre de Racine</i>			<i>0,314</i>	<i>0,311</i>	<i>0,376</i>

Corpus Molière : 34 pièces, 394 963 occurrences, 16 735 formes graphiques normalisées et 8 088 vocables.

Corpus Racine : 12 pièces, 166 626 occurrences, 10 120 formes graphiques normalisées et 4 323 vocables

Annexe III.

Classification arborée sur l'ensemble des œuvres théâtrales de Corneille et Molière
(extrait du JQL, VIII, 3, p 227)



Ce graphe a été tracé par M. X. Luong de l'université de Nice. Nous lui avons fait parvenir le tableau de données sans lui indiquer les auteurs et textes sur lesquels portait l'expérience.
Pour les numéros des pièces, se reporter aux annexes précédentes.

Les traits en gras :

- N° 06 Corneille : Comédies des Tuileries (Richelieu, 1634)
- N° 15 et 16 Corneille : Le menteur et la suite du menteur (1642 et 1643)
- N° 34 : passages de Psyché écrits par Corneille
- N° 35 : passages de Psyché écrits par Molière
- N° 36 : prologue de Psyché écrit par Quinault
- N° 43 : Dom Garcie de Molière

Annexe IV.

Distances séparant Dom Garcie (Molière) et Psyché (Corneille et Molière) des dernières pièces de Corneille.

Ultime pièces de Corneille	Dom Garcie (Molière,1661)	Psyché (Corneille, 1671)
Rodogune (1644)	0,245	0,231
Theodore (1645)	0,234	0,245
Heraclius (1647)	0,248	0,273
Andromède (1650)	0,241	0,218
DonSanche (1650)	0,224	0,251
Nicomède (1651)	0,244	0,264
Pertharite (1651)	0,235	0,263
Œdipe (1659)	0,223	0,226
Toison d'or (1661)	0,221	0,220
Sertorius (1662)	0,230	0,238
Sophonisbe (1663)	0,228	0,236
Othon (1664)	0,235	0,240
Agésilas (1666)	0,234	0,233
Attila (1667)	0,235	0,227
Tite et Bérénice (1670)	0,227	0,235
Psyché (1671)	0,230	—
Pulcherie (1672)	0,230	0,226
Surena (1674)	0,216	0,224
<i>Moyenne Corneille</i>	<i>0,243</i>	<i>0,244</i>
<i>Moyenne Molière</i>	<i>0,286</i>	<i>0,297</i>

Annexe V

Principales distances caractéristiques entre Corneille et Racine à l'époque de Tite et Bérénice.

	Tite et Bérénice (Corneille, 1670)	Bérénice (Racine, 1670)
CORNEILLE :		
Agésilas (1666)	0.159	0.278
Attila (1667)	0.180	0.289
Tite et Bérénice (1670)	0	0.256
Pulchérie (1672)	0.155	0.271
Suréna (1672)	0.156	0.264
RACINE :		
Andromaque (1667)	0.259	0.225
Britannicus (1669)	0.251	0.209
Bérénice (1670)	0.256	-
Bazajet (1672)	0.262	0.220
Mithridate (1673)	0.248	0.206

Annexe VI
Par qui ont été écrites les pièces de Molière ?

16 pièces sont attribuées à Corneille
(par ordre chronologique de création)

Titre	Actes	Genre	Date	Taille (mots)
L'étourdi	5	Vers	1658	18 674
Le Dépit amoureux	5	Vers	1656	16 243
Sganarelle ou le cocu imaginaire	1	Vers	1660	6 042
Dom Garcie de Navarre	5	Vers	1661	17 049
L'Ecole des maris	3	Vers	1661	10 536
Les fâcheux	3	Vers	1661	7 922
L'Ecole des femmes	5	Vers	1662	16 625
La princesse d'Elide	5	Vers et prose	1664	11 333
Le Tartuffe	5	Vers	1664	18 272
Dom Juan	5	Prose	1665	17 454
Le Misanthrope	5	Vers	1666	17 182
Mélicerte	2	Vers	1666	5 540
Amphytrion	3	Vers libres	1668	15 117
L'Avare	5	Prose	1668	21 033
Psyché	5	Vers	1671	16 182
Les Femmes savantes	5	Vers	1672	16 865

9 pièces de Molière n'ont pas été écrites par Corneille
(par ordre chronologique de création)

Titre	Actes	Genre	Date	Taille (mots)
La jalousie du barbouillé	1	Prose	1659	3 501
Le médecin volant	1	Prose	1659	3 876
Les précieuses ridicules	1	Prose	1660	6 651
Critique de l'école des femmes	1	Prose	1663	8 613
Impromptu de Versailles	1	Prose	1663	7 170
Le mariage forcé	1	Prose	1664	6 059
L'amour médecin	3	Prose	1665	6 148
Le médecin malgré lui	3	Prose	1666	9 319
La comtesse d'Escarbagnas	1	Prose	1671	5 565

7 pièces sont douteuses (elles peuvent avoir été écrites par Corneille mais elles sont trop éloignées des *Menteurs* ou de *Psyché* pour que l'on puisse conclure avec une certitude raisonnable)

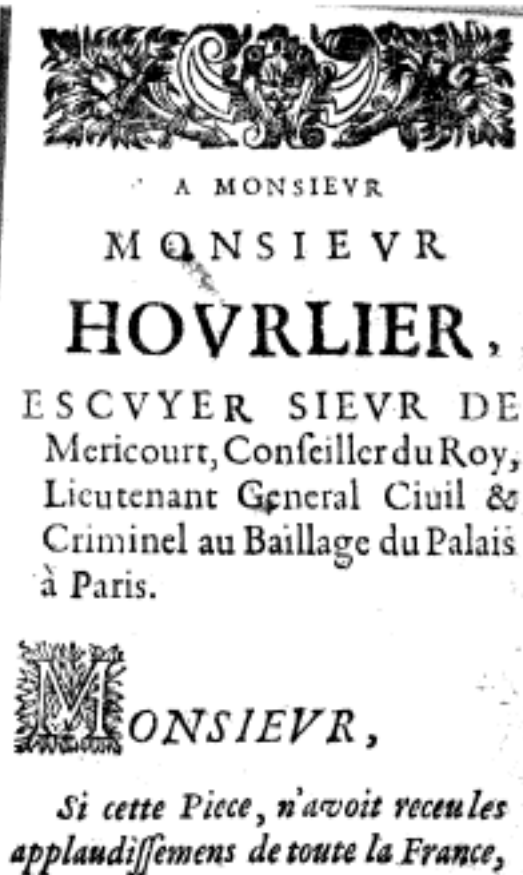
Titre	Actes	Genre	Date	Taille (mots)
Le sicilien ou l'amour peintre	1	Prose	1667	5 375
Georges Dandin	3	Prose	1668	11 009
Monsieur de Pourceaugnac	2	Prose	1669	11 803
Les amants magnifiques	5	Prose	1670	11 983
Le bourgeois gentilhomme	5	Prose	1670	17 136
Les fourberies de Scapin	3	Prose	1671	14 245
Le malade imaginaire	3	Prose	1673	19 920

NB : le *Bourgeois gentilhomme* et le *Malade imaginaire*, bien que plus éloignés des *Menteurs*, peuvent être rattachés à la première catégorie — celles des pièces écrites par Corneille car en enlevant les dernières scènes du *Bourgeois* ou les passages en latin de cuisine du *Malade* (ainsi que les intermèdes en italien), leur distance avec les deux *Menteurs* tombe en dessous de 0.25.

Annexe VII. Les syntagmes "pseudo-auxiliaire + infinitif" (fréquence pour 100.000 mots)

Corneille		Molière		Racine	
Syntagmes	Fréquence	Syntagmes	Fréquence	Syntagmes	Fréquence
<i>faire voir</i>	33,8	<i>faire voir</i>	31,5	aller voir	12,0
<i>pouvoir être</i>	18,8	<i>pouvoir être</i>	25,5	<u>pouvoir voir</u>	9,6
<i>pouvoir faire</i>	18,4	<i>pouvoir faire</i>	25,5	faire entendre	9,0
faire naître	13,9	vouloir dire	24,9	pouvoir faire	8,4
<u>pouvoir voir</u>	13,4	<i>vouloir faire</i>	19,5	aller chercher	7,8
devoir être	12,7	pouvoir dire	14,5	faire parler	7,8
pouvoir souffrir	10,8	pouvoir avoir	13,7	pouvoir être	7,8
<i>vouloir faire</i>	9,9	aller faire	13,2	venir chercher	7,2
faire connaître	9,6	avoir faire	13,2	faire éclater	6,6
devoir faire	8,7	<u>pouvoir voir</u>	12,3	falloir partir	6,6

Racine ne partage que "pouvoir voir" avec les deux autres qui, eux, en ont cinq en commun dont les trois premiers dans le même ordre et avec des densités voisines. Etant donné le nombre des combinaisons possibles, la probabilité pour qu'une telle "coïncidence" survienne au hasard est infinitésimale.



si elle n'avoit esté le charme de Paris, & si elle n'avoit esté le divertissement du plus grand Monarque de la Terre, ie ne prendrois pas la liberté de vous l'offrir. Il y a long-temps que i'avois resolu de vous presenter quelque chose qui vous marquast mes respects; Mais ne trouvant rien qui fut digne de vous estre offert, & qui fut proportionné à vos merites, i'avois toujours differé le iuste & respectueux hommage que ie m'étois proposé de vous rendre; & i'eusse peut-estre encore tardé long-temps à le faire, si le Depit Amoureux de l'Auteur le plus approuvé de ce siècle ne me fut tombé entre les mains. J'ay crû, Monsieur, que ie ne devois

Notes :

En me fondant sur cette couverture, j'ai écrit que l'édition originale du *Dépit amoureux* est de 1663. On m'a fait justement remarquer que l'impression a été achevée à l'automne 1662 : il ne faut pas se fier aux couvertures !

Il reste à savoir qui est "l'auteur le plus approuvé de ce siècle". Interrogé sur ce point, notre principal contradicteur a bien voulu, dans un document mis en ligne sur son site personnel durant l'été 2003, convenir que, dans les années 1660, Corneille "dominait de la tête et des épaules le théâtre français depuis vingt ans".

Annexe IX

Psyché (1671)

Le libraire au lecteur

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

Annexe X

Une moquerie de Molière contre les frères Corneille ?

L'Ecole des femmes (Acte 1, vers 165 sq)

CHRYSALDE.

Je me réjouis fort, seigneur Arnolphe...

ARNOLPHE.

Bon !

Me voulez-vous toujours appeler de ce nom ?

CHRYSALDE.

Ah ! malgré que j'en aie, il me vient à la bouche,

Et jamais je ne songe à Monsieur de la Souche.

Qui diable vous a fait aussi vous aviser,

A quarante et deux ans, de vous débaptiser,

Et d'un vieux tronc pourri de votre métairie

Vous faire dans le monde un nom de seigneurie ?

ARNOLPHE.

Outre que la maison par ce nom se connoît,

La Souche plus qu'Arnolphe à mes oreilles plaît.

CHRYSALDE.

Quel abus de quitter le vrai nom de ses pères

Pour en vouloir prendre un bâti sur des chimères !

De la plupart des gens c'est la démangeaison ;

Et, sans vous embrasser dans la comparaison,

Je sais un paysan qu'on appeloit Gros-Pierre,

Qui n'ayant pour tout bien qu'un seul quartier de terre,

Y fit tout à l'entour faire un fossé bourbeux,

Et de Monsieur de l'Isle en prit le nom pompeux.

ARNOLPHE.

Vous pourriez vous passer d'exemples de la sorte.

Mais enfin de la Souche est le nom que je porte :

J'y vois de la raison, j'y trouve des appas ;

Et m'appeler de l'autre est ne m'obliger pas.