Inter-Textual Distance and Authorship Attribution. Corneille and Molière

Cyril Labbé, Dominique Labbé

Inter-textual distance and autorship attribution
Corneille and Molière

Cyril LABBE*, Dominique LABBE**

* Université Grenoble I
cyril.labbe@imag.fr

** CERAT-IEP, BP 48 – 38040 GRENOBLE Cedex 9.
dominique.labbe@iep.upmf-grenoble.fr

English preliminary version of:

"Inter-Textual Distance and Authorship Attribution. Corneille and Molière".

Published in: Journal of Quantitative Linguistics. 8-3, December 2001, p 213-231.

Summary :

The calculation proposed in this paper, measures neighborhood between several texts. It leads to a normalized metric and a distance scale which can be used for authorship attribution. An experiment is presented on one of the famous cases in French literature : Corneille and Molière. The calculation clearly makes the difference between the two works but it also demonstrates that Corneille contributed to many of Moliere's masterpieces.

> «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.

The authorship research of an unknown or doubtful text is one of the oldest statistical problems applied to literature. The unknown text is to be compared with other texts where we are sure that we know the author or we are sure that he wrote at least a part of it. Usually, the study concerns the most frequent words or a selection of them, often the «function words ». On this topic, see (Holmes, 1995), (Baayen and al, 1996) and (Binongo, 1999). In this paper, we propose a calculation which considers the entire text and which gives a standardized measure of the actual distance between it and another text. This is known as «lexical connection» defined as «the intersection of two texts vocabularies» (Muller 1977). Therefore, connection is the complement of distance, a colloquial term in statistics ; for this reason, we have chosen it.

To understand our calculation, one may consider the difference between «token» and «type». The token is the smallest measurable element in a text, and the «type» forms the vocabulary's basic element. For instance, the longest novel in French, *Les misérables* is made up of half a million tokens : its length or extent (noted N), while its vocabulary (noted V) is made up of less than 10 000 normalized and tagged types.

Usually, the «connection» measure is done on the vocabulary regardless of the type frequency (see Brunet, 1988). Here , we suggest to consider the frequency of each type, that is to say, the entire texts (we use the adjective «textual» in order to show that the calculation is on N and not only on V or on a part of V).

Our metric measures whether two or several texts are relatively far from one another. It has been applied to a lot of corpora and used to set up a useful distance scale for authorship attribution. We present an application to one of the most famous cases in French literature : Corneille and Molière. The measure makes clear the difference between their works but it also proves that Corneille probably wrote a lot of Moliere's plays.

I**ntertextual distance**

To be allowed to say whether two texts are rather near or far from one another, if we consider their extents, we must use a «metric» with the following properties:

— non sensitive to length differences of the compared texts;

— applicable to several texts and, if possible, to all texts in the same language;

— varying in the same way – between 0 (the same vocabulary and similar frequency of each type in the 2 texts) and 1 (no common type) – without jump, nor threshold effect around some values.

— symmetric (given 2 texts A and B then : $\delta_{(A,B)} = \delta_{(B,A)}$);

— as «transitive» as possible: when we «agregate» 2 texts, the distance of this «corpus» regarding other texts must reflect the prior distance in the ordering ($\delta_{(A,B)} < \delta_{(A,C)} < \delta_{(B,C)}$ then $\delta_{(A,B)} < \delta_{\{A,(B \cup C)\}}$;

— as «robust» as possible (ie: a marginal change in one of the 2 texts must be reflected by a marginal change in their distance…)

Some previous studies in this field, especially Muller's and Brunet's ones, suggest the following method.

Given 2 texts A and B:

V_a and V_b: number of types in A and B

F_{ia}: frequency of the ith type in A

F_{ib}: frequency of the ith type in B

N_a and N_b: number of tokens in A and B with $N_a = \sum F_{ia}$ and $N_b = \sum F_{ib}$

The absolute distance between A and B will be the union of the 2 texts less their intersection, $(N_a \cup N_b) - (N_a \cap N_b)$, that is to say the sum of the differences between the absolute frequencies of each type in the 2 texts.

The relative distance can be computed in two ways:

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

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

Formula (2) is the one given by E. Brunet (1988). Two objections to these formulae can be found.

— (1) and (2) are equivalent only when the texts lengths are equal (Na = Nb). If no type is shared, the two formulas actually give a result of 1 whatever the text length is (which is one of the conditions for our idealistic metric);

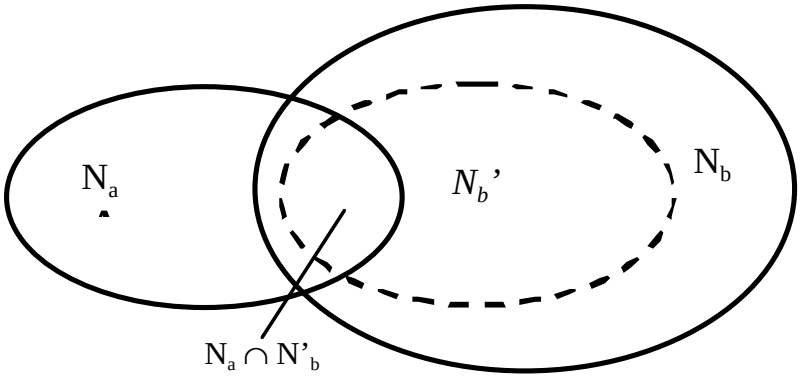
— on the other hand, the theoretical minimum can reach 0 only in the specific case of equal lengths. As a matter of fact, the greater the difference in length between the two texts, the further the minimal numerator will be from 0. For instance, in Molière's corpus, the shortest text counts 732 tokens and 274 types (it is a piece from a lost play : *Pastoral Comedy*). On the other hand, the longest play is the *Malade imaginaire* (19 920 tokens and 2 082 types). Even if the small text was completely included in the large one, the distance would not be null since there is not enough room in the small text for all the types of the long one;

— in (1) as in (2), the intersection of the 2 texts is counted twice. Therefore, more importance is given to the common types rather than to the specific vocabulary of each text.

Is it possible to overcome these two objections and allow a good approximation of the distances between several texts ?

An approximation of intertextual distance

In order to get an accurate estimate of the distance between several texts, we propose to «reshape» the largest to the size of the smallest. Define B' this reduction of B to the size of A:



The mathematical expectation of every type of B with a frequency f_i is:

$$E_{ia(u)} = F_{ib} * U_{(a,b)} \text{ with } U_{(a,b)} = \frac{N_a}{N_b}$$

This gives the value of N'_b:

$$N'_b = \sum_{V_b} E_{ia(u)}$$

Consequently, we can reformulate (1) and (2) replacing F_{ib} by $E_{ia(u)}$ and N_b by N'_b.

Zero, the theoretical minimum will be reached when the small text is like a model of the largest. In this case, all the types of A are present in B with a frequency $F_{ia} = E_{ia(u)}$ and, consequently, the numerators of the formulae will be equal to zero. In fact, $2N_a$ is the maximum token population that the two texts can share if they have the same size, the same vocabulary and equal frequencies for each type. Conversely, the theoretical maximum (one) means that A and B do not share any type : in this case, both numerator and denominator are equal to $N_a + N'_b$.

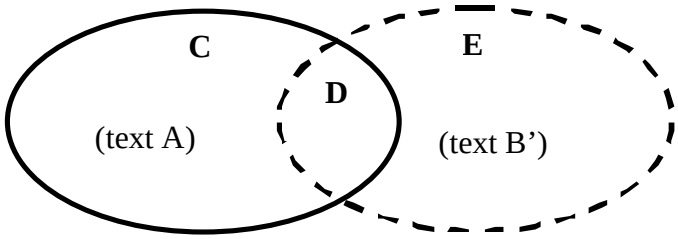
However, this new formulation gives no answer to the double count objection about the intersection of the two texts and does not entirely solve the «physical» problem noted above : if the lengths of the two texts are very different, all the types of the largest cannot be used in the smallest.

To accomodate and to allow an unbiased measurement of the distance, we propose to:

— consider the intersection of the two texts only once;

— limit calculations to all the types of A *and* the only types of B whose frequency is high enough to expect almost one in A ($E_{ia(u)} \geq 1$). The sum of these expectations is N'_b.

Consequently, calculation is done in two steps (see the figure below):



Firstly, the V_a types : C and D (in the C set, $E_{ia(u)} = 0$) and, secondly, the types of E : V'b(e) (in this case : $F_{ia} = 0$). The absolute distance between A and B' is :

$$D_{Va,\,b(u)} = \sum_{Va,\ V'b(E)} \left| F_{ia} - E_{ia(u)} \right|$$

When A and B share no type, this distance will be equal to : $N_a + N'_b$. This will be the numerator of the relative distance formula since the metric maximum is 1 and the actual result must be less than 1 when the intersection of A and B is not empty.

$$(3)\quad D_{(a,b)} = \frac{\sum_{Va,\ V'b(e)} \left| F_{ia} - E_{ia(u)} \right|}{\sum_{Va} F_{ia} + \sum_{V'b} E_{ia(u)}} = \frac{\sum_{Va,\ V'b(e)} \left| F_{ia} - E_{ia(u)} \right|}{N_a + N'_b}$$

It is worth noting that:

— the same result, rounding excepted, can be obtained by subtracting the relative frequencies of the two texts, if one considers all the vocabulary of the smallest text (A) and only the B types whose frequencies are high enough to expect at least one if B is reduced to the size of A.

— the metric accuracy is slightly reduced by rounding. In fact, the observed frequencies are always integers whereas mathematical expectations include decimals which will contribute to the distance. This drawback will increase when low frequency types are an important part of the texts, that occurs in the case of small texts. To overcome this, it is convinient not to apply the calculation to too small texts —we never applied this calculation under the limit of 1 000 tokens (so that the small excerpt of the *Comédie pastorale* cannot be examined) — and to avoid a too large scale of sizes (under 1/10). In the application above, the shortest text counts 3 500 tokens — it is Molière's first comedy (see the appendix) and

the largest counts 20 300 (Corneille's *Toison d'or*)[1]. For the same reasons, all results under .50 are eliminated from the numerator ($|F_{ia} - E_{ia(u)}| < .5$).

— this calculation means that, beforehand, the texts are normalized and – from our point of view – that all the tokens are tagged (in French : «lemmatisés»), i.e. attached to their dictionary entries (Muller, 1977 and Labbé, 1990). For example, comparing prose and poetry pieces, without reducing to lower case the verses initial upper case, automatically creates a distance of around 1/8 since a verse counts around 6-10 words. The distance calculation applied to a non-normalized corpus will place on one hand all pieces of prose and on the other hand all poetry, even if both content are not different… Other examples exist: in his letters, an author may use a lot of abbreviations (Mr for «mister», initials for names, etc.) but not in his works: is this an actual vocabulary difference ? One can see that the distance calculation implies a prior agreement on standards.

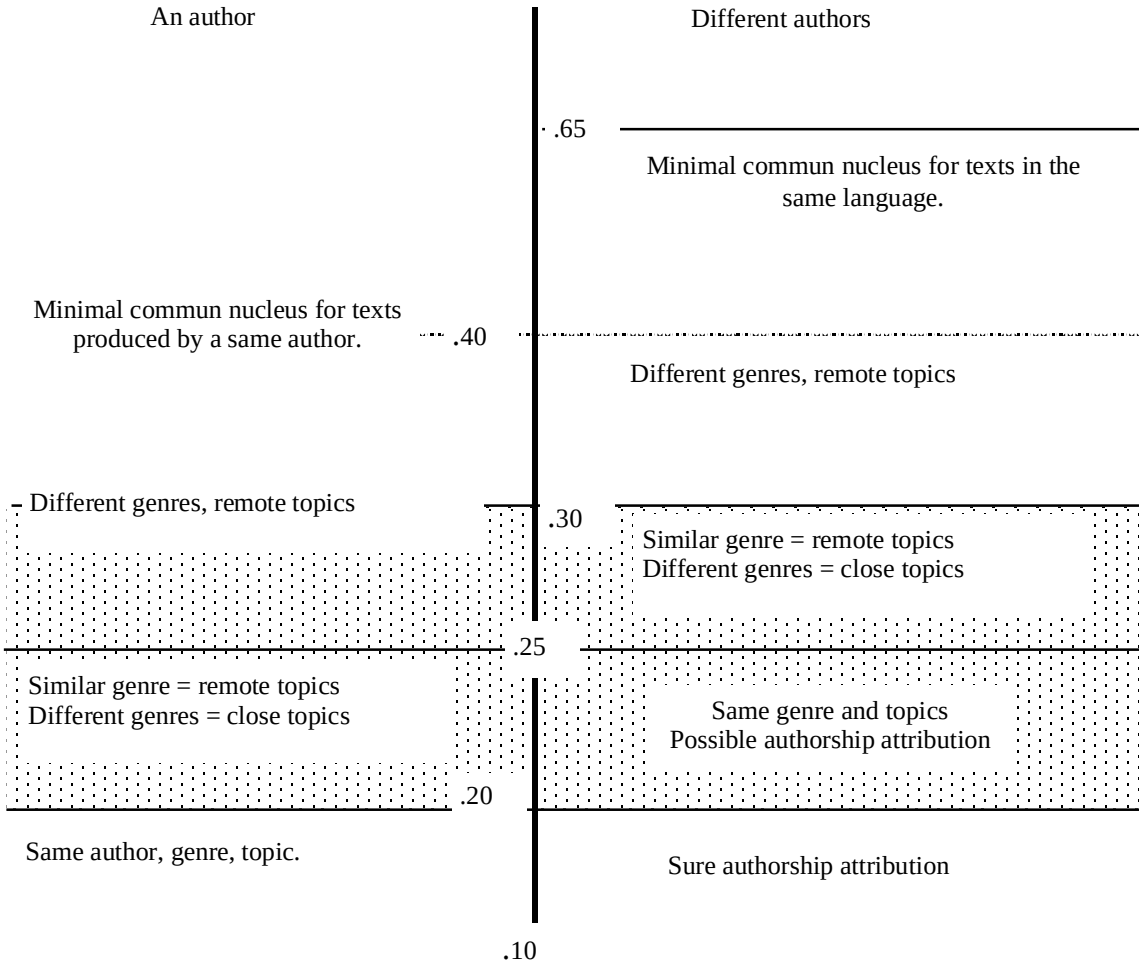
— the interpretation of the results is very easy. For example, a metric value of .50 means that we can estimate that the two texts share half of their whole extent; .25 that three quarters of the two texts are common, etc. Thus, a scale of distances can be established, which can be useful for authorship attribution.

Distance scale

This calculation has been applied to various corpora the total size of which is about 10 million of tokens all counted with the same standards : General de Gaulle's and F. Mitterrand's speeches, Canadian and French Prime minister's adresses to parliament since 1945 (Labbé-Monière, 2000), several novels from the last 3 centuries (with E. Brunet), Trade Unions newspapers editorials (Labbé-Brugidou, 1999), economic press articles, transcription of interviews (Bergeron-Labbé, 2000). These experiments have been used to establish the following empirically distance scale.

[1] (2007) In fact : the longest play is l'*Avare* (21 033 tokens).

Table 1. Intertextual distance standardized scale[2].



— for the same author, we always notice distances smaller than those existing between two different and contemporary authors (when they are dealing with the same topic).

— distances smaller than .20 usually do not exist between two different authors (concerning texts of the same kind with close topics). In the case of an unknown writer, authorship attribution is quite sure. If it is known that both authors are different, then one of them was «inspired» by the other.

— between 0.20 and 0.25 represents the case where the texts are very similar. In the case of only one author, a change in themes and genre is indicated. If one of the authors is unknown, attribution is possible but it will be sure only if it is proved that there are no other texts nearer and if one can provide other proofs, particularly stylistic.

[2] (NB 2007) : this distance scale has been calibrated with the help of contemporaneous texts the lengths of which were over 3500 and under 20 000 tokens.

— above .25, authors are probably different or genre and topics are too far to allow a comparison ;

As an example, we present an application of the calculation to Corneille and Molière's plays. These works have often been analysed by critics and even in statistic studies (especially, Muller 1967; Kylander, 1995) that gives some useful references.

From the very beginning, it was rumoured that Molière was not the writer of his plays. These rumours were intensified by a publisher's «warning» placed at the head of one play : *Psyché* (1671). It was said that although Corneille wrote two thirds of the verses, it had previously been played under Molière's name (this play and the publisher's warning are published in the second volume of Corneille's complete works in La Pléiade, Gallimard). Since then, the problem has been discussed many times; most often, Corneille is said to be the virtual author; among others, the poet P. Louys at the beginning of the XXth century, and more recently, two Belgian writers have underlined how similar the two works are (Wouters and Ville de Goyer, 1990).

Moliere's plays

Intertextual distance calculus gives some interesting information. Firstly, as an example, one can find below the distances separating the most well known Molière's plays (table II).

Table II. Distances between Molière's well known works.

	Ecole des femmes	Tartuffe	Dom Juan	Le Misanthrope	L'Avare	Bourgeois gentilh.	Femmes savantes	Malade imaginaire
Ecole des femmes	0	.183	.205	0.194	0.200	.231	.198	.223
Le Tartuffe		0	.199	.167	.199	.230	.170	.219
Dom Juan			0	.204	.170	.207	.219	.205
Le Misanthrope				0	.210	.239	.173	.239
L'Avare					0	.194	.214	.187
Bourgeois gentilh.						0	.234	.196
Femmes savantes							0	.226
Malade imaginaire								0

The calculation shows an important similarity between all these plays, although their topics are very different. The smallest (.167) is between *Tartuffe* and *le Misanthrope*, two

plays in Alexandrines in which Molière does not use farce nor colloquial language, nor jargon. The greatest (.239) is between *le Misanthrope* and *le Bourgeois gentilhomme* or *le Malade imaginaire*. The first one is in verse, the two others in prose and they contain a lot of inventions in «turkish» or in «latin». More generally, distances greater than .20 separate *l'Ecole des femmes, Tartuffe, le Misanthrope* and *les Femmes savantes* — written in verse — and *Dom Juan, l'Avare, le Bourgeois gentilhomme* and *le Malade imaginaire,* written in prose. Considering these differences, it is obvious that all these masterpieces are from the same author. Some cases seem particularly clear : *Tartuffe* and *Dom Juan* —two plays which caused scandal and were withdrawn — are written with, the first in verse and the second in prose. In spite of a lot of «patois» in the second one, which increases their distance, they remain very close (.199): this confirms that they have only one author and that they were written during the same period (the same comment can be said for *l'Avare* and *Tartuffe*).

Moliere's plays are too numerous to reproduce here their distances matrix (33 lines and 33 columns). The mean of the distances separating each play from all the others gives some information (table III). The overall mean is .249, with a small relative variation coefficient (15%). Thus Molière's works are rather homogeneous, less than Corneille's (.230), but more than Racine's (.290) although half of Molière's plays are in verse and the rest in prose and although he used a lot of «latin» words, some «patois» and imaginary language.

Table III. Overall distances between one play and all the others in Molière's works.

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

Comedies spread out in a very caracteristic way : the main masterpieces — *l'Avare*, *Dom Juan*, *l'Ecole des femmes*, *l'Ecole des maris*, *les Femmes savantes*, *le Tartuffe*, *le Misanthrope*, *le Malade imaginaire* — stay in the center and at small mean distances (it would be the same for the *Bourgeois* if the «Turkish» language was not put at the end of this play). On the other hand, some plays are apart : the first comedies Molière played before living in Paris (la *Jalousie du barbouillé*, *le Médecin volant*) or some small occasional creations like *la Critique de l'Ecole des femmes* et *l'Impromptu de Versailles*. In the same case, we find *les Précieuses ridicules* (first of Molière's success) and *Dom Garcie*, a serious verse comedy which was unsuccessful. Except these few plays, it is quite sure that all the work is from a single author.

The bottom of the table shows that Corneille's admitted contribution appears rather far from the rest of the work but it contains a surprising fact : Molière's *Psyché* is further apart than Corneille's one. As a matter of fact, the only conclusion to be drawn from this last measure concerns the atypical position of *Psyché* in Molière's work (as well as in Corneille's one).

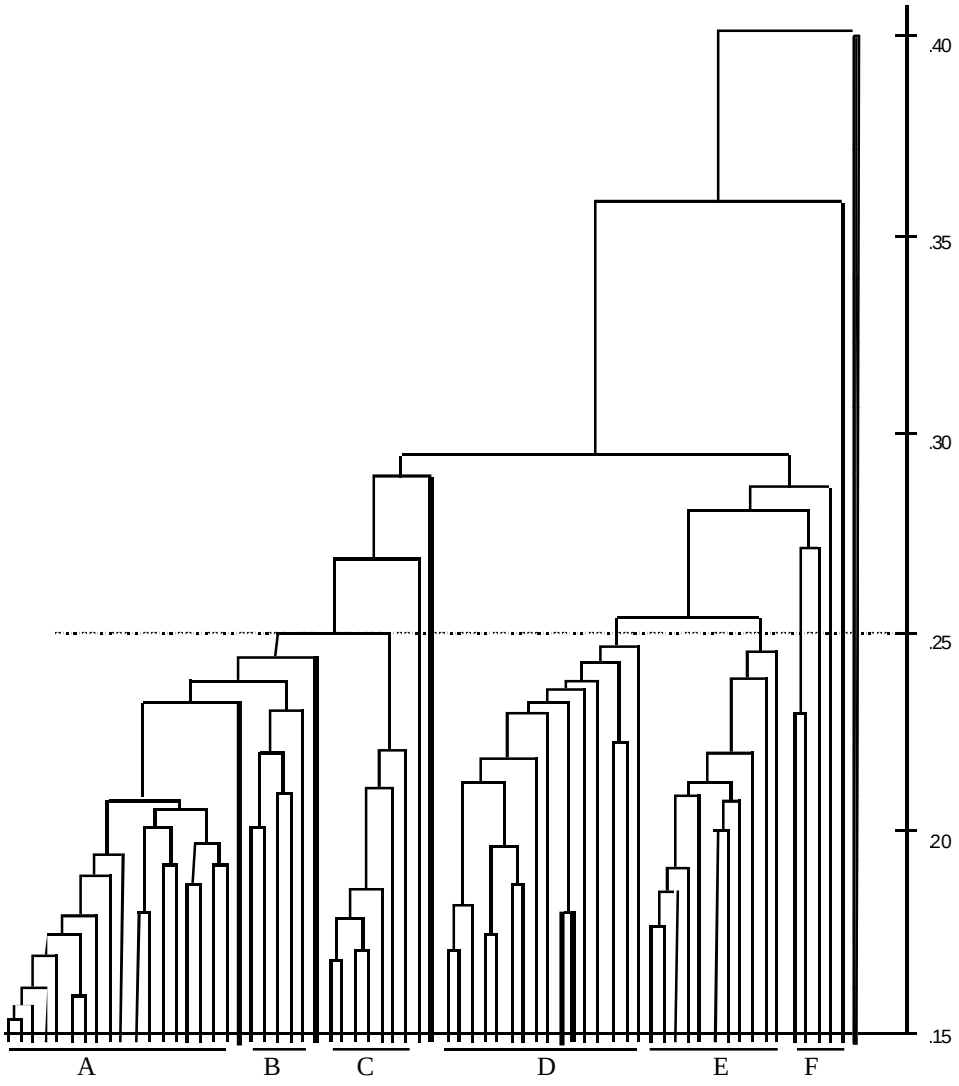
Corneille and Molière

The two works were merged in a single corpus (see the annex list). Besides Psyché, this corpus consists of 64 plays that is to say 917 000 tokens, the writing of which spreads over 44 years (1630-1673). It mingles comedies, tragedies, verses and prose plays, and tackles extremely diverse themes. However, the whole work remains more homogeneous than Racine's only dramatic work, which is much smaller (166 000 words) and all in alexandrine verses, and than all large corpora — even with a single author — we have dealt with until now.

To obtain an overall view, two classification experiments were carried out. The first one was a cluster analysis on the distances matrix. The two nearest plays are merged and the distances of this new set with respect to all other plays are calculated again for the following grouping. The classification steps are summed up in a dendrogram (Table IV): from left to right, the regrouping order and, as ordinate, distances corresponding to the different agregation stages. The origin is placed on .15 in order to enable an easy reading of the graph, but it must not be forgotten that all these plays are very near.

This first experiment brings to light that the works are different but near; the two corpora link at .28. Moreover, regrouping fits with what is expected. On the left, the most homogeneous group is made up of Corneille's mature tragedies (group A), then come his first tragedies (B) that made him famous (*le Cid*, *Horace*, *Cinna*…), and, finally, his comedies (C). As for Molière, the classification separates his verse comedies (D) and prose ones (E, F). In finer detail, most basic groupings correspond to thematical proximity already remarked by critics. For instance, Molière's *les Femmes savantes*, *l'Ecole des Mar*is and *l'Ecole des femmes*.

Table IV Cluster analysis on Corneille and Molière's plays



From left to right :

A. Corneille :
Tite et Bérénice
Pulchérie
Suréna
Agésilas
Othon
Sertorius
Sophonisbe
Atilla
Nicomède
Don Sanche
Polyeucte
Théodore
Héraclius
Pertharite
Andromède
Toison d'Or
Rodogune
Oedipe

Dom Garcie

B : Corneille
Cinna
Pompée
Le Cid
Horace
Médée

Psyché Corneille

C Corneille comédies
Galerie du Palais
La Suivante
Mélite
La Veuve
La Place Royale
L'illusion comique
Clitandre
Comédie des Tuileries

Psyché Molière

D. Molière (verse)
Le Tartuffe
Le Misanthrope
Femmes savantes
L'étourdi
Dépit amoureux
L'école des maris
L'école des femmes
Amphytrion
Sganarelle
Le Menteur 1
Le Menteur 2
Mélicerte
Les fâcheux
Princesse d'Elide

E. Molière (prose)
Amants magnifiques
Le sicilien
Georges Dandin
L'avare
Dom Juan
Fourberies Scapin
Médecin malgré lui
M. Pourceaugnac
Malade imaginaire
Bourgeois gentil.
L'amour médecin
Mariage forcé
Ctesse d'Escarb.

F. Molière :
Critique de l'école
L'impromptu
Précieuses ridicules
Médecin volant
La jalousie

Psyché Quinault

The distance calculation combined with cluster analysis, thus provides an accurate and reliable tool. However , in this corpus, this tool detects some «anomalies» (bold lines) :

— one of Molière's play is found in the middle of Corneille's ones : *Dom Garcie*. Very probably, this play is from Corneille. As a matter of fact, it is very near to those he wrote in the period when *Dom Garcie* was created.

— the two *Psyché* are placed together in Corneille's work. The one ascribed to Corneille is almost at an equal distance between comedies and tragedies — this is logical — and Molière's one nearly between the two works.

— two of Corneille's comedies (the *Menteur* and the *Suite du Menteur*) are found in the middle of Molière's verse plays. This classification is very surprising because these comedies (the two last ones officially written by Corneille) dated 1642-43, while Molière's first plays were supposed to have been written at the earliest in 1656 and were played in Paris only since 1660. So, we can say that, as Corneille is the undisputed author of the two *Menteurs*, he probably also wrote the plays found on the dendrogram on the left of these two plays and which are all very near to one another:

Tartuffe, le *Misanthrope*, les *Femmes savantes*, l'*Etourdi*, le *Dépit amoureux*, l'*Ecole des maris*, l'*Ecole des femmes, Sganarelle, Amphytrion*, la *Princesse d'Elide, Mélicerte* and les *Fâcheux*. That is to say all Molière's verse plays.

On the other hand, the relationship of the Menteur with Molière's prose plays is not so clear; although it has been noted that these Molière comedies are nearer than *Psyché* is to Corneille's work. Therefore, the authorship of Corneille is possible but not so clear as for the alexandrine works. These overall findings suggest a thorough examination of the existing neighbourhood particularly around the two *Menteurs*. Besides verse plays, *Dom Juan* and the *Avare* appear under the .25 level which make us think they were written by the same author as the *Menteurs*. The *Psyché* part written by Corneille, as well as *Dom Garcie* are suggesting an even higher level (.273) that is to say a probable contribution by Corneille to *Amphytrion and Fourberies de Scapin* and even to the *Malade imaginaire* (but here dog latin and italian interludes increase the distance).

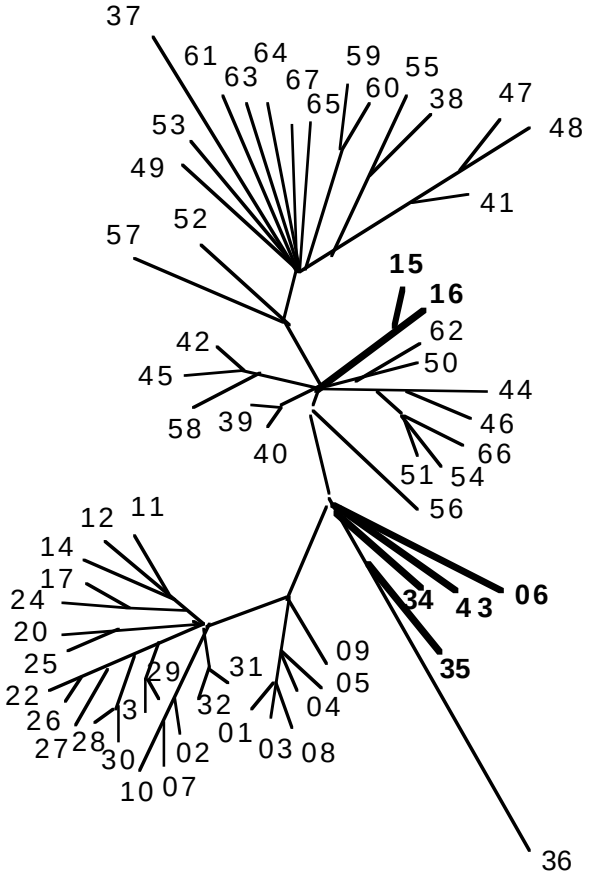
Table V Two Menteurs distances compared to Molière's plays

Plays	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
Molière's entire work mean	*0,275*	*0,266*
Molière's verse plays mean	*0,241*	*0,234*
Corneille's mean	*0,252*	*0,249*

The two *Menteurs* were probably used as a model for a great number of Molière's plays, especially all the verse ones, so that the *Menteurs* are clustered with Molière's and not with Corneille's works, even his comedies. However, except for *l'Etourdi* (1660), all other distances exceed the level of .20 above which authorship becomes unsure. Time lag must also

be considered : Molière's plays were written twenty years and even later the *Menteurs*. To check these findings, another method was used : the tree-analysis. Texts or groups of texts are no longer classified one by one, but, for each one, the best representation of its neighbourhood is compared with all the others (Barthélémy and Guénoche 1988; Luong 1994). Each text forms an end (leaf) linked to others by branches and by trunk sections. The path to link two texts measures their proximity (Table VI).

Table VI. Tree-classification on Molière's and Corneille's plays



We thank M. Xuan Luong (Nice University) for this graph. Each play number is given in the annex. For a better view, we substract .15 from all the distances (as in the dendrogram above). This substraction exagerates neighborhoods between leaves and distances between main nodes).

Allmost all Corneille's plays (numbered from 1 to 33) appear closely gathered on the left, at the bottom of the graph (with two sub-sets : comedies on one hand, tragedies and tragi-comedies on the other). Molière's plays (from 37 to 67) all appear at the top of the graph, clearly divided in two groups : prose writings (the top set) and verse ones (in the middle). On the right, at the bottom, five "anomalies" (represented in bold): besides a short piece of *Psyché* by Quinault (36) the two other *Psyché* by Corneille (34) and Molière (35) ; *Dom Garcie* (43) and (06) the *Comédie des tuileries* fifth act (Corneille's work commissioned by Cardinal Richelieu 37 years before *Psyché).*

Above all, X. Luong's tree-analysis strengthens the main conclusion : Corneille's *Menteurs* (15-16) stand quite in the centre of Molière's works, though they are rather far from the rest of Corneille's plays. In other words, the *Menteurs* authorship is clearly the same as most of Molière's masterpieces. This finding brings out interesting questions. For instance, it can be asked which particularities may be found in Corneille's style and vocabulary that are also present in the plays he wrote for Molière. And conversely, which are their differences.

Molière's historic importance is not at all minimized by it being almost certain that Corneille contributed to most of his masterpieces. He was indeed the first French modern "theater businessman", at the same time : company manager, artistic director, theater director and actor, and, as is shown, an excellent scenario "hunter". Historians and critics will explain how and why the two men worked together and why they concealed it.

This analysis does not put an end to the question of how to establish Molière's authorship. One may go on thinking that Molière summurarised the idea of a contemporaneous life satire (Corneille seemed to have abandoned this idea after the second *Menteur* failure). It can be added that Molière usually directed and played Corneille's works and he could have been " immersed " in Corneille's language and ready to write in the same way as the author he prefered and of whom he knew thousands of verses. More studies will have to be carried out to answer these questions. It could be interesting to look at Molière's main contemporary authors. Some of them could have helped him to write some prose comedies that are too different to be from the pen of Corneille such as : *les Précieuses ridicules*, *l'Impromptu* or *la Critique de l'école des femmes*…

Finally, the easiness of distance calculation programming is to be noticed. Futhermore, the results and their interpretation are within the reach of everyone, without any statistics culture. That is why we hope other studies will strengthen of intertextual distance significance combined with cluster analysis as a tool for literary criticism especially in authorship attribution.

Acknowledgements

We are grateful to Professor Pierre Hubert for suggesting us the Molière and Corneille case, to Jean-Guy Bergeron, Pierre Hubert and Denis Monière for their helpfull comments, and to Mrs Kathleen Milsted for her accurate reading of our first translation.

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 statitistique des données textuelles*, Université de Nice, 1998, p. 49-71.

BERGERON Jean-Guy, LABBE Dominique, 2000, "L'évaluation de la négociation raisonnée par les acteurs: une analyse lexicométrique", XVIe congrès de l'Association Internationale des Sociologues de Langue Française, Québec (à paraître aux Presses de l'Université Laval).

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.

BRUGIDOU Mathieu, LABBE Dominique, 1999, *Le discours syndical français contemporain (CGT, CGT, FO en 1996-98)*, Grenoble-Paris, CERAT-EDF(GRETS).

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.

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.

KYLANDER Britt-Marie, *Le vocabulaire de Molière dans les comédies en alexandrins*, Göteborg, Acta Universitatis Gothoburgensis, 1995.

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 (1967), *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse, (réédition : Genève-Paris, Slatkine-Champion, 1979).

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

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

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

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

Annex Corneille’s and Molière’s Plays.

Corneille		Year of création	Genre	Length (tokens)
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 libres	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