

**University of Dublin  
Trinity College**

**French Department Research Seminar  
(6 April 2004)**

**Corneille in the shadow of Molière**

Dominique Labbé

[dominique.labbe@iep.upmf-grenoble.fr](mailto: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*)

In December 2001, the Journal of Quantitative Linguistics published an article by Cyril Labbé and me (see bibliographical references at the end of this note, before appendixes). This article presented a new method for authorship attribution and gave the example of the main Molière plays which Pierre Corneille probably wrote (lists of these plays in appendix I, II and VI). An essay, intended for the general French public, was published in June 2003 : *Corneille dans l'ombre de Molière* (Corneille in the shadow of Molière). This essay tells the story of this research and tries to answer the main questions. My presentation briefly summaries the method and the conclusions on Corneille-Molière. Of course, from these results have arisen a large number of questions and objections. This presentation discusses the following points: what is the signification of the results? How was this method validated? What do these results prove?

## Intertextual distance and authorship attribution

The aim of our software is to compare two texts by superposing one on the other to calculate differences and similarities in the vocabularies of these two texts. The algorithm involved in this process is the following.

Given two texts A and B.

- $N_a$  and  $N_b$  represent sizes of A and B in tokens (words);
- $F_{ia}$  and  $F_{ib}$  represent the absolute frequency of type ("vocalbe") i in texts A and B.

If their lengths (in tokens) are equal, distance calculation can be directly performed by subtracting the two frequencies of each type and by making a sum of the results.

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

If the two texts are not of same lengths (in tokens), we propose to "reduce" the longer to the size of the shorter one. What is the probability of i occurring in this reshaped text B? This expected value — or "mathematical expectancy" ( $E_{ia}$ ) — can be easily calculated:

$$(1) E_{ia(u)} = F_{ib} * U \text{ with } U = \frac{N_a}{N_b}$$

Considering all the types of A and B, the difference between these theoretical frequencies ( $E_{ia}$ ) and the observed ones ( $F_{ia}$ ) gives the absolute distance between A and B:

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

And the relative distance:

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

The values of relative distance vary evenly between 0 (the two texts have the same types with the same frequencies) and 1 (the two texts share no words) with no jump, nor threshold effect around some values.

Of course, the spelling of all words is carefully checked — and, for French corpora, each token is tagged in order to reduce the effect of the numerous variable endings of words and the high density of homographs in this language.

In formula (3), the denominator is equal to  $2N_a$ : this calculation gives the two texts the same weightings. However, this method, almost always but only slightly favours longer texts, because the extraction is performed on all the vocabulary — the size of which is a function of length — and also as a result of the effects of vocabulary specialisation (see Hubert & Labbé,

1988). In French texts, it appears that this effect can be overlooked under three conditions: 1) the size of the smallest texts is not too small (in any case: more than 1,000 tokens); 2) the scale between the different texts to be compared is not too large (in any case less than 1:10) and, 3) calculations are performed on texts the spellings of which are normalised, and words are lemmatised.

Within these limits, the result falls into a margin of uncertainty equal to less than 5% (Labbé & Labbé, 2003). If these limits are taken into account, intertextual "distance" can be considered as an euclidian metric, in the same way that, for example, the everyday distance between two objects - which is expressed in meters -, or between two cities (in miles).

### **What is intertextual distance measuring exactly ?**

Given that the conditions described above have been respected, intertextual distance measures exactly the more or less similar distance existing between two texts. Four factors determine this similarity.

— **genre**. One does not speak as one writes. Fiction has its rules, theatre has others, etc. The mould imposed by the genre is more or less rigid. Of course, "technical" matters — where the author must adopt an impersonal way of writing and must follow strict rules — are not considered here. It is for this reason that some of our contradictors have argued that the rules of comedy (theatre) — especially for plays in alexandrine verses — was, during the 17th century, so restrictive that it is impossible to know who the author is. But, if *Tartuffe* had been written in the manner of a report for the Science Academy, its success might have been very improbable!

— **vocabulary** of the period. For example, Corneille's work spreads over a period of more than 40 years ; Molière's work over 15 years. It is the length of time which separates the last comedy officially written by Corneille (le Menteur and la Suite du Menteur) and the first play in verses by Molière (l'Etourdi issued in 1658). Over such a period, the style and vocabulary used by an author necessarily changes as does the language itself, especially in the lexical field. Thus, contemporaneous texts tend to be nearer than those written in different epochs.

— the **theme** treated. Each theme has its own specific vocabulary, using places, people, particular nouns and adjectives. For example, Roman tragedy will necessarily deal with "emperor", "senate", "forum", "legions" and many other components that will not be found in Greek mythology.

— and lastly, but not the least important factor: the **author**...

Thus to find the authorship of an anonymous or doubtful text, this text must be compared with some uncontested ones, written during the same epoch and treating similar themes in a same genre (poetry, novel, theatre...) This is an important point: theatre must be compared with theatre, comedies with other comedies, etc...

We applied this calculation to thousands of texts (theatre, novels, plays, poetry, press articles, political speeches, interviews...) These experiments have validated the method and allowed the calibration of a distance scale for French texts.

### *A distance scale*

For texts of which lengths are comprised between 5,000 and 20,000 tokens (words):

- a value less or equal to .20 never occurs when the authors are different;
- between .20 and .25, it is quite sure that the author is the same. If not, the two texts were written during the same period, in the same genre, on similar subject and themes. For example, it sometimes occurs in press articles, about one event, because the journalists have worked with the same sources and have spoken about the same events, people and places. In the case of two different authors of literary works, it is very probable that the second, in chronological order, was "inspired" by the first one;

- over .25, a "grey" area exists: two hypothesis should be considered: a single author and two different themes (and/or genre) or two different contemporaneous authors treating a similar theme in the same genre. So, the more the distance goes beyond this threshold, the more authorship attribution will be difficult, despite the fact that this authorship should not be rejected;

- over .40, authors are certainly different or, for a same author, the genres of the texts are very far from each others. This is for example, the case with spoken and written languages.

Of course,

- these figures are calibrated for French texts the spellings of which are without errors and are strictly standardised. Given the fact that, in any text written in this language, an average of more than one-third of the words are "homographs" (one spelling, several dictionary meanings) and that two-third of the words — often the same ones — have flexible endings, each token is given a tag which indicates its dictionary headword and its part of speech ("lemmatisation"). All calculations are done on these tags;

- the distance calculation must exactly follow the algorithm and respect the length scale presented above;

- these figures must not be considered as thresholds but as milestones along a continuum;

At least, when a great number of texts are analysed, the interpretation of these figures needs the help of some classification algorithms.

## **Classifications**

The "superposition" of the 66 plays by Corneille and Molière — all considered in groups of two— generates a matrix of 4,290 cells. This amount is too large to allow a direct examination and a synthetic view. This obstacle becomes impossible to overcome when several hundred texts are under analysis. Incidentally, it is noteworthy that these experiments were conducted in order to calibrate the tools which are necessary for managing large text bases and that the plays of the French 17th century were some test files among several thousand others...

Some classification methods enable the establishment of the "best possible order" in such a large population of distances. These three words are placed in inverted commas because, despite the calculation capacities of modern computers, none of these techniques is perfect. For the experiments on Corneille and Molière, two of them were used: the well-known "cluster analysis" and the newest one: "tree-classification" (which is almost exclusively used by biologists or geneticists). More precisely, it was the technique of "valued trees" developed by a French mathematician (Xuan Luong, University of Nice). Applied to the Corneille-Molière case, these two methods lead exactly to the same conclusions (which are comprehensively presented in our article of December 2001). Appendix III reproduces Luong's tree.

It is important to remember that a graph is not "proof". Firstly, original data used to draw these graphs must be verifiable (our data and software are published). Secondly, the classification algorithms should be well known and adapted to the data. And thirdly, riggings are possible. In order to avoid this suspicion, distance matrix was sent to X. Luong in July 2000. The titles of the plays were replaced by numbers so that Luong ignored the nature of the data he treated. He returned the chart reproduced in the appendix.

On this chart, plays are the vertex of the tree. Distances between two plays is represented by the length of the path that needs to be gone along in order to link the two vertices. For example, the farthest texts are Psyché prologue (36), written by Quinault, and the first Molière play (37).

Tree classification clearly isolated three major clusters: at the bottom, Corneille's plays; in the middle, Molière's plays in verses and, at the top, Molière's plays in prose which are far more dispersed. But some "anomalies" can be noted, especially:

— in the south-east part of the tree, Psyché (1671) — which is an official collaborative play between Corneille and Molière — joins Dom Garcie de Navarre (1661) — officially by Molière — and the third act of Comédie des Tuileries, written by Corneille for Cardinal Richelieu in 1634, that is to say 37 years before Psyché;

— in the middle of Molière's plays in verses, are to be found the two last comedies of Corneille: le Menteur and la Suite du Menteur (1642-43). This fact is surprising from two points of view. On the one hand, these two plays are not officially by Molière (he was 18 years old). On the other hand they were created 15 years before l'Etourdi (first Molière's play) and 30 years before the Femmes savantes (last play in verses by Molière...)

### **Who wrote Molière's plays ?**

These "anomalies" have highlighted some "strategic" areas of data matrix. In other words, automatic clustering, or tree analysis, are exploratory methods: they provide help in decryption of large text data bases. They facilitate questioning and formulating hypothesis the verification of which should only be done with the help of a detailed analysis of these crucial areas of data matrix, as has been done for Corneille and Molière.

Firstly, the two Menteurs and Molière's plays in verses. Appendix II gives the distances between these two Corneille's plays and all Molière's work. Given the large length of time separating the creation of these plays, distances between them are the smallest that can be observed on a single author work. For example, in Corneille's work, all plays of which the creations are separated by at least 15 years have distances over .20. These little distances are uncommon even for a single author when his work spreads over a large period. Remarkable is the fact that the two Menteurs are nearer the work of Molière than that of Corneille. This proximity is true even if one considers only Corneille's Comedies which were created a few years before these two Menteurs: from Mélite (1629) to Illusion comique (1636).

Consequently, these two Menteurs are the “eldest sisters” of all Molière's plays in verse, and also very probably of Dom Juan and l'Avare. For these two last plays, another factor of differentiation is to be considered: they are in prose and this difference should generate greater distances in regard to plays in verse.

Secondly, the strange position of Dom Garcie and Psyché. Critics usually underline that Dom Garcie stands apart in Molière's work. In fact, appendix IV shows that Dom Garcie and Psyché are twin sisters - even if Psyché is created 10 years after Dom Garcie — and that this sistership includes Corneille's last tragedies which are contemporaneous. Consequently, all these plays were written by the same author and this author is P. Corneille. For Psyché, this fact is certain because of an indiscretion made by the first editor (appendix IX).

Conclusions:

— Dom Garcie and Psyché have the same ancestors which are especially: Andromède (1650) and la Toison d'or (1661);

— other Molière's plays in verses, Dom Juan and l'Avare, are descended from the two Menteurs;

Once again, it must be underlined, that in French literature, such a similar case of mutual links between two works by different authors does not exist. Moreover, such shorts distances are very rare in a single work of an equivalent size: nearly 900,000 words in 49 plays written over a period of more than 40 years.

Please excuse this long demonstration which is the main point in the attribution to Corneille of the major masterpieces of Molière.

### **Trials and signification of results**

Considering the novelty of this method (calculation capacities of modern computers have enabled it only very recently), it is reasonable to consider how it can be tested.

First of all, sceptical people can try to demonstrate that formulae or reasoning are false. Until now, no error has been found. It is the main utility of international reviews: in addition to the editor, two readers have rigorously examined and vetted our text...

Secondly, the method has also been empirically tested.

#### *A crucial experiment*

One can devise a difficult trial like the following one: ask two famous authors to write a novel or a play about the same theme, giving them the same deadline and isolating them to avoid one copying from the other. In this way, it will thus be possible to neutralize three of the four factors presented above (genre, epoch, theme), so that the influence of the fourth (authorship) can be easily measured. Is this impossible to manage? Two famous French dramatists agreed to do so: P. Corneille and J. Racine wrote simultaneously two tragedies —

in alexandrine verses and respecting the same drastic rules — on the same theme: an impossible love between a Roman emperor (Titus) and an oriental queen (Bérénice). It is said that the implicit model was based on the same characters: Louis XIV and his sister in law (Henriette d'Angleterre). Thus, place of action, nouns and names of characters *etc.*, are the same, generating a large common vocabulary. Their distance (.256) is higher than all values between the two Menteurs and all Molière's plays in verse, despite the fact that, for these comedies, themes were always different, dates of creation spread over 30 years and rules were less constraining than for the "great" alexandrine tragedies.

Appendix IV suggests a wider conclusion: the distances between texts pertaining to the same genre, written during the same epoch by a single author, on different themes, are systematically shorter than the distances between two texts written by different authors even if they wrote at the same time, in the same genre and on a single theme... Up until now, nobody has been able to find a counter-example...

Jean Racine and Pierre Corneille have offered us the possibility to conduct the most extraordinary experiment that one can dream about!

*What is the impact of genre ?*

The experiment presented above has been generalised in order to measure precisely how intertextual distance varies when one, two, or three factors are neutralised. For example, it has been said that proximities between the two Menteur and some Molière's plays are explained by the genre "comedy in verse". In order to verify this explanation, Plaideurs by J. Racine can be used: written in 1668, this comedy in alexandrine verses is exactly contemporaneous to that of Molière. By comparing Plaideurs to these plays, we can measure the impact of two factors (author and theme) and neutralise the two others (genre and epoch). If this contention is correct, distances between Plaideurs and Molière plays in verse should be shorter than those between these plays and the Menteurs (because in this last case, time is also an important factor). Appendix II shows that distances over .25 always separate Racine's Plaideurs and all Molière's plays in verses. The shortest distance is with Ecole des femmes (.26), and secondly with Etourdi, Dépit amoureux and Avare (this last one is in prose but it was created the same year than Plaideurs). It is noticeable that J. Racine drew his inspiration from the plays written by Corneille but avoided the sensitive ones like Tartuffe or Dom Juan...

### *A blind test*

Thirdly, to empirically validate a theory, some more sophisticated procedures are possible, such as "blind tests" which are well known in the context of medicine: "placebos" are mixed with the active product, several sectors of the population are also blended together and the tests are conducted anonymously.

In the last days of 2001, E. Brunet, a literature professor (University of Nice) agreed to undertake such an experiment with us. He selected 50 texts drawn from 22 novels by 11 different authors and sent them without any indication of titles or authors. Our algorithm isolated, without any errors, all the excerpts drawn out of a same book and they identified all the "placebos" which E. Brunet had imagined. The only limit — which was foreseen by the model — was on a small number of different books, by a same author, the creations of which are separated by a long period of time. Of course, E. Brunet knew this problem well and, for some authors, he logically chose novels that were separated by many years. But this is not a failure, because it is preferable to give less "sure" answers about authorship attribution (acceptation or rejection) even, if no answer is given sometimes because distances are too high to be accepted and too low to be rejected. Our report for this experiment is on line:

<http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeExperience.pdf>

Of course, should these tests not be considered sufficient, we invite colleagues to send us other texts which, in their opinion, should be able to "fool" our algorithms. These texts will be transparently processed and all the files will be made public. This invitation has been extended for 18 months now without being taken up. Is not this silence proof in itself?

### **What do these results prove?**

Three things are important to bear in mind about Molière.

Firstly, there are no manuscripts by Molière in existence except around 20 signatures on official documents: no dedication, no letter, not a single note in his handwriting.

Secondly, during his life, there is not a single description of him at work, no explanation of his creative methods or details of the sources he used, the books he read...

Finally, the first edition of the "complete works" of Molière was printed 9 years after his death and this compilation gave no indication about the way he conceived and wrote these plays. The first "biography" of Molière was published 32 years after his death. This very short book was founded on the single testimony of Baron, a very young actor who entered the

troupe very late on in its existence. This biography gave no indication about the way Molière created his plays. Boileau - who knew Molière very well - dismissed the whole book as being false.

Consequently, traditional methods of authorship attribution are of very little help in this case.

Let us use the example of scientific police investigations. If fingerprints or DNA found on a crime scene are not those of the suspect, then he is automatically exonerated without any discussion. However, if the fingerprints or DNA found are his, but there is no other evidence against him, it would be difficult to condemn this suspect because a very slight but real doubt remains in such a case, despite the great accuracy of those methods.

So do we have other evidence to strengthen our case? Of course! And this evidence is considerable and corroborates our hypothesis. If not, we would not have published these results, even as a curiosity.

Lexical and stylistic evidence should be distinguished from that taken from the two men's lives.

**As far as texts are concerned**, Pierre Louÿs at the beginning of twentieth century, Henry Poulaille during the 1950' and H. Wouters in the 1990' have pointed to an unusually large amount of similarities between the two works. They are the real discoverers. Our analysis is a simple addition to their demonstration, in the same way, for instance, that "scientific police laboratories" help criminal investigators.

Our conclusions are reinforced by two statistical indices which are perhaps more important than intertextual distance, because these calculations take into account not only words but also their combinations. First, our experiments enlighten the very personal way of combining two verbs, of which the pattern is: "vouloir dire", "savoir faire"... (the first verb is modal and the second is an infinitive). These combinations which are very frequent in French, depend on a very personal view of word and creation. Of course, this experiment needs a large amount of words — here, the question is not the analysis of each play considered alone. On the other hand, it should be admitted that, sometimes, some authors might have changed their mind during their life. Thus similar combinations belong almost surely to a single author but different combinations do not allow to conclude to two different authors, except when the two works are exactly contemporaneous. If these limits are accepted, until now, no two different authors — with the same favourite combinations and similar frequencies — have been paired except... Corneille and Molière. The table in appendix VII also gives the results for Racine.

These usual "modal" verbs are several hundred in number: thus this result cannot occur by chance. It must be noticed that lemmatisation is necessary: if not, the flexible endings of modal verbs or, for example, the function words, placed between modal and infinitive verbs, will prevent these constructions from being found.

Better still, the analysis of semantic networks enable us to understand the specific meaning that each author gives to the words he uses. As a result of this study, it can be observed that the most important words have the same meanings in both the works of Corneille and Molière. Indeed, Molière's meanings appear to be included in those of Corneille. Here again, there is no other similar case in the whole of French literature available on electronic files. For an example, see the case of the word "amour" ("love") which is the most frequently used substantive in 17<sup>th</sup>-century French theatre:

*<http://www.cavi.univ-paris3.fr/lexicometrica/article/numero3.html>*

**As far as history of the period is concerned**, H. Poulaille and H. Wouters have pointed out some strange facts about the lives of the two men. For example, for the period before 1658, there are only 3 Molière comedies in existence. They are in prose, very short and quite bad (*La Jalousie du barbouillé*, *Gorgibus dans le sac*, le *Médecin volant*). In 1658 Molière came and lived in Rouen for six months, where the brothers, Thomas and Pierre Corneille, were living. After this stay in Rouen, a new author was born, who was very different from the first one. In 1662, Corneille and his family moved to Paris and then, masterpieces were created one after the others (see dates for plays in Appendix II). To this suggestive (revealing) chronology, some facts can be added:

— three "collaborative efforts" are evident. Two editors named Corneille as the author of *Dépit amoureux* (1658) and of *Psyché* (1671) (see this documents in appendix VIII and IX). After Molière's death, Thomas Corneille changed *Dom Juan* from prose into verses.

— The author most played by Molière (after himself) was Corneille. However, Corneille gave to Molière the direction of several of his last tragedies. So the statement – often repeated without proof - that the two men quarrelled does not seem to be substantiated.

— Finally, here we have a contemporaneous witness who knew Molière very well:

*Dans ce sac ridicule où Scapin s'enveloppe*

*Je ne reconnais plus l'auteur du Misanthrope*

(Boileau, *Art poétique*, chant III, 1674)

Boileau's dilemma has a plain answer: Molière, who played Scapin, is not the author of the *Misanthrope*...

So a large amount of contradictory evidence exists. It only proves that Corneille wrote these plays. But it does not answer another question: who had the idea for Tartuffe or Dom Juan? Is it possible that Corneille was a kind of "ghost writer" for Molière? Statistics cannot provide an answer to such questions...

### **How was this research received?**

The intertextual distance has been presented since 1998 in various seminars, and, for the first time, during a conference held in Lausanne in March 2000 (with D. Monière) and then, in July 2000 (with J.-G. Bergeron). Several researchers, in different countries are using and testing this method. To date, two articles on W. Shakespeare have been published (Merriam, 2002 and 2003a) and a third one on modern authors (Merriam, 2003b). Others are forthcoming.

Several objections had been made. We have already answered two of the most important ones. Firstly, the impossibility of distinguishing contemporaneous theatre authors writing in a same genre. Secondly, the supposed "quarrel" between Corneille and Molière about Ecole des femmes: no serious facts were reported about this "quarrel".

In our article published in the Journal of Quantitative Linguistics, we suggested a third objection. Corneille was obviously the favourite author of Molière and he certainly knew thousands of verses by Corneille. Thus he should have been similarly "immersed" in Corneille's style and vocabulary... But it would be curious that such an influence acted in an irregular way: considerable in the Fâcheux (1661) and almost slight in the Précieuses ridicules (1661), two plays which are supposed to have been written at the same time; and again considerable in the Ecole des femmes (1662) and undetectable in the Critique de l'Ecole des femmes (1663), and so on. It seems that there are no other cases in literary history of such an irregular influence spread out over a 15 year period.

Two other objections have been raised.

Firstly, it has been said that our calculation is not taking prosody into account. In fact, prosody belongs to genre and its influence is measured by our method. However, it is noticeable that there is often a "prejudice" among literary critics who think that prosody, and rime, are characteristics of an author. Of course, at a given date, they can be slightly different in each work, but above all, they are techniques that are used or "forgotten", very consciously, according to the needs and to fashion. This is verified with the help of a software programme

called "métromètre", which gives a precise analysis of each verse. V. Beaudouin used this software on Corneille and Racine. The results are clear: for contemporaneous plays, in the same genre, there are hardly any differences in prosody between authors. Of course, this tool is interesting for other purposes in literary studies.

Secondly, 6 verses in Ecole des femmes and a short text by Abbé d'Aubignac actually raise a problem (appendix X). P. Louÿs and H. Poulaille thought that these 6 verses were not written by Corneille — as a matter of fact, they are not as well done as the rest of the play. It is an idea that could be looked at in greater detail. For example, our calculation suggests that the Bourgeois gentilhomme and the Malade imaginaire may well be "collaborative" plays as is the case of Psyché (here again, most of it was written by Corneille). Why should it not be the case in other plays?

Intertextual distance, combined with cluster analysis and tree analysis, provides a reliable tool for the classification of large textual data bases and... for literary history and criticism.

### **Acknowledgements**

I am especially grateful to:

my son Cyril who helped me to conceive the method, to write the software, to conduct the first experiments and to write the first two articles;

Edward Arnold (University of Dublin) for his accurate reading of this text;

Jean-Guy Bergeron (Université de Montréal), Pierre Hubert, Jean and Nelly Leselbaum (Université de Paris-X), and Denis Monière (Université de Montréal) who undertook the initial experiments with us;

Charles Bernet (Ecole normale supérieure de Lyon) who provided the electronic files of all the plays by Corneille, Molière and Racine;

Pierre Hubert (Ecole des Mines de Paris) who indicated to us works of P. Louÿs and H. Wouters;

Reinhart Köhler (Editor of Journal of Quantitative Linguistics) who published our first article about Corneille and Molière;

Xuan Luong (University of Nice, F) for charts and tree-classifications;

Tom Merriam (Basingstoke, UK) for his constant help and encouragement.

## Bibliography

- Beaudouin Valérie (2002). Mètre et rythmes du vers classique : Corneille et Racine. Paris. Champion.
- Bergeron Jean-Guy, Labbé 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 in Bernier Colette et Al (eds). Formation, relations professionnelles à l'heure de la société-monde. Paris-Québec. L'Harmattan-Les Presses de l'Université Laval. 2002. p 239-252.
- Hubert Pierre, Labbé Dominique (1988). Un modèle de partition du vocabulaire". in Dominique Labbé, Philippe Thoiron, Daniel Serant, Études sur la richesse et la structure lexicale. Paris-Genève. Slatkine-Champion. p 93-114.
- Labbé Cyril and 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 and 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<sup>e</sup> 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.
- Love Harold (2002), Attributing Authorship: An Introduction, Cambridge, Cambridge University Press.
- Luong Xuan (1988). Méthodes d'analyse arborée. Algorithmes, applications. Thèse pour le doctorat ès sciences. Université de Paris V.
- Merriam Thomas (2002). "Intertextual Distances between Shakespeare Plays, with Special Reference to Henry V (verse)". Journal of Quantitative Linguistics. 9-3. December 2002. p 260-273.
- Merriam Thomas (2003a). "An Application of Authorship Attribution by Intertextual Distance in English". Corpus. 2. 2003. p 167-182.
- Merriam Thomas (2003b). "Intertextual Distances, Three Authors". Literary and Linguistic Computing. 18-4. November 2003.
- 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.

## Appendix I.

Corneille's plays.

(All plays by Corneille are in verse)

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

Psyché	Year of creation	Genre	Size in tokens
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

This corpus comprises 34 plays, a total of 553,190 tokens and 6,258 different types

## Appendix II

Distances between the two *Menteurs* (Corneille) and the *Plaideurs* (Racine)  
and all plays officially attributed to Molière

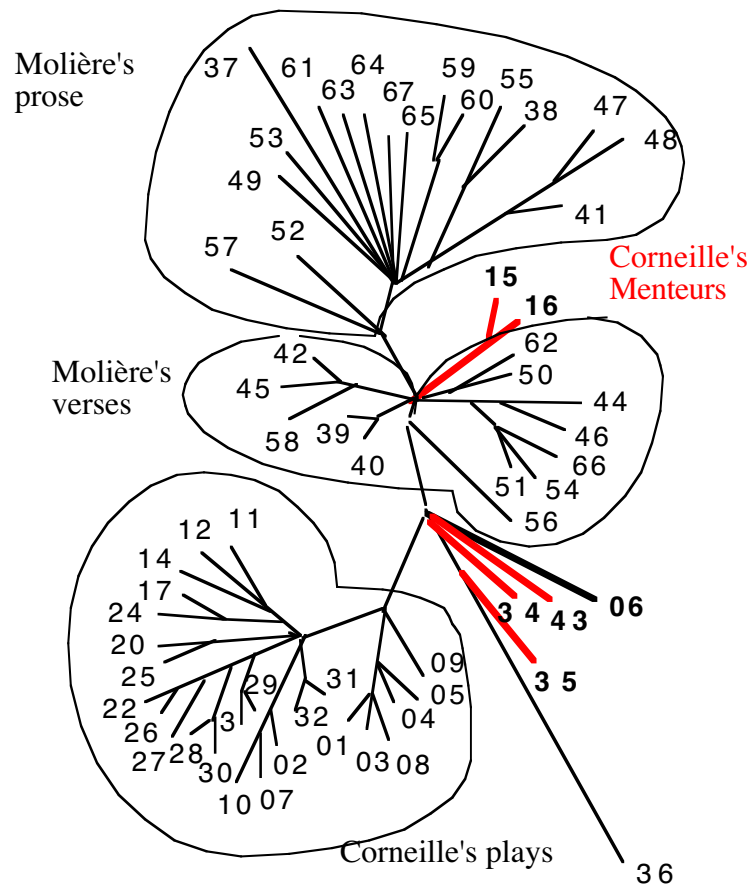
N°	Pièces	Genre	Le Menteur (Corneille 1642)	Suite du Menteur (Corneille 1643)	Les plaideurs (Racine : 1668)
15	Le Menteur (1642)	Vers	0,000	0,180	0,296
16	La suite du Menteur (1643)	Vers	0,180	0,000	0,293
34	Psyché Corneille (1671)	Vers	0,288	0,273	0,348
36	Psyché Molière (1671)	Vers	0,329	0,325	<b>0,354</b>
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	<b>L'étourdi (1658)</b>	Vers	<b>0,205</b>	<b>0,206</b>	<b>0,269</b>
40	<b>Dépit amoureux (1658)</b>	Vers	<b>0,215</b>	<b>0,212</b>	<b>0,270</b>
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	<b>L'école des maris (1661)</b>	Vers	<b>0,223</b>	<b>0,217</b>	0,279
45	<b>Les fâcheux (1661)</b>	Vers	<b>0,248</b>	<b>0,248</b>	0,306
46	<b>L'école des femmes (1662)</b>	Vers	<b>0,226</b>	<b>0,217</b>	<b>0,261</b>
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	<b>Princesse d'Elide (1664)</b>	Vers Prose	0,251	<b>0,243</b>	0,314
51	<b>Le Tartuffe (1664)</b>	Vers	<b>0,242</b>	<b>0,228</b>	0,275
52	<b>Dom Juan (1665)</b>	Prose	0,259	<b>0,248</b>	0,281
53	L'amour médecin (1665)	Prose	0,292	0,289	0,287
54	<b>Le Misanthrope (1666)</b>	Vers	0,252	<b>0,234</b>	0,283
55	Médecin malgré lui (1666)	Prose	0,298	0,289	0,296
56	<b>Mélicerte (1666)</b>	Vers	0,257	<b>0,250</b>	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	<b>L'Avare (1668)</b>	Prose	0,256	<b>0,244</b>	<b>0,270</b>
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	<b>Femmes savantes (1672)</b>	Vers	0,260	<b>0,248</b>	0,283
67	Malade imaginaire (1672)	Prose	0,282	0,270	0,278
<i>Distance mean with Molière's work</i>			<i>0,275</i>	<i>0,266</i>	<i>0,299</i>
<i>Mean with Molière's plays in verses</i>			<i>0,241</i>	<i>0,234</i>	<i>0,290</i>
<i>Distance mean with Corneille's work</i>			<i>0,252</i>	<i>0,249</i>	<i>0,347</i>
<i>Distance mean with Racine's work</i>			<i>0,314</i>	<i>0,311</i>	<i>0,376</i>

Corpus Molière : 34 plays, 394,963 tokens and 8,088 types.

Corpus Racine : 12 plays, 166,626 tokens and 4,323 types.

### Appendix III.

Tree classification of entire works of Molière and Corneille  
(Journal of Quantitative Linguistics, VIII, 3, p 227)



This chart was drawn by M. X. Luong (University of Nice). We sent him the data without details of authors and titles (see appendix I and II)

Plain lines :

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 : parts of Psyché by Corneille

N° 35 : parts of Psyché by Molière

N° 36 : prologue of Psyché by Quinault

N° 43 : Dom Garcie by Molière

#### Appendix IV.

Distances between Dom Garcie (Molière), Psyché (Corneille & Molière) and all the last plays of Corneille.

Last plays of Corneille	<u>Dom Garcie</u> (Molière,1661)	<u>Psyché</u> (Corneille, 1671)
Rodogune (1644)	0,245	0,231
Theodore (1645)	0,234	0,245
Heraclius (1647)	0,248	0,273
Andromède (1650)	0,241	<b>0,218</b>
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)	<b>0,221</b>	<b>0,220</b>
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)	<b>0,227</b>	0,235
Psyché (1671)	<b>0,230</b>	—
Pulcherie (1672)	0,230	0,226
Surena (1674)	<b>0,216</b>	0,224
<i>Mean Corneille</i>	<i>0,243</i>	<i>0,244</i>
<i>Mean Molière</i>	<i>0,286</i>	<i>0,297</i>

#### Appendix V

Main characteristic distances between Corneille and Racine at the epoch of Tite et Bérénice.

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

**Appendix VI**  
Who wrote Molière's plays

16 plays are attributed to Corneille  
(chronological order)

Titles	Acts	Genre	Date	Size (tokens)
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 Molière's plays were not written by Corneille  
(chronological order)

Title	Acts	Genre	Date	Size (Tokens)
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 Molière's plays are not attributed  
(they are too far from *Menteurs* or *Psyché*)**

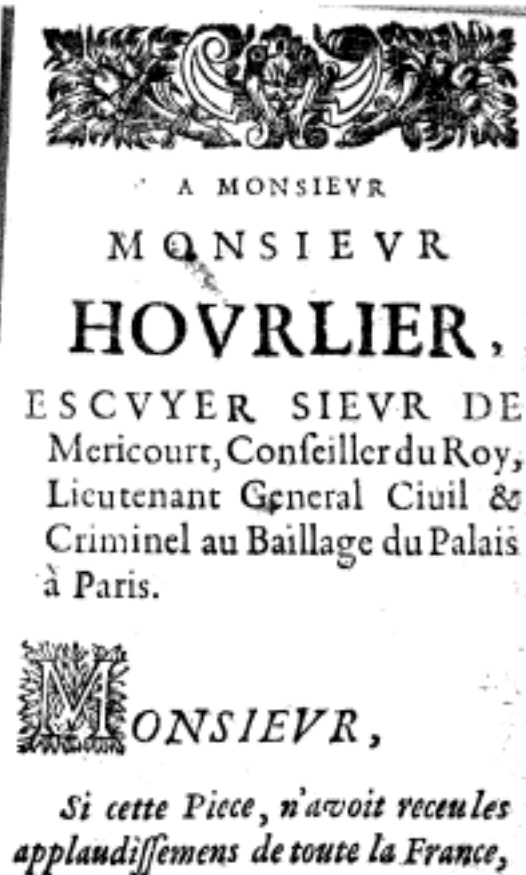
Titles	Acts	Genre	Date	Size (tokens)
Le sicilien ou l'amour peintre	1	Prose	1667	5 375
Georges Dandin	3	Prose	1668	11 009
Monsieur de Pourceaugnac	2	Prose	1669	11 803
Les amants magnifiques	5	Prose	1670	11 983
Le bourgeois gentilhomme	5	Prose	1670	17 136
Les fourberies de Scapin	3	Prose	1671	14 245
Le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

NB : The *Bourgeois gentilhomme* and the *Malade imaginaire*, even if they are farther from the *Menteurs*, can be considered as belonging to the first class — the plays written by Corneille — because when some scenes are withdrawn from these plays (for example, the passages in strange latin in the *Malade*), their distances with the two *Menteurs* fall under .25.

Appendix VII. Combinations of verbs "modal + infinitive" (frequency for 100,000 mots)

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

Racine shares three combinations with "pouvoir" (to be able to see, to do, to be) with the two others. Molière and Corneille share 5/10 and the first three in the same order with very close densities. Given the very large number of possible combinations, this situation cannot occur by chance...



*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 j'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, j'avois toujours differé le iuste & respectueux hommage que ie m'étois proposé de vous rendre; & j'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 :

Because of the cover, I wrote that the first edition of this book was published in 1663. One observed that the printing was done in the autumn of 1662. What is printed on the cover of this book is therefore not reliable!!!

Who was "l'auteur le plus approuvé de ce siècle" (most famous author of this century)? One of the professors at the Sorbonne (G. Forestier) indicates, in a document placed on his website in July 2003 that, by the 1660's, Corneille "dominait de la tête et des épaules le théâtre français depuis vingt ans" (Corneille had been the most prominent author of theatre for the previous 20 years).

## Appendix 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.

## Appendix 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.