La distance intertextuelle et l'attribution d'auteur
Corneille et Molière

(Version préliminaire en français de l'article "Inter-Textual Distance and Authorship Attribution. Corneille and Molière" paru dans : Journal of Quantitative Linguistics. 8-3, December 2001, p 213-231).

Cyril LABBE
Université Grenoble I
cyril.labbe@imag.fr

Dominique LABBE
Institut d'Etudes Politiques de Grenoble
dominique.labbe@iep.upmf-grenoble.fr

Résumé :

La distance intertextuelle quantifie les proximités entre plusieurs textes. Elle peut être mesurée grâce à un indice normalisé et à une échelle de la distance. Ces outils peuvent être utilisés pour l'attribution d'auteur. Une application est présentée sur l'un des cas célèbre de la littérature française : Corneille et Molière. Le calcul fait clairement la différence entre les deux oeuvres mais il démontre aussi que Corneille a contribué à de nombreux chefs d'oeuvre de Molière.

`

> « Molière aurait confié à Nicolas Despréaux : *Je dois beaucoup au Menteur. Lorsqu'il parut j'avais bien envie d'écrire, mais j'étais incertain de ce que j'écrirais ; mes idées étaient confuses : cet ouvrage vint les fixer* » .
> André Le Gall, *Corneille*, Paris, Flammarion, 1997, p 469.

La recherche de l'auteur d'un texte inconnu ou douteux est l'un des plus vieux problèmes de la statistique appliquée à la littérature. Il s'agit toujours de rapprocher ce texte d'autres dont les auteurs sont certains et dont on soupçonne qu'ils ont pu participer à sa rédaction. Habituellement, l'étude porte sur les mots les plus fréquents ou sur une sélection de ceux-ci, souvent les mots outils (function words). Voir à ce sujet (Holmes, 1995 et Baayen et al, 1996). Nous proposons ci-dessous un calcul qui prend en compte la totalité des textes et donne une mesure standardisée de la distance existant entre eux.

La question est connue sous le nom de « **connexion** lexicale ». Celle-ci est définie comme « l'intersection du vocabulaire de deux textes » (Muller 1977, p 145-154). La connexion est donc le complémentaire de la **distance**, terme plus familier en statistique et que nous retenons pour cette raison.

Pour comprendre la portée de ce calcul, il faut rappeler la différence existant entre « **mot** » et « mot différent » (ou **vocable**). Le mot est le plus petit élément mesurable d'un texte et le vocable, forme l'élément de base du **vocabulaire**. Par exemple, le plus long roman en langue française, *Les misérables,* compte un demi-million de mots (c'est sa **taille** ou son « étendue », notée N) et son vocabulaire (noté V) comporte moins de 10.000 vocables.

Jusqu'à maintenant, l'étude de la « connexion » a été faite sur le vocabulaire sans tenir compte de la fréquence des mots (par exemple Brunet, 1988). Nous proposons ici de considérer la fréquence d'emploi de chacun des vocables, c'est-à-dire l'ensemble de l'étendue des textes comparés. Dans le terme « distance intertextuelle », l'adjectif **textuel** indique donc que les calculs portent sur l'ensemble des textes (N) et non sur leur seul vocabulaire (V).

Après avoir présenté le calcul nous proposons une application à l'un des cas les plus célèbres de la littérature française : Corneille et Molière.

La distance intertextuelle

Pour pouvoir dire si deux textes sont « plutôt proches » ou « plutôt éloignés », quant à l'utilisation du vocabulaire, il faut transformer la mesure absolue de leur distance en un indice.

On recherche donc un indice δ :

— insensible aux différences de taille entre les textes comparés ;

— applicable à plusieurs textes et, potentiellement, à tous les textes d'une même langue ;

— variant uniformément — entre 0 (même vocabulaire et fréquence semblable de chacun des mots dans les deux textes) et 1 (aucun vocable en commun) — sans saut ni effet de seuil autour de certaines valeurs ;

— symétrique (soit deux textes A et B alors $\delta_{(A,B)} = \delta_{(B,A)}$) ;

— aussi « transitif » que possible : quand on agrège le vocabulaire de deux textes, les distances de ce nouveau vis-à-vis des autres textes doit refléter l'ordre des distances antérieures (si $\delta_{(A,B)} > \delta_{(A,C)} > \delta_{(B,C)}$ alors $\delta_{(A,B)} > \delta_{\{A,(B \cup C)]}$) ;

— aussi "robuste" que possible (ie une modification marginale dans le vocabulaire d'un des deux textes doit se traduire par une variation marginale de leur distance)…

Quand on examine les travaux classiques en ce domaine, on trouve habituellement les calculs suivants :

Soit deux textes A et B et,

V_a et V_b : nombre de vocables dans A et B (vocabulaire) ;

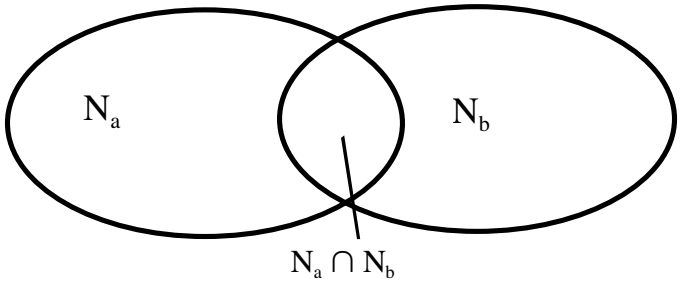
F_{ia} : fréquence du vocable i dans A ;

F_{ib} : fréquence du vocable i dans B.

N_a et N_b : nombre de mots dans A et B (taille) ;

avec $N_a = \sum F_{ia}$ et $N_b = \sum F_{ib}$

La distance absolue entre A et B sera la surface des deux textes moins leur intersection, c'est-à-dire la somme des différences entre les fréquences absolues de chacun des mots des deux textes.



La distance relative pourra être calculée de deux manières :

$$(1)\ \delta_{(a,b)} = \frac{\sum_{Va} | F_{ia} - F_{ib} | + \sum_{Vb} | F_{ib} - F_{ia} |}{N_a + N_b}$$

ou :

$$(2)\ \delta_{(a,b)} = \frac{1}{2} \left(\frac{\sum_{Va} | F_{ia} - F_{,b} |}{N_a} + \frac{\sum_{Vb} | F_{ib} - F_{ia} |}{N_b} \right)$$

La formule (2) est, à la notation près, celle qui est suggérée par E. Brunet dans Brunet (1988). La distance maximale absolue est égale à $N_a + N_b$.

Ces formules classiques ont soulevé deux objections :

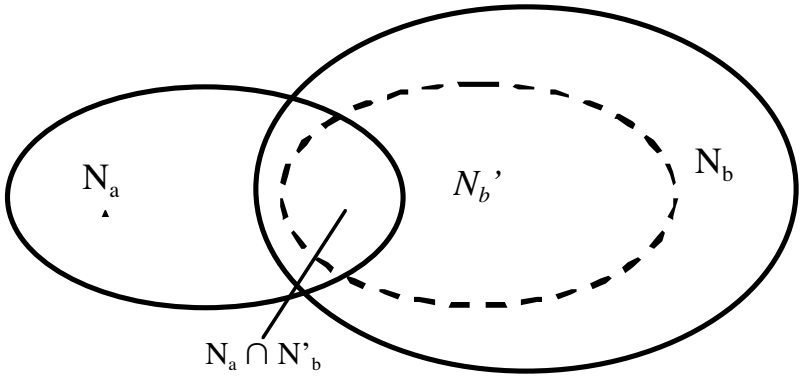
— (1) et (2) ne sont équivalentes que quand les textes ont des tailles égales ($N_a = N_b$). Si les deux textes comparés ne partagent aucun vocable, les formules (1) et (2) donnent bien un indice de 1 quelle que soit la taille des textes (ce qui est une des conditions requises pour l'indice idéal). En revanche, le minimum théorique ne peut atteindre zéro que dans le cas particulier de tailles égales. En effet, plus les textes comparés seront de tailles différentes, plus le numérateur minimal possible s'éloignera de zéro. Par exemple, dans le discours politique québécois : le texte de 1965 qui est le plus court du corpus a une taille de 1 006 mots et un vocabulaire de 419 vocables alors que le texte de 1984 (le plus long du corpus) contient 12 828 mots et 2 790 vocables. Physiquement parlant, les 2 790 vocables du texte de 1984 ne peuvent pas tous entrer dans le texte de 1965. Même si le petit texte était totalement inclus dans le grand, la distance ne serait pas nulle, puisque le calcul porte également sur les (2 790 - 419) vocables absents du plus petit et ne pouvant pas tous y figurer.

— dans (1) comme dans (2), l'intersection des deux textes est comptée deux fois. On donne donc plus d'importance aux vocables communs qu'aux vocables propres à chacun des textes.

Comment surmonter ces deux objections pour une donner une mesure plus fine de la distance entre plusieurs textes ?

Une approximation de la distance intertextuelle

Il est proposé de simuler la réduction du plus grand des deux textes à la taille du plus petit. Soit B' cette réduction de B en fonction de la taille de A :



Soit $U_{(a,b)}$, le coefficient de proportionnalité entre A et B :

$$U_{(a,b)} = \frac{N_a}{N_b}$$

Tout mot de fréquence F_i dans B aura une fréquence attendue (espérance mathématique) dans A égale à :

$$E_{ia(u)} = F_{ib} * U_{(a,b)}$$

D'où l'on tire que :

$$N'_b = \sum_{\forall b} E_{ia(u)} = N_a$$

On propose donc de remplacer dans les formules (1) et (2) les termes F_{ib} par $E_{ia(u)}$ et N_b par N'_b.

Le minimum théorique (zéro) sera atteint quand le petit texte sera une sorte de modèle réduit du grand. Dans ce cas, tous les vocables de A se retrouvent dans B avec une fréquence telle que :

$$F_{ia} = E_{ia(u)}$$

Le numérateur de la formule (2) sera égal à zéro et le dénominateur à :

$$N_a + N'_b = 2\ N_a$$

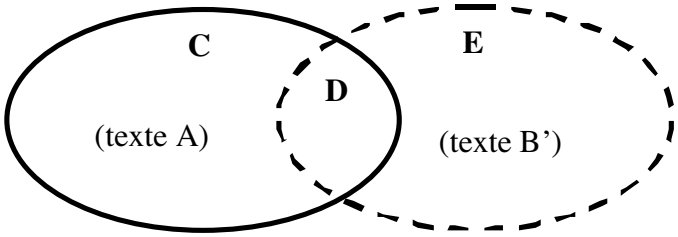
C'est en effet l'effectif maximum des mots que les deux textes peuvent avoir en commun s'ils ont même dimension, même vocabulaire et pour chacun des vocables, même fréquence.

Le maximum théorique (l'unité) devrait être atteint quand les deux textes n'ont aucun mot en commun. Au numérateur, comme au dénominateur, figureront N_a et N'_b.

Toutefois, cette nouvelle formulation ne répond pas à l'objection concernant le double compte de l'intersection des deux textes et ne résout pas totalement le problème physique mentionné ci-dessus : tous les vocables de B ne peuvent pas théoriquement figurer dans A. Pour tenir compte de ces deux objections, il est proposé de :

- ne considérer qu'une seule fois l'intersection des deux textes ;

- limiter le calcul à l'ensemble des vocables de A mais aux seuls vocables de B dont la fréquence est telle que l'on en attend au moins 1 dans A ($E_{ia(u)} \geq 1$). La somme de ces espérances donne N'_b.



La procédure de calcul se déroule en trois temps (voir figure ci-dessus).

Pour les V_a vocables (ensemble C), la contribution à la distance est égale à :

$$D_{V_a,\, b(u)} = \sum_{V_a} \left| F_{ia} - E_{ia(u)} \right|$$

Pour que l'indice maximal soit effectivement égal à 1 (quand A et B' n'ont aucun vocable en commun) et toujours inférieur à l'unité si leur intersection n'est pas vide, il faut considérer successivement :

— les J mots de A pour lesquels : $F_{ia} \geq E_{ia(u)}$. Ce premier ensemble comprend C (mots pour lesquels $E_{ja(u)} = 0$), et la partie de D que l'on peut rattacher à A. Pour ce groupe de mots, la distance maximale théorique possible, qui devra figurer au dénominateur de l'indice, sera atteinte quand les J mots seront tous absents de B'. Elle est donc égale à :

$$D_{max(j)} = \sum_{J} F_{ia}$$

— les K mots de A pour lesquels, : $E_{ia(u)} > F_{ia}$. C'est la partie de D que l'on peut rattacher à B'. Pour ce second ensemble, le maximum théorique possible est égal à :

$$D_{max(k)} = \sum_{k} E_{ia(u)}$$

Il faut enfin envisager l'ensemble E composé des vocables de B' absents de A et qui devraient s'y trouver si les deux textes étaient identiques (l'espérance mathématique du nombre de leurs occurrences dans A en fonction de leur fréquence dans B est au moins égale à l'unité).

Il y a L vocables qui répondent aux deux conditions :

$F_{ia} = 0$ (absent de A) ;

$E_{ia(u)} \geq 1$ (fréquence attendue dans A, en fonction de la fréquence dans B, au moins égale à l'unité).

Pour ces L vocables, la distance observée et la distance maximale théorique seront identiques. D'où :

$$D_{max\ (L)} = \sum_{L} E_{ia(u)}$$

La distance absolue séparant A et B sera égale à la somme des trois contributions pour les J, K, L vocables. Et la distance relative s'obtiendra en divisant cette somme par celle des trois maxima :

$$(3) \quad D_{(a,b)} = \frac{\sum_{j} \left| F_{ia} - E_{ia(u)} \right| + \sum_{k} \left| F_{ia} - E_{ia(u)} \right| + \sum_{l} \left| E_{ia(u)} \right|}{D_{max(j)} + D_{max(k)} + D_{max(l)}}$$

Dans le cas d'une intersection (D) vide, les premiers membres du numérateur et du dénominateur seront égaux à Na, le second membre sera nul et les troisièmes égaux à N'b. Ce qui donne bien un indice de 1.

On remarquera que le même résultat, aux arrondis près, peut être obtenu en soustrayant les fréquences relatives de chacun des vocables dans les deux textes comparés, à condition de limiter le calcul à tout le vocabulaire du plus petit des deux textes et à ceux des vocables qui, dans le plus grand, ont une fréquence suffisante pour qu'on en attende au moins un s'il avait la taille du plus petit.

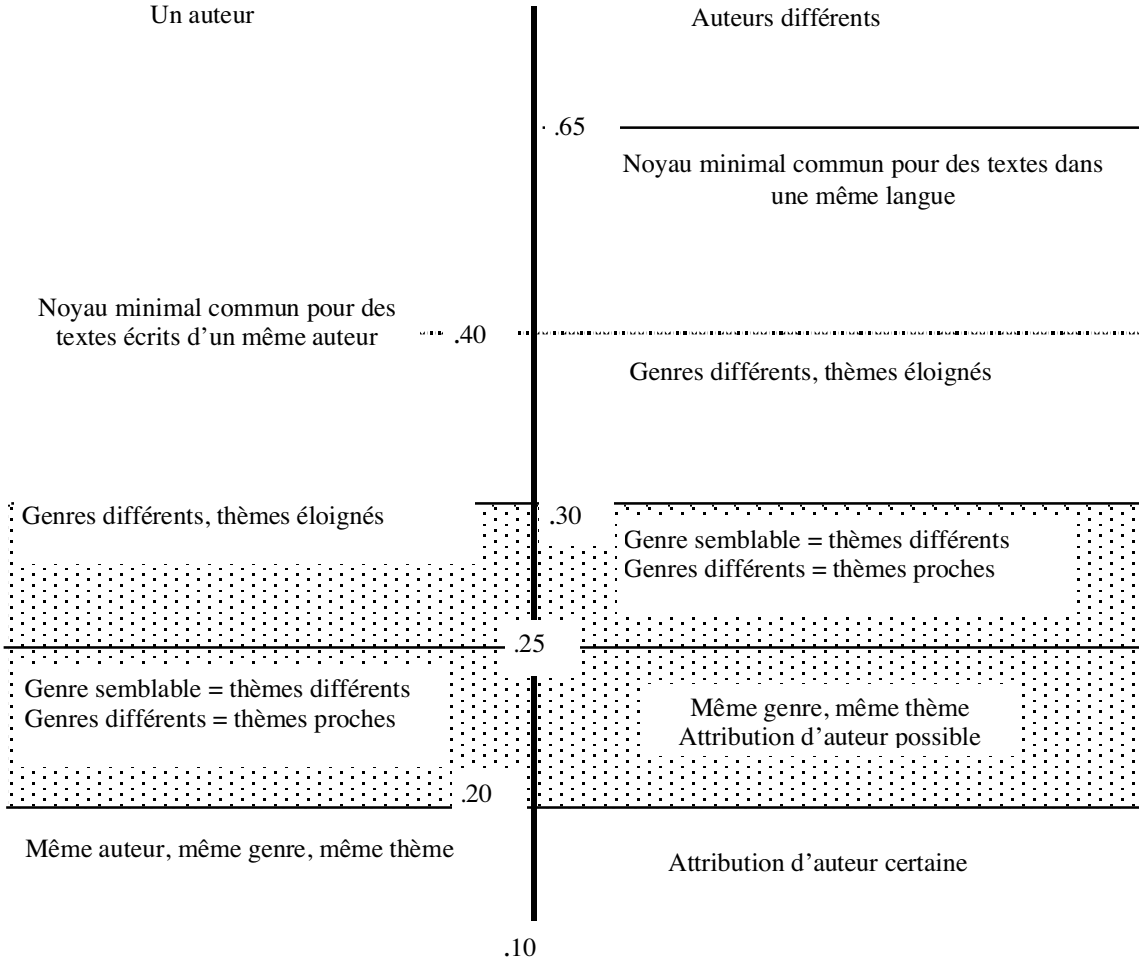
Les arrondis introduisent une légère incertitude dans les résultats. Alors que les fréquences observées sont toujours des entiers, les fréquences théoriques auront presque toujours des décimales qui entreront dans la distance. Ce défaut sera d'autant plus sensible que les mots de basses fréquences occuperont une surface importante, ce qui est le cas quand les textes sont brefs. Pour limiter partiellement ce premier inconvénient, on n'applique pas le calcul à de trop

petits textes. Dans l'application ci-dessous, le plus petit texte comporte 3.500 mots (il s'agit de la première comédie de Molière) et le plus long 20.300 (La toison d'or de Corneille)[1]. De manière plus générale, il est préférable de ne pas traiter de taille inférieure à 1.000 et de s'en tenir à une échelle des dimensions inférieur à 1/10 environ. Pour les mêmes raisons, il faut éliminer du numérateur tous les résultats inférieurs à 0.5.

Echelle des distances

Le calcul a été appliqué à divers corpus (tous dépouillés selon la même norme) ce qui permet d'établir empiriquement une échelle des distances (Table I).

Table I. Echelle normalisée des distances entre textes[2]



— pour un même auteur, on constate toujours des distances inférieures à celles qui peuvent exister entre deux auteurs différents (quand ils traitent d'un même thème à peu près à

[1] Note 2007 : la pièce la plus longue est l'Avare (21 033 mots)
[2] Note 2007 : cette échelle a été calibrée avec des échantillons dont les longueurs étaient comprises entre 5000 et 20000 mots.

la même époque). La procédure de reconnaissance d'auteur nécessite donc un choix raisonné des textes comparés pour neutraliser autant que possible les effets du genre et des thèmes qui augmentent les distances.

— Les distances inférieures à .20 ne se constatent généralement que chez un même auteur et pour des textes appartenant à un même genre avec des thèmes proches. En cas d'auteur inconnu, l'attribution d'auteur est quasi-certaine. Si les deux textes ont officiellement des auteurs différents, l'un des deux s'est « inspiré » de l'autre…

— Entre 0,20 et 0,25 s'étend une zone grise où la parenté entre les textes demeure forte. Si l'auteur est unique, les thèmes ou les genres changent. Si l'un des auteurs est inconnu, l'attribution est probable mais ne pourra être avérée que si l'on peut démontrer qu'il n'existe pas d'autres textes plus proches et si d'autres indices, notamment stylistiques, viennent conforter la conclusion.

— Au-dessus de 0,25 les genres et/ou les thèmes sont trop éloignés pour qu'on puisse valablement utiliser ces textes pour une attribution d'auteur.

A titre d'illustration, nous proposons une application du calcul aux pièces de théâtre de Corneille et Molière. En effet, dès l'origine, des rumeurs ont couru sur la paternité des pièces de Molière. Ces rumeurs ont notamment été nourries par un « avertissement de l'éditeur » placé en tête de la publication d'une des pièces (*Psyché*, 1671), avertissement qui attribuait à Corneille les deux tiers des vers alors que la pièce avait été jouée auparavant sous le seul nom de Molière (cette pièce ainsi que l'avertissement de l'éditeur sont reproduits dans le deuxième volume des œuvres complètes de Corneille publiées dans la Pléaïde chez Gallimard). Depuis lors, la question a refait surface à plusieurs reprises, le nom de Corneille étant le plus souvent cité comme « plume de l'ombre ». Au début du XXe siècle, le poète P. Louys et plus récemment deux auteurs belges ont souligné la parenté frappante entre les deux oeuvres (Wouters et Ville de Goyer, 1990).

Le théâtre de Molière

Le calcul de la distance intertextuelle apporte quelques informations intéressantes à ce sujet. Voici d'abord, à titre d'exemple, les distances séparant les pièces de Molière les plus connues et les plus jouées (tableau II).

Table II. Distances entre les principales œuvres de Molière

	L'école des femmes	Tartuffe	Dom Juan	Le Misanthrope	L'avare	Bourgeois gentilh.	Femmes savantes	Malade imaginaire
Ecole des femmes	0	0,183	0,205	0,194	0,200	0,231	0,198	0,223
Le Tartuffe		0	0,199	0,167	0,199	0,230	0,170	0,219
Dom Juan			0	0,204	0,170	0,207	0,219	0,205
Le Misanthrope				0	0,210	0,239	0,173	0,239
L'avare					0	0,194	0,214	0,187
Bourgeois gentilh.						0	0,234	0,196
Femmes savantes							0	0,226
Malade imaginaire								0

Le calcul fait donc apparaître une nette proximité entre toutes ces pièces malgré la grande diversité des thèmes traités. Cependant, certaines distances dépassent 0,20. Elles séparent *l'Ecole des femmes*, *le Tartuffe*, *le Misanthrope* et les *Femmes savantes* — qui sont écrites en vers — et *Dom Juan*, *l'Avare*, le *Bourgeois gentilhomme* et le *Malade imaginaire* qui sont en prose. Au total, en tenant compte de cette différence, il est évident que tous ces chefs d'œuvre sont du même auteur… Cela est particulièrement net dans certains cas. Ainsi, le *Tartuffe* et *Dom Juan* — les deux pièces qui firent scandale lors de leur présentation — sont l'une en vers (*Tartuffe*), l'autre en prose (*Dom Juan*). De plus, la seconde comporte plusieurs passages en « patois » ce qui augmente encore la distance. Malgré cela, leur proximité est grande (0,199) ce qui indique avec certitude un auteur unique et une contemporanéité de la rédaction (la même remarque vaut également pour *l'Avare* et *Tartuffe*, etc).

Le nombre des œuvres de Molière est trop important pour que l'on puisse reproduire ici la matrice des distances (33 lignes * 33 colonnes). La moyenne des distances séparant chaque pièce à toutes les autres fournit une indication de synthèse (Table III). La moyenne générale est de 0,249, avec un coefficient de variation relative faible (15%). L'œuvre de Molière est donc assez homogène (moins que celle de Corneille mais plus que celle de Racine par exemple) alors que la moitié des pièces sont en vers et l'autre moitié en prose et que l'auteur n'hésitait pas à introduire de nombreux mots en latin, en patois ou de son invention comme le « turc » du *Bourgeois gentilhomme*.

Table III. Distance moyenne d'une pièce à toutes les autres dans le théâtre de Molière.

Titre	Date de création	Nature	Distance moyenne
L'Avare	1668	Prose	0,216
Dom Juan	1665	Prose	0,220
L'Ecole des femmes	1662	Vers	0,220
Le Tartuffe	1664	Vers	0,224
Le Misanthrope	1666	Vers	0,229
L'Ecole des maris	1661	Vers	0,230
Femmes savantes	1672	Vers	0,232
Dépit amoureux	1658	Vers	0,235
Malade imaginaire	1673	Prose	0,235
Fourberies de Scapin	1671	Prose	0,237
L'étourdi	1656	Vers	0,238
Monsieur de Pourceaugnac	1669	Prose	0,239
Bourgeois gentilhomme	1670	Prose	0,239
Georges Dandin	1668	Prose	0,240
Princesse d'Elide	1664	Vers & prose	0,241
Le Sicilien ou l'amour peintre	1667	Prose	0,243
Amphytrion	1668	Vers libres	0,244
L'amour médecin	1665	Prose	0,245
Médecin malgré lui	1666	Prose	0,246
Amants magnifiques	1670	Prose	0,252
Les fâcheux	1661	Vers	0,255
Sganarelle	1660	Vers	0,256
Mélicerte	1666	Vers	0,256
Comtesse d'Escarbagnas	1671	Prose	0,257
Mariage forcé	1664	Prose	0,265
L'impromptu	1663	Prose	0,266
Précieuses ridicules	1660	Prose	0,267
Médecin volant	1659	Prose	0,279
Critique de l'Ecole	1663	Prose	0,280
Dom Garcie	1661	Vers	0,284
La jalousie	1660	Prose	0,310
Psyché Corneille		Vers	*0,293*
Psyché Molière		Vers	*0,305*
Moyenne Molière			**0,249**

Les œuvres s'échelonnent de manière caractéristique : les principaux chefs d'œuvre — l'*Avare*, *Dom Juan*, l'*Ecole des femmes*, l'*Ecole des maris*, les *Femmes savantes*, le *Tartuffe*, le *Misanthrope*, le *Malade imaginaire* — figurent au centre et à des distances moyennes très faibles (il en serait de même du *Bourgeois* sans les « turqueries » placées à la fin de la pièce).

En revanche, d'autres pièces sont plus décalées : les premières comédies que jouaient Molière avant de s'installer à Paris (*La jalousie du barbouillé* et le *Médecin volant*) ou comme les petites pièces de circonstance, à l'image de la *Critique de l'école des femmes* et de *l'Impromptu de Versailles*. Se trouvent également dans ce cas : les *Précieuses ridicules* qui fut le premier succès de Molière et *Dom Garcie*, comédie en vers « sérieuse » qui fut un échec. En dehors de ces quelques pièces, il est pratiquement certain que toute l'œuvre est bien de la même plume.

Le bas du tableau montre que la collaboration avec Corneille apparaît fortement décalée par rapport au reste de l'œuvre, mais il comporte une surprise : la partie de Psyché attribuée à Molière est encore plus décalée que celle due à la plume de Corneille… Au fond, la seule conclusion qu'on puisse tirer de ce dernier constat concerne le caractère atypique de *Psyché* dans l'œuvre de Molière (comme dans celle de Corneille d'ailleurs).

Corneille et Molière

Nous avons opéré la fusion des deux œuvres dans un corpus unique (voir la liste en annexe). Outre Psyché, ce corpus comporte 64 pièces, soit 917.000 mots, dont la rédaction s'étend sur 44 ans (1630-1673). Il mêle comédies, tragédies, pièces en vers et en prose, et aborde des thèmes extraordinairement divers. L'ensemble reste malgré tout assez homogène (distance moyenne entre les pièces : 0,280), plus homogène que la seule œuvre théâtrale de Racine (0,289) pourtant entièrement versifiée en alexandrins et que tous les corpus de cette taille — même à auteur unique — qu'il nous a été donné de traiter jusqu'à maintenant…

Pour obtenir une vision d'ensemble, deux expériences de classification ont été menées.

En premier lieu, on a procédé à une classification automatique ascendante sur la matrice des distances. Les deux pièces les plus proches sont regroupées et les distances de ce nouvel ensemble avec toutes les autres pièces sont recalculées pour le regroupement suivant. Les étapes de la classification sont résumées dans un dendrogramme (figure ci-dessous). L'ordre des regroupements se lit de la gauche vers la droite avec, en ordonnées, les distances correspondantes aux différents niveaux d'agrégation (l'origine est placée à 0,15 afin de rendre le graphe lisible mais cela ne doit pas faire oublier la grande proximité de la plupart de ces pièces).

Cette expérience montre que les oeuvres sont distinctes quoique proches : les deux corpus se rejoignent à 0,280. Il y a donc une parenté avec au moins une partie des pièces de Corneille. Mais surtout, il s'opère un curieux chassé-croisé entre les deux oeuvres :

— une pièce de Molière s'inscrit au milieu des pièces de Corneille : *Dom Garcie*. Cette pièce est donc très probablement de la main de Corneille. Sa proximité avec *Pertharite* (1651) ne laisse pas de doute sur l'époque où Corneille l'aurait écrite (après l'échec de *Pertharite*, Corneille a abandonné le théâtre pendant près de dix ans…)

— deux comédies de Corneille (*Le Menteur* et la *Suite du Menteur)* viennent se placer au milieu des pièces de Molière. Ce qui surprend relativement puisque ces comédies (les deux dernières écrites officiellement par Corneille) datent de 1642-43 alors que les premières

pièces de Molière sont supposées avoir été écrites au plus tôt en 1656 et n'ont été jouées à Paris qu'à partir de 1660. Autrement dit, puisque Corneille est l'auteur incontestée des deux *Menteurs*, il est très probablement aussi celui du bloc de pièces situées sur le dendrogramme à la gauche de ces deux pièces et qui sont toutes fort proches les unes des autres : le *Tartuffe*, le *Misanthrope*, les *Femmes savantes*, *L'étourdi*, le *Dépit amoureux*, *l'Ecole des maris*, *l'Ecole des femmes*, *Sganarelle*, *Amphytrion*, la *Princesse d'Elide*, *Mélicerte* et les *Fâcheux*. C'est-à-dire toutes les pièces en vers de Molière…

En revanche, la parenté est moins évidente pour les pièces en prose, bien qu'on puisse remarquer qu'elle rejoignent les pièces en vers à une distance inférieure à celle où les deux *Psyché* rejoignent l'oeuvre de Corneille. La chose est donc très possible mais moins claire que pour les pièces en vers.

Ces constats d'ensemble suggèrent d'examiner plus en détail les proximités existantes notamment autour des deux *Menteurs* (Table IV).

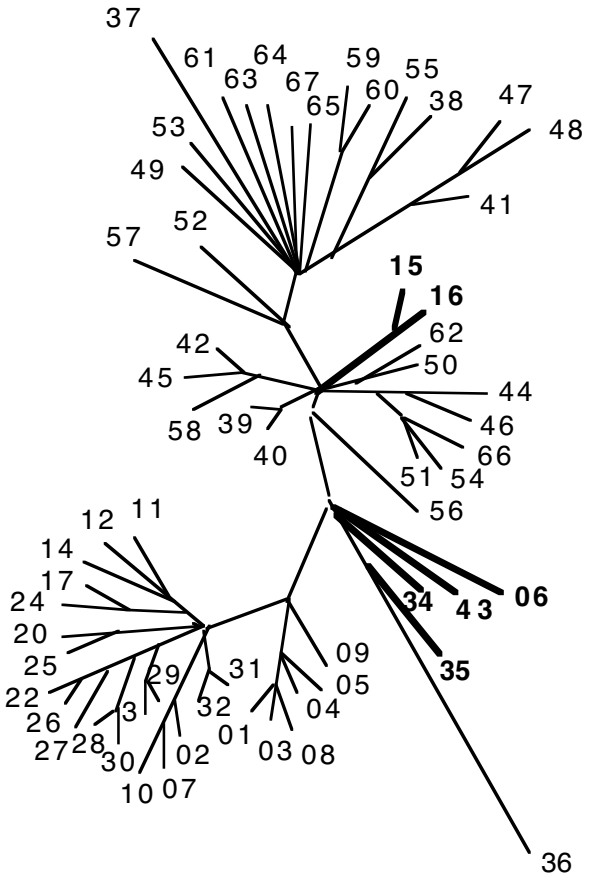
Outre les pièces en vers, *Dom Juan* et l'*Avare* figurent en-dessous du seuil de 0,25 et peuvent donc être supposés avoir été écrits par le même auteur que les *Menteurs*… La partie de *Psyché* écrite par Corneille ainsi que *Dom Garcie* suggèrent même un seuil plus élevé (0,273), c'est-à-dire une collaboration probable de Corneille à *Amphitryon*, aux *Fourberies de Scapin*, voire au *Malade imaginaire* (mais ici, le « latin de cuisine » et les intermèdes en italiens contribuent à augmenter la distance).

Les deux *Menteurs* ont probablement fourni le modèle pour les pièces de Molière, spécialement celles en vers, ce qui explique que la classification rattache ces deux *Menteurs* à l'œuvre de Molière et non à celle de Corneille. Cependant, toutes les distances sur le tableau IV excèdent nettement 0.20 (sauf *l'Etourdi*) de telle sorte que la paternité de Corneille sur les comédies en vers de Molière est probable et non certaine. Le temps est une explication possible. En effet, les comédies en vers de Molière ont été écrites 20 ans et plus après les *Menteurs*.

Pour vérifier ces conclusions, nous avons eu recours à une seconde technique de classification : l'analyse « arborée » qui classe les textes ou groupes de textes, non plus deux à deux, mais en considérant, pour chacun, la meilleure représentation possible de sa distance par rapport à tous les autres (Barthélémy et Guénoche, 1988, Luong, 1994, Juilland et Luong,1997). Chaque texte constitue une terminaison (feuille) qui est reliée aux autres par des branches plus ou moins longues et par des sections de tronc. La proximité relative de deux textes se mesure par le chemin à parcourir pour les unir (graphe ci-dessous).

Table IV. Distances des deux Menteurs avec les pièces de Molière

Texte	Le Menteur 1	Le Menteur 2
Le Menteur 1	0,000	0,180
Le Menteur 2	0,180	0,000
Psyché Corneille	0,288	0,273
Psyché Molière	0,329	0,325
La jalousie	0,341	0,331
Médecin volant	0,310	0,293
L'étourdi	**0,205**	**0,206**
Dépit amoureux	**0,215**	**0,212**
Précieuses ridicules	0,315	0,314
Sganarelle	0,259	0,253
Dom Garcie	0,280	0,273
L'école des maris	**0,223**	**0,217**
Les fâcheux	**0,248**	**0,248**
L'école des femmes	**0,226**	**0,217**
Critique de l'école	0,323	0,319
L'impromptu	0,321	0,316
Mariage forcé	0,322	0,302
Princesse d'Elide	0,251	**0,243**
Le Tartuffe	**0,242**	**0,228**
Dom Juan	0,259	**0,248**
L'amour médecin	0,292	0,289
Le Misanthrope	0,252	**0,234**
Médecin malgré lui	0,298	0,289
Mélicerte	0,257	**0,250**
Le sicilien	0,277	0,260
Amphytrion	0,253	0,256
Georges Dandin	0,292	0,279
L'Avare	0,256	**0,244**
M. de Pourceaugnac	0,292	0,283
Amants magnifiques	0,282	0,279
Bourgeois gentilhomme	0,294	0,280
Fourberies de Scapin	0,269	0,263
Ctesse d'Escarbagnas	0,311	0,300
Femmes savantes	0,260	**0,248**
Malade imaginaire	0,282	0,270
Moyenne Molière	0,275	0,266



Le graphique est dû à l'obligeance de M. Xuan Luong de l'Université de Nice. Pour les titres des pièces, se reporter à l'annexe. Pour améliorer la lisibilité, ce graphe a été établi en retranchant 0,15 à toutes les distances. Tout comme le dendrogramme ci-dessus, il exagère donc les proximités entre les feuilles terminales et l'éloignement des principaux noeuds.

Les pièces de Corneille (numérotées de 1 à 33) figurent pratiquement toutes en bas à gauche du graphe ; celles de Molière (de 37 à 67) se trouvent presque toutes dans la partie haute du graphe. En bas, à droite, figurent en gras, quatre « anomalies » : les deux parties de *Psyché* écrites par Corneille (n° 34) et par Molière (n° 35) ; *Dom Garcie* (43) et le cinquième acte de la *Comédie des Tuileries* (06) qui est une commande écrite par Corneille pour le Cardinal de Richelieu. Pour les comédies de Molière, le graphe partage nettement celles écrites en prose (le groupe situé tout en haut) et celles écrites en vers (au milieu).

L'analyse arborée de X. Luong confirme surtout la principale conclusion de la classification hiérarchique : les deux *Menteurs* de Corneille (15 et 16) se placent pratiquement au centre de l'œuvre de Molière alors qu'ils sont nettement plus éloignés du reste de celle de Corneille. Autrement dit, le rédacteur de ces deux pièces est très probablement aussi celui de la plupart des œuvres signées par Molière…

La contribution très probable de P. Corneille à tous les chefs d'oeuvre de Molière n'enlève rien à l'importance historique de ce dernier. Il fut en quelque sorte le premier « entrepreneur de spectacle » moderne, à la fois directeur de troupe, metteur en scène, acteur et, comme on le voit, excellent chasseur de « scénarios »…

Naturellement, on peut aussi voir dans ces résultats la preuve de l'influence considérable de P. Corneille sur le théâtre du XVIIe siècle et considérer que Molière a repris le projet d'une vaste satire des mœurs de son époque, projet que Corneille avait semblé abandonner après l'échec du deuxième *Menteur*… On peut ajouter que Molière mettait en scène et jouait régulièrement les pièces de Corneille, ce qui a pu contribuer à son « imprégnation » et le conduire à écrire « comme » son auteur préféré dont il connaissait par cœur des milliers de vers. Des études plus approfondies seront nécessaires pour répondre en détail à ces objections à notre calcul.

Nous ferons d'ailleurs remarquer qu'on ne pourra vraiment affirmer que Corneille est bien l'auteur principal des pièces de Molière qu'après avoir examiné les pièces des principaux auteurs contemporains susceptibles d'avoir joué le même rôle auprès de l'illustre comédien notamment pour la rédaction de certaines des comédies en prose trop éloignées pour avoir été l'œuvre de Corneille (comme les *Précieuses ridicules*)…

Notre analyse ne prétend donc pas clore un débat séculaire. Elle débouche aussi sur plusieurs questions intéressantes. On peut se demander notamment quelles sont les particularités du style ou du vocabulaire de Corneille que l'on retrouve chez Molière et, à l'opposé, quelles sont les différences et les singularités.

Au-delà de cet exemple, nous espérons que d'autres travaux viendront confirmer la grande puissance de la distance intertextuelle, combinée à la classification automatique, comme outil pour l'attribution d'auteur. A ce sujet, nous voudrions souligner qu'un tel calcul exige au préalable que les graphies des textes comparés aient été normalisées mais aussi, à notre avis, que les mots aient été « lemmatisés », c'est-à-dire rattachés à leurs entrées de dictionnaire ou « vocables » (voir Labbé, 1990).

Par exemple, comparer de la prose et de la poésie, sans réduire les majuscules initiales des vers, engendre automatiquement une distance d'environ au moins un septième entre ces deux sous-corpus puisque, en moyenne, un vers contient entre cinq et sept mots… Dès lors on peut être certain qu'un tel calcul, effectué sur un corpus non normalisé, mettra d'un côté toutes les œuvres en prose et, de l'autre, tous les poèmes, sans que le contenu soit forcément en cause. D'autres exemples viennent à l'esprit : dans sa correspondance un auteur utilisera

abondamment les abréviations (M. pour monsieur, les initiales des noms et des prénoms, etc.) facilités qu'il proscrira de ses œuvres… Est-il légitime de considérer qu'il s'agit d'une différence de vocabulaire ?

Tout calcul de la distance exige donc au préalable que l'on se mette d'accord sur une norme unique de mesure un peu comparable au mètre étalon…

Bibliographie

BARTHELEMY Jean-Pierre, GUENOCHE Alain (1988), *Les arbres et les représentations des proximités*, Paris, Masson.

BARTHELEMY Jean-Pierre, LUONG Xuan (1998), "Représenter les données textuelles par des arbres", in Sylvie MELLET (ed), *4e journées internationales d'analyse statititistique des données textuelles*, Université de Nice, 1998, p. 49-71.

BINONGO José N., SMITH M.W.A. (1999), « The Application of Principal Component » Analysis to Stylometry, *Literary and Linguistic Computing*, 14-4, p 445-465.

BRUNET Etienne (1988), "Une mesure de la distance intertextuelle : la connexion lexicale", *Le nombre et le texte. Revue informatique et statistique dans les sciences humaines*, Université de Liège.

FORSYTH Richard S. (1999), « Stylochronometry with Substings, or : a Poet Young and Old », *Literary and Linguistic Computing*, 14-4, p 467-477.

HOLMES David (1995), « The Federalist revisited : new directions in autorship attribution », *Literary and Linguistic Computing*, 10-2, p 111-127.

JACCART P. (1908), "Nouvelles recherches sur la distribution florale", *Bull. Soc. Vand. Sci. Nat.*, 44.

JUILLARD Michel, LUONG Xuan (1997), « Words in the Hood : a New Look at the Distribution of Word in Texts », *Literary and Linguistic Computing*, 12-2, p 71-78.

LABBE Dominique, MONIERE Denis (2000), « La connexion intertextuelle. Application au discours gouvernemental québécois », Martin RAJMAN et Jean-Cédric CHAPPELIER (eds), *Actes des 5e journées internationales d'analyse des données textuelles*, Lausanne, Ecole polytechnique fédérale, vol 1, p 85-94.

LUONG Xuan (1994), « L'analyse arborée des données textuelles : mode d'emploi », *Travaux du cercle linguistique de Nice*, 1994, 16, p 25-42.

MULLER Charles, BRUNET Etienne (1988), « La statistique résout-elle les problèmes d'attribution ? », *Strumenti Critici*, septembre 1988, p 367-387.

MULLER Charles (1997), *Principes et méthodes de statistique lexicale*, Paris, Hachette université, 1977.

ROBERTS F.S. et Al (1971), *Measurement Theory*, Addison-Wesley, Reading.

TOMASSONE Richard et Al (1988*), Discrimination*

WOUTERS Hippolyte, VILLE DE GOYET, Christine de (1990), *Molière ou l'auteur imaginaire ?*, Bruxelles, Eds Complexe.

Annexe I. Les œuvres de Corneille et de Molière

Corneille		Année de création	Genre	Longueur (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	1 299
Molière				
37	La jalousie	1660	Comédie en prose	3 501
38	Médecin volant	1660	Comédie en prose	3 876
39	L'étourdi	1660	Comédie en vers	18 671
40	Dépit amoureux	1660	Comédie en vers	16 242
41	Précieuses ridicules	1660	Comédie en prose	6 648
42	Sganarelle	1660	Comédie en vers	6 042
43	Dom Garcie	1661	Comédie en vers	17 049
44	L'école des maris	1661	Comédie en vers	10 536
45	Les fâcheux	1661	Comédie en vers	7 922

46	L'école des femmes	1662	Comédie en vers	16 625
47	Critique de l'école	1663	Comédie en prose	8 610
48	L'impromptu	1663	Comédie en prose	7 168
49	Mariage forcé	1664	Comédie en prose	6 058
50	Princesse d'Elide	1664	Comédie en vers et prose	11 333
51	Le Tartuffe	1664	Comédie en vers	18 271
52	Dom Juan	1665	Comédie en prose	17 452
53	L'amour médecin	1665	Comédie en prose	6 147
54	Le Misanthrope	1666	Comédie en vers	17 180
55	Médecin malgré lui	1666	Comédie en prose	9 317
56	Mélicerte	1666	Comédie en vers	5 540
57	Le sicilien	1667	Comédie en prose	5 375
58	Amphytrion	1668	Comédie en vers libre	15 117
59	Georges Dandin	1668	Comédie en prose	11 009
60	L'avare	1668	Comédie en prose	21 033
61	M. de Pourceaugnac	1669	Comédie en prose	11 803
62	Amants magnifiques	1670	Comédie en vers et prose	11 983
63	Bourgeois gentilhomme	1670	Comédie en prose	17 132
64	Fourberies de Scapin	1671	Comédie en prose	14 245
65	Comtesse d'Escarbagnas	1671	Comédie en prose	5 564
66	Femmes savantes	1672	Comédie en vers	16 863
67	Malade imaginaire	1673	Comédie en prose	19 919