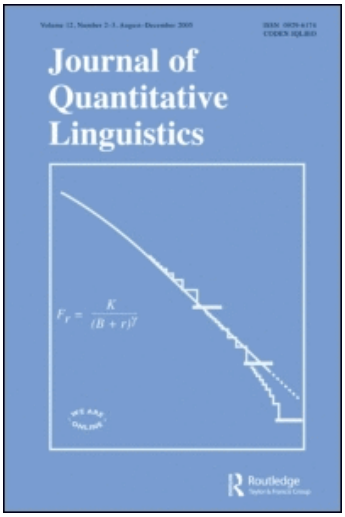

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100702

Mathematical Methods for Attributing Literary Works when Solving the "Corneille-Molière" Problem*

Mikhail Marusenko [a];Elena Rodionova [a]
[a] Saint-Petersburg State University, Russia

Online publication date: 03 February 2010

To cite this Article Marusenko, Mikhail andRodionova, Elena(2010) 'Mathematical Methods for Attributing Literary Works when Solving the "Corneille-Molière" Problem*', Journal of Quantitative Linguistics, 17: 1, 30 — 54
To link to this Article: DOI: 10.1080/09296170903395924
URL: http://dx.doi.org/10.1080/09296170903395924

Journal of Quantitative Linguistics
2010, Volume 17, Number 1, pp. 30–54
DOI: 10.1080/09296170903395924

Routledge
Taylor & Francis Group

Mathematical Methods for Attributing Literary Works when Solving the "Corneille–Molière" Problem*

Mikhail Marusenko and Elena Rodionova
Saint-Petersburg State University, Russia

ABSTRACT

This research work focuses on developing a linguistic approach in the field of attribution of literary works using the material of plays written in verse for which Molière is thought to be the author. In this research work, a solution of the problem "Corneille–Molière" is suggested by using methods of mathematical modelling and quantitative description of individual author styles on a syntactic level.

INTRODUCTION

The question of the authorship of comedies written under the pseudonym Molière was raised at the beginning of the 20th century and is widely discussed today both in France and in other countries. According to the existing hypothesis, the possible real authors of plays which have been attributed to Molière include such well-known French playwrights as Corneille and Quinault. The most heated debates take place over the question of the possibility that Corneille wrote the best plays in verse attributed to Molière. The hypothesis that Corneille wrote works that have been attributed to Molière has traditionally been called the "Corneille–Molière" problem in the works of various researchers. One of the most recent research works done on this topic was that of D. Labbé, whose work is based on an analysis of the lexicons used by the two writers (Labbé, 2003). That said, an analysis of the writers' lexicons alone is not an adequately reliable basis for attributing the authorship of

*Address correspondence to: Elena Rodionova, Udarnikov 43/2 kv. 143, Saint-Petersburg 195030, Russia.

0929-6174/10/17010030

texts since it is the lexicon of a language in particular, connected to narration, which can most easily be imitated. Other researchers of Molière's works in various years have used other literary and biographical arguments which supported one or another point of view. Thus a situation arose in which a linguistic question could not be solved using methods based only on philological analysis, or with the use of methods of a quantitative description of the work's lexicon.

The goal of this article is to use a mathematical method of attribution of anonymous or pseudonymous works when solving the "Corneille–Molière" problem.

The theoretical base for this paper comes from the following assumption:

An individual author's style is a complex hierarchal system and should be described by a set of style-differentiating parameters using a method of multidimensional classifications. The problem of attributing anonymous and pseudonymous texts is one of the tasks of pattern recognition (Marusenko, 1990).

The following texts were used as the *materials* for this research:

(1) Thirteen comedies in verse ascribed to Molière (*L'Étourdi*, *Le Dépit amoureux*, *Sganarelle*, *Dom Garcie de Navarre*, *L'École des maris*, *Les Fâcheux*, *L'École des femmes*, *La Princesse d'Élide*, *Tartuffe*, *Le Misanthrope*, *Mélicerte*, *Pastorale comique*, and *Les Femmes savantes*);
(2) Eleven comedies in verse written by Corneille (*Mélite*, *La Veuve*, *La Galerie du Palais*, *La Suivante*, *La Place royale*, *Comédie des Tuileries*, *L'Illusion comique*, *Le Menteur*, *La Suite de Menteur*, *Don Sanche d'Arago*, and *Tite et Bérénice*);
(3) Three comedies in verse written by Quinault (*Les Rivales*, *L'Amant indiscret*, *La Mère Coquette*).

THE HISTORY OF THE "CORNEILLE–MOLIÈRE" QUESTION

The "Corneille–Molière" problem was first examined in 1919 by the famous French poet Louÿs, who was an expert on 17th-century poetry (Louÿs, 1919). Having made an in-depth analysis of the works of the great French playwright Corneille (1606–1684), who wrote 34 plays in verse, Louÿs reached the unexpected conclusion that there was a large

similarity between Corneille's verse and that of Poclain (1622–1673), another famous playwright who worked under the pseudonym of Molière. Molière's theatrical opus consists of 33 works written in both verse and in prose. Judging by the stylistic characteristics of the works of the two playwrights, Louÿs suggested that Corneille was the author of such Molière masterpieces as *Le Misanthrope*, *L'École des femmes*, *Tartuffe*, and *Dom Juan*. Louÿs stated the assumption that one of the reasons why Corneille could have refused authorship of his own works in favour of the actor Molière was his desire to speak of his love, his life, and his feelings in his comedies, but to do so anonymously, without revealing his real name.

The main theses stated in Louÿs' articles were further developed and justified in 1957 in the book of Poulaille, a novelist, entitled *Corneille under the Mask of Molière* (Poulaille, 1957). Poulaille wrote an alternative biography of Molière, starting from the latter's early childhood. The researcher compares the facts from the biographies of Molière and Corneille, and notes the times when the two playwrights met, thereby proving the possibility and likelihood of their creative union. According to Poulaille, one of the main reasons forcing Corneille, one of the greatest poets of his time, to use the services of Molière, was Corneille's desire to settle scores with his many enemies, using them as characters in topical satirical plays.

This research work did not receive much attention in its day either from simple admirers or from professional researchers of Molière and Corneille. Only a handful of researchers of French literature of the 17th century had heard of Louÿs' hypothesis until 1990, when the problem of the authorship of Molière's theatre was brought up once again by two lawyers from Brussels, Wouters and de Ville de Goyet.

In their work Wouters and de Ville de Goyet gave a detailed analysis of the ideological–stylistic characteristics of such plays as *Les Précieuses ridicules*, *Les Fâcheux*, *L'École des femmes*, *Tartuffe*, *Dom Juan*, and *Amphitryon*, and, also taking into account the histories of the writing of these plays, prove the impossibility that these plays were written by Molière (Wouters & de Ville de Goyet, 1990). Writing the large numbers of plays signed by Molière would have required working non-stop from morning until night, but there is no way that Molière had so much time, since he worked constantly as a theatre director, as a director–producer, and as an actor. Researchers have also pointed to Corneille's difficult financial situation, which was connected to his dependence on cash

payments from the king. Corneille's financial dependence on the court, in the two researchers' opinion, forced the playwright to make fun of his enemies under a pseudonym. Besides Corneille, the two researchers also name the playwright Quinault as one of the possible authors who worked under the name of Molière, since Quinault's style is close to the comedies and ballets attributed to Molière.

The appendix to Wouters and de Ville de Goyet's book features a work by the great ethnologist Vernaud. Having made a lexicological and stylistic study of the plays, Vernaud found a strong resemblance between the texts of Corneille with those written by Molière in their words, versification and style. For example, in Molière's plays Vernaud found Norman words and a large number of terms and expressions connected to jurisprudence. Only Corneille, a resident of Rouen and a famous lawyer, could use these terms. Furthermore, these plays use quotes from little-known works in Latin and make reference to religious literature, which is absent from Molière's library. Corneille, on the other hand, being a very religious man, would certainly have studied these works. Vernaud found coincidences between the texts of the plays written by Molière and Corneille, such as references to Aristotle and Horatio, the mention of Armenia or Armenians, the use of the rare name Nicandre, and others.

Modern literary and theatre critics have had a very critical view on these research works, and have made their own counter-arguments. Forestier, the head of the department of 18th-century theatre studies at the Sorbonne, made the most in-depth critical review of the main assumptions made by supporters of the idea that there was cooperation between Molière and Corneille. In Forestier's opinion, Molière's theatre is rather uniform in its content despite the fact that there are many genre subtypes: one can find the influence of farces in any large comedy written by Molière. Also, Forestier explains Molière's ability to write a large number of plays in a short period of time by the fact that many of Molière's works are short divertissements and are written in prose.

However, besides ideological–stylistic and biographical arguments, the supporters of attribution of Molière's plays to Corneille also now have the first mathematically-based indication of such cooperation.

In 2001 D. Labbé, a professor at the Institute of Political Research in Grenoble, and a specialist in speech analysis, offered a mathematical method for attribution which calls for calculating ''intertextual distance''. In 2003 D. Labbé started using a new method for researching the lexical

content of the theatre of Molière and Corneille. Calculating the ''intertextual distance'' allowed Labbé to attribute about 18 comedies signed by Molière to Corneille (Labbé, 2003).

D. Labbé's research received a great deal of attention in France and abroad, and created a large interest on the part of specialists of various scientific fields, including both mathematicians and literary critics. Furthermore, D. Labbé's work served as an impulse for new research on the problem of authorship of Molière's theatrical pieces.

The hypothesis of secret cooperation between Molière and Corneille was proven by the results of research by Vidal, who made an analysis on the basis of biographical data of the two playwrights (Vidal, 2001).

Then Boissier wrote what is the most in-depth and structured study on the Molière case. Following previous research of the works of Molière and Corneille, Boissier suggests that the plays that make up Molière's *oeuvre* came from three different sources: part of the plays were put together by actors who read French, Spanish, and Italian comedies; another part was bought from various poets who needed money, or from their widows; and lastly, part of the plays were ordered from Corneille (Boissier, 2004).

The idea of Corneille being the author of Molière's plays was accepted by researchers of Corneille: in 2006 the Association of French Researchers of the Work of Corneille launched an official website on the ''Corneille–Molière'' case in honour of the 400th anniversary of Corneille's birth (www.corneille-moliere.org).

Heated discussion over the ''Corneille–Molière'' problem continues both in France and outside the country to the present day.

Objects of Attribution

The majority of research on Molière's *oeuvre* takes all of Molière's literary works into account, including his plays both in poetry and prose, when analysing the question of attribution. In the case of using mathematical methods for attributing literary works, it is necessary to make a clear differentiation between various groups of attributed objects. In the research the comedies in poetry were selected as the attributed objects (see Table 1).

Forming an Attribution Hypothesis

The above-given history of the ''Corneille–Molière'' problem testifies to the fact that there are several hypotheses at present on the real author of

Table 1. Attributed objects.

No.	Work	First performance	N	N_Ω
1	*L'Étourdi, ou Le Contre-Temps*	1658	1084	10530
2	*Le Dépit amoureux*	1658	1210	
3	*Sganarelle, ou Le Cocu imaginaire*	1660	473	
4	*Dom Garcie de Navarre, ou Le Prince jaloux*	1661	853	
5	*L'École des maris*	1661	765	
6	*Les Fâcheux*	1661	546	
7	*L'École des femmes*	1662	1266	
8	*La Princesse d'Élide (parts in verses)*	1664	224	
9	*Tartuffe, ou L'imposteur*	1664	1293	
10	*Le Misanthrope*	1666	1112	
11	*Mélicerte*	1666	421	
12	*Pastorale comique*	1667	72	
13	*Les Femmes savantes*	1672	1211	

the plays ascribed to Molière. The attribution hypothesis may be formulated in the following way:

Null hypothesis (H_0): The texts of plays in verse ascribed to Molière belong completely to Molière.

Alternative hypotheses go as follows:

(H_a^1): The texts of plays in verse ascribed to Molière belong completely to Corneille;
(H_a^2): The texts of plays in verse ascribed to Molière are works of Corneille, Quinault, and of one or several other unknown authors. In this case one must determine the number of authors and the potential share of participation of each of these authors.

METHODS OF ATTRIBUTION

The classification of attribution principles written by the academician Vinogradov divides all attribution methods into groups based on their subjectivity and objectivity. The use of mathematical methods of linguistic analysis is the most fruitful objective method of attribution.

The plethora of mathematical methods of analysis of objects makes it possible to successfully use these mathematical methods for attribution.

Historical–documental and philological methods of research dominated in the field of attribution for a long time. Subjective methods of attribution were used in order to find the particularities of each author. In accordance with this method, the external details of an author's individual style, such as the author's favourite words, terms, or expressions, were subjectively selected.

Mathematical–statistical methods of attribution were first used in the beginning of the 19th century to attribute authorship of works written in ancient times. The research work done by Campbell (1867) and Lutoslawski (1897) was based on establishing the rarest (''original'') words used by a given author, as well as establishing the position of determining and determined words. Linguistic–mathematical methods gradually came to be used even more widely when describing an author's individual style. As a result, data was gathered about the properties of language units and a special scientific apparatus for attribution of texts was formed. The use of specially designed indices for evaluating the lexical structure of a text resulted from researchers' striving to develop a new universal apparatus for objective lexical analysis. Many researchers developed new methods for evaluating the lexical content of texts. These new methods each had their own faults, while none of them took into account already-existing solutions in joint fields of knowledge, for example in mathematics, where the theory of recognition of patterns for describing objects of a various nature had already been successfully used for a long time.

One of the most recent research works on attribution of texts based on lexical analysis was done by D. Labbé, who suggested a formula in 2001 for calculating the ''intertextual distance''. The research done by D. Labbé reveals the serious methodological and statistic-probability shortcomings in previously-used methods. These shortcomings are related first of all to the lack of authenticity of results of attribution made on the basis of an analysis on just one lexical level. When making a stylistic analysis with the aim of attributing texts, a study of the lexical contents of a text should be supplemented with additional data on other language levels as well, and foremost on the syntactic structure of the text being analysed.

The period from the end of the 1970s to the present day has been characterized by strong development of computer calculation technology

and software, as a result of which more and more researchers have become interested in using computer technologies for data processing when analysing texts in their syntactical, grammatical, morphemic, and lexical aspects. Researchers' efforts to employ automatic stylistic diagnostics of texts and an automated search of individual characteristics of an author's style have led to preference being given in stylistic analyses to any other language level besides the syntactic level. The dependence of stylistic analysis on computerized data processing and on methods which employ computer equipment leads to a simplification of the methodological basis of research. This, in the end, makes methods of attribution less effective.

The history of the development of methods of attribution has led to an understanding that an effective method for stylistic analysis with the aim of determining authorship should have the following characteristics:

(1) Characteristics of the whole text, and not of individual sentences, should be determined with the help of stylistic analysis.
(2) Description of the text should cover various levels of the language system, and the structure of the text should be analysed as well as the text's vocabulary.
(3) It is necessary to use multidimensional classifications.

Furthermore, studying the links between parameters has shown that simply increasing the number of parameters does not lead to an increase in the effectiveness of analysis. Thus it is necessary to have a special mathematical apparatus for estimating the links between parameters. Those informational parameters should be chosen which exclude excess parameters which have a strong correlation between each other.

All of these requirements are fulfilled by such a ready-to-use mathematical apparatus as the theory of pattern recognition. Methods of pattern recognition were first used when attributing anonymous and pseudonymous works on the basis of an author's individual style characteristics in the work of Marusenko in 1990 (Marusenko, 1990). Since then this method of attribution was successfully applied for attribution of a number of literary works of doubtful authorship. Among them are *Quiet Flows the Don* attributed to Sholohov, the works attributed to Romain Gary, and others (Marusenko et al., 2001; Chepiga 2007).

In the present research work a text is viewed as a complex linguistic object which features a wide inventory of elements that can be analyzed on many levels. Judging by the requirements of an adequate description of a

text, a multidimensional statistical analysis – pattern recognition theory – was used as the base of the new method of attribution of anonymous and pseudonymous literary works. If we use pattern recognition then style is considered to be "a set of properties, characterizing the content, the ways of connecting, and the statistical-probability regularity of the use of language means which form the given individual writer's style" (Marusenko, 1990, p. 24). The set of properties which characterize the structure of a text in its syntactical aspect becomes in this case the sum total of informative parameters. The informative parameters are meant to distinguish the woks by different authors and the make-up of their sum total is determined by executing a special procedure for selecting informative parameters for each concrete case.

An important theoretical position of the given research work is that the procedure of attribution is divided into three relatively independent stages:

(1) Formation of a literary-critical attribution hypothesis which is executed using methods of traditional philological analysis employing all accessible methods for attribution.
(2) Rejecting/not rejecting the literary-critical hypothesis. Means of the theory of pattern recognition are used to test the hypothesis.
(3) Interpretation of the results of testing the attribution hypothesis.

The hypothesis is considered to be statistically corroborated if the results of recognition coincide with the original literary-critical attribution hypothesis (under the established level of meaning). In the opposite case the hypothesis is considered to be disproved, and either an alternative hypothesis is made, or the original hypothesis is modified. When fulfilling such a scheme of attribution, statistical-probability methods of analysis of language and style are used only as supplemental means for testing the original attribution hypothesis with the help of philological methods of attribution. Testing the literary-critical hypothesis takes place in several stages using a certain set of procedures:

(1) Determining the *a priori* set of individual stylistic parameters. Considering that parameters from the *a priori* dictionary of parameters should be determined by style in its structural–syntactical aspect (Martynenko, 1988); these parameters are taken from works of those researchers who studied sentence structure and

make-up using mathematical methods (among those researches are Sevbo, 1981; Fucks, 1975; Khetso et al., 1989; Vašák, 1980).

(2) Determining the *a priori* set of classes. The make-up of *a priori* classes is determined by the requirements of uniformity of time and genre, while volume is measured in the main syntactical units – sentences.

(3) Description of classes from the *a priori* alphabet of classes in the language of parameters from the *a priori* dictionary of parameters. Each linguistic object which is subject to analysis with the aim of making stylistic diagnostics is presented in accordance with the mathematical object p, characterized by an n-dimensional vector, where n is the number of parameters.

During the stage of describing the attributed objects in the language of parameters from the *a priori* dictionary of parameters, the researcher must process data by hand. This makes it possible to adequately describe the text in its syntactical aspect. Of course it is very important to exclude the accidental mistakes – that is why the processing by hand entails forming general rules for analysing texts, introducing rules for parameterization of the text for each parameter and, finally, making a calculation of linguistic phenomena according to the rules.

(4) Determining the information set of parameters. This stage of attribution consists of separating a necessary and adequate number of parameters for linking the object to a class from the informational parametric space. Excess parameters are eliminated as a result. When forming the set of informative parameters Bongard's scheme is used (Bongard, 1967). The Bongard's scheme is a set of successive stages and the research work can be replicated by other researches. The detailed description of this stage is given below by the example of the problem "Corneille–Molière".

(5) Choice of a deciding rule. The task of determining the author of an anonymous or pseudonymous text in this research paper is seen as a task of finding the distance between the multidimensional vector which suits the *a priori* class M_1, and the multidimensional vector which suits the *a priori* class M_2 of the unknown author. The deciding rule is the function chosen to measure this distance and to take a decision on the sameness or discrimination of sameness of these objects. The recognition algorithm used should provide for a separation of distance of signs into fields which

correspond to classes with a minimum of recognition mistakes. In the given research work the algorithm of recognition calls for a two-stepped recognition procedure: determining and probability.

(6) Appraisal of the quality of qualification. Since the classes received as a result of the mathematical procedure of classification can be artefacts, it is necessary to make an appraisal of the quality of classification. This appraisal may lead to correction of the structure of the classes received.

Use of the mathematical apparatus developed by Marusenko on real historical–literary material showed the apparatus' high effectiveness (Marusenko, 2001). One can make a conclusion about the stability of the recognition system to fluctuations in the volume of texts and to a temporary evolution of parameters of the author's style based on the results of tests of real attribution hypotheses described in several research works. In the majority of cases the recognition system provides a complete separation of objects into according classes. In the opposite case, a hypothesis can be stated after appraising the quality of classification that another one or several classes of authors also exist who were not accounted for in the original attribution hypothesis. That said, sequential use of determining and probability algorithms of recognition excludes impossibility of recognition.

ATTRIBUTION OF PLAYS IN VERSE ATTRIBUTED TO MOLIÈRE

Testing of the attribution hypothesis calls for selection of informative parameters from the *a priori* dictionary of parameters and making a recognition procedure.

A Priori Dictionary of Parameters

The inventory of parameters which form the original description was made up of 51 parameters which are relevant for describing 17th-century texts in French (see Table 2). The language of parameters is a system for determining and fixating a series of the most significant structural particularities of organization of sentences.

The values of parameters received as a result of describing *a priori* classes are signs which characterize the according classes. This research

Table 2. *A priori* dictionary of parameters.

Code	Parameter
1	2
X1	The number of words in a simple independent sentence
X2	The number of simple sentences
X3	The number of main clauses
X4	The number of complex sentences
X5	The number of complex sentences without a conjugated form of the verb
X6	The number of subordinate clauses
X7	The number of 1st degree subordinate clauses
X8	The number of 2nd degree subordinate clauses
X9	The number of 3rd degree subordinate clauses
X10	The number of 4th degree and higher subordinate clauses
X11	The number of simple sentences without a nominative subject
X12	The number of subordinate clause without a conjugated form of the verb
X13	The number of parenthetic phrases
X14	The number of cross-sectional sentences
X15	The number of words of the 1st group (lexical)
X16	The number of words of the 2nd group (auxiliary)
X17	The number of nouns
X18	The number of adjectives
X19	The number of pronouns
X20	The number of numerals
X21	The number of conjugated verb forms
X22	The number of name forms of verbs
X23	The number of adverbs
X24	The number of prepositions
X25	The number of conjunctions
X26	The number of subordinating conjunctions
X27	The number of coordinate conjunctions
X28	The number of predicatives
X29	The number of direct objects
X30	The number of indirect objects
X31	The number of subjects
X32	The number of pronouns – subjects
X33	The number of groups of uniform parts
X34	The number of parts of uniform groups
X35	The number of uniform predicates
X36	The number of similar groups of objects
X37	The number of participle turns of speech
X38	The number of parts of participle turns
X39	The number of distributed participle modifiers
X40	The number of members of distributed participle modifiers
X41	The number of matched modifiers

(*continued*)

Table 2. (*Continued*).

Code	Parameter
X42	The number of participles – matched modifiers
X43	The number of non-matched modifiers
X44	The number of nouns – non-matched modifiers
X45	The number of isolated parts
X46	The number of parts in groups of isolated parts
X47	The number of nouns without a group
X48	The number of groups of nouns
X49	The number of parts of groups of nouns
X50	The number of lexical words in groups of nouns
X51	The number of helping words in groups of nouns

work forms general rules for analyzing texts in French, and introduces rules for parameterization of the text for each parameter. The full list of general rules is accessible on the Internet page (http://www.corneille-moliere.com/), and a fragment of one of them is shown below for example:

X_{45} – the number of isolated parts.

The isolation is a notional and inflexion accentuation of parts of a sentence, getting in some way the syntactical independence in a phrase. In the research the isolation is determined on the basis of punctuation marks: a comma or commas, brackets, a dash.

In the French language there are the so-called "affiliated elements" (Wilmet, 1998), which do not have any grammatical connection with the rest of a sentence. Among these are amplifying parts, addresses, parenthetical clauses and parenthetical clauses of direct speech pointing at the author of a remark. In this research work the amplifying parts and the addresses are considered to be isolated parts of a sentence, while the parenthetical clauses which amplify different sides of situation are considered to be coordinate clauses. Then the sentences with parenthetical clauses of direct speech are not included in the text corpora.

The examples:

En un lieu, l'autre jour, où je faisois visite,
Je trouvai quelques gens d'un très-rare mérite,

Qui, parlant des vrais soins d'une âme qui vit bien,
Firent tomber sur vous, Madame, l'entretien. (Misanthrope) [$X_{45} = 2$].
Et moi, j'ai de la faim et de l'inquiétude
De voir qu'un sot amour fait toute votre étude. (Sganarelle) [$X_{45} = 1$].

A Priori Classes

The make-up of the *a priori* alphabet of classes determines the mathematical models of *a priori* classes relative to which attribution of debatable works is done. In this research work two *a priori* classes have been formed: Corneille – Ω_1(Corneille) and Quinault – Ω_2(Quinault), with a power of 11 and 3 texts, accordingly (see Tables 3 and 4).

The Informative Set of Parameters

An experiment was performed at describing the *a priori* classes in the language of parameters from the *a priori* dictionary of parameters in order to determine the informative set of parameters.

Two random selections with a volume of 100 sentences each were made from two *a priori* classes. A volume of 100 sentences is sufficient for estimative selections which are made to determine the order of dispersion value. In mathematical statistics it is recommended to make no less than 30 measurements (Doerffel, 1990). In the research, the real volumes of the necessary selection is determined using formulas taking into account a standard deviation and a volume of general totality of texts (Formula (2)).

Table 3. Contents and structure of the *a priori* class Ω_1 (Corneille).

No.	Work	Year written	N	N_Ω
1	*Mélite*	1629	1036	11103
2	*La Veuve*	1631	1169	
3	*La Galerie du Palais*	1632	1156	
4	*La Suivante*	1633	974	
5	*La Place royale*	1634	950	
6	*Comédie des Tuileries*	1634	209	
7	*L'Illusion comique*	1636	1035	
8	*Le Menteur*	1642	1268	
9	*La Suite de Menteur*	1643	1337	
10	*Don Sanche d'Arago*	1650	966	
11	*Tite et Bérénice*	1670	1003	

Table 4. Contents and structure of the *a priori* class Ω_2 (Quinault).

Code	Work	Year written	N	N_Ω
1	*Les Rivales*	1653	866	3125
2	*L'Amant indiscret ou le Maistre étourdi*	1654	1112	
3	*La Mère Coquette, ou les Amants Brouillez*	1665	1147	

The results of the experiment were presented in two object-sign data matrices with a dimensionality $n \times N = 100 \times 51$, where n is the number of parameters, and N is the number of objects.

When forming the set of informative parameters, Bongard's scheme was used, which calls for two-step reduction of the parametric space (Bongard, 1967). In the first stage the *a priori* set of informative parameters is broken down into two subsets of parameters, the relevant and irrelevant ones for distinguishing *a priori* classes. The relevance of parameters for distinguishing the two *a priori* classes is determined using the Student t-criterion (Formula (1)),

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{1/2}}, \tag{1}$$

where $\bar{x}_1$, $\bar{x}_2$ are average arithmetical values, σ_1, σ_2 are standard deviations, and n_1, n_2 are the sizes of the selections.

In the research, σ_1, σ_2 are not known, but only estimated, and the threshold value at $\alpha = 0.05$ of the Student t-criterion is approximately 1.96 (see Table 5).

The values of the t-criteria for five parameters turned out to be higher than the critical level, which made it possible to determine the parameters X02 (the number of simple sentences), X04 (the number of complex sentences), X21 (the number of conjugated forms of the verb), X31 (the number of subjects), and X32 (the number of pronouns–subjects) as informative.

The second stage of Bongard's scheme calls for a procedure of reducing the parametric space into a subset of informative parameters.

The calculation results showed that for all five parameters $E_j < 1$, from which one can conclude that on the second stage of selection of informative parameters there was no further reduction in the number of informative parameters, and that the informative set of parameters included the five parameters received during the first stage.

Table 5. Calculation of the Student *t*-criteria.

	A priori classes						
	Ω_1 (Corneille)			Ω_2 (Quinault)			
Parameter	$\bar{x}_i$	σ_i	n	$\bar{x}_i$	σ_i	n	t
1	2	3	4	5	6	7	8
X1	3.28	4.92	100	2.84	4.19	100	0.68
X2	1.80	0.89	100	2.17	1.41	100	**2.22**
X3	0.44	0.56	100	0.38	0.65	100	0.70
X4	0.53	0.94	100	1.22	1.43	100	**4.04**
X5	0.04	0.20	100	0.02	0.14	100	0.83
X6	0.52	0.75	100	0.44	0.78	100	0.74
X7	0.50	0.70	100	0.44	0.78	100	0.57
X8	0.02	0.14	100	0.00	0.00	100	1.42
X9	0.00	0.00	100	0.00	0.00	100	–
X10	0.00	0.00	100	0.00	0.00	100	–
X11	1.41	0.91	100	1.65	1.10	100	1.68
X12	0.04	0.20	100	0.02	0.14	100	0.83
X13	0.06	0.28	100	0.08	0.34	100	0.46
X14	0.05	0.22	100	0.07	0.29	100	0.56
X15	9.61	5.28	100	10.50	7.29	100	0.99
X16	3.58	2.94	100	3.46	3.14	100	0.28
X17	2.38	1.95	100	2.31	2.00	100	0.25
X18	1.36	1.44	100	1.38	1.50	100	0.10
X19	2.28	1.89	100	2.42	1.99	100	0.51
X20	0.04	0.20	100	0.06	0.28	100	0.59
X21	1.76	1.01	100	2.14	1.52	100	**2.09**
X22	0.77	1.06	100	0.97	1.15	100	1.28
X23	0.93	0.98	100	1.04	1.13	100	0.74
X24	1.49	1.45	100	1.17	1.30	100	1.64
X25	0.96	1.16	100	0.79	1.01	100	1.10
X26	0.42	0.79	100	0.32	0.63	100	0.98
X27	0.52	0.69	100	0.47	0.67	100	0.52
X28	0.15	0.41	100	0.28	0.65	100	1.69
X29	1.27	1.06	100	1.13	1.30	100	0.83
X30	1.40	1.25	100	1.11	1.35	100	1.58
X31	1.47	0.98	100	1.91	1.46	100	**2.50**
X32	1.08	0.95	100	1.39	1.05	100	**2.19**
X33	0.37	0.85	100	0.34	0.87	100	0.25
X34	1.12	2.85	100	1.08	2.91	100	0.10
X35	0.14	0.51	100	0.17	0.59	100	0.38
X36	0.15	0.56	100	0.07	0.41	100	1.16
X37	0.10	0.39	100	0.07	0.26	100	0.64

(*continued*)

Table 5. (*Continued*).

	A priori classes						
	Ω_1 (Corneille)			Ω_2 (Quinault)			
Parameter	$\bar{x}_i$	σ_i	n	$\bar{x}_i$	σ_i	n	t
1	2	3	4	5	6	7	8
X38	0.43	1.85	100	0.36	1.43	100	0.30
X39	0.02	0.14	100	0.00	0.00	100	1.42
X40	0.02	0.20	100	0.00	0.00	100	1.00
X41	1.37	1.45	100	1.31	1.43	100	0.29
X42	0.00	0.00	100	0.02	0.14	100	1.42
X43	0.42	0.78	100	0.45	0.73	100	0.28
X44	0.20	0.55	100	0.23	0.55	100	0.39
X45	0.34	0.57	100	0.35	0.58	100	0.12
X46	1.00	2.14	100	0.75	1.47	100	0.96
X47	1.08	1.06	100	0.93	1.08	100	0.99
X48	1.21	1.34	100	1.26	1.33	100	0.26
X49	3.44	3.77	100	3.64	4.42	100	0.34
X50	2.83	2.98	100	2.93	3.20	100	0.23
X51	0.61	1.21	100	0.71	1.57	100	0.50

Mathematical Models

The stage of the transition from the real object to its mathematical model is based on describing the object in the informative parameters of the recognition system. The next stage of our research was to form a data matrix of attributed objects and *a priori* classes. The extent of the attributed objects and *a priori* classes was determined beforehand (Formula (2)).

$$n = \frac{N}{1 + \left(\frac{V_{\bar{x}}}{V}\right)^2 N}, \quad \text{where } V = \frac{s}{\overline{x}},\ V_{\bar{x}} = \frac{V}{\sqrt{n}}\sqrt{1-f}. \tag{2}$$

The calculation of $n(V_{\bar{x}} = 0.05)$ for one of the plays (*L'Étourdi*) is shown in Table 6, while the calculations of $n(V_{\bar{x}} = 0.05)$ for all attributed objects and their coordinates are accessible on the Internet page (http://www.corneille-moliere.com/).

The co-ordinates of standards of classes are shown in Table 7.

"The algorithm of recognition" includes determinate and probabilistic attribution.

Table 6. Calculation of $n(V_{\bar{x}} = 0.05)$ for an attributed object 1M (*L'Étourdi*).

Parameter	$\bar{X}_i$	σ_i	$n(V_{\bar{x}} = 0.05)$
X02	2.06	1.52	183
X04	0.79	1.20	**498**
X21	1.95	1.67	231
X31	1.81	1.67	261
X32	1.43	1.45	300
$N = 1084$			$n = 100$

"The determinate algorithm of recognition" determines the belonging of attributed objects to one or another class of texts. The fact that we use the sufficiently high sample size allows us to use the Student *t*-criterion as a classifying function of the determinate algorithm of recognition (Formula (1)).

Positive decisions to attribute an object to one or another class were taken in the case that the observed values of the *t*-criteria were less than the critical level in all five measurements of classification space. Classification of objects with the help of the determinate algorithm of recognition is an iterative procedure, in which after each iteration there is a change in the make-up and the capacity of classes, as well as a re-calculation of the co-ordinates of classes.

The values for standards of a *j* class with a power of *i* objects after each iteration are calculated by the following formulas:

$$\bar{x}_j = \frac{(\bar{x}_1 n_1 + \bar{x}_2 n_2 + \cdots + \bar{x}_i n_i)}{\sum n_i}, \tag{3}$$

$$\sigma_j^2 = \frac{\sum \sigma_i^2 n_i}{\sum n_i}. \tag{4}$$

The determinate algorithm was stopped after the 3rd iteration. The co-ordinates of classes after each of the three iterations are available on the Internet page (http://www.corneille-moliere.com/) and the decisions of the determinate algorithm of recognition are shown in Table 8.

The authorship of six of the thirteen objects analysed was determined as a result of the work of the determinate algorithm: *Le Dépit amoureux*, *L'École des maris*, *Les Fâcheux*, *L'École des femmes*, *Tartuffe*, and *Les Femmes savantes* were attributed to Corneille with a probability within

Table 7. Co-ordinates of the standards of classes at the 0th iteration.

	Class			
	Ω_1^0 (Corneille)		Ω_2^0 (Quinault)	
Parameters	$\bar{X}_i$	σ_i	$\bar{X}_i$	σ_i
X02	1.91	1.08	2.02	1.38
X04	0.56	0.98	0.92	1.31
X21	1.90	1.25	2.01	1.49
X31	1.69	1.13	1.83	1.44
X32	1.25	1.01	1.40	1.16
	$N = 1122$		$N = 465$	

Table 8. Decisions of the determinate algorithm of recognition.

Object		Class	
Code	Work	Ω_1 (Corneille)	Ω_2 (Quinault)
1M	*L'Étourdi*	−	−
2M	*Le Dépit amoureux*	+	−
3M	*Sganarelle*	−	−
4M	*Dom Garcie de Navarre*	−	−
5M	*L'École des maris*	+	−
6M	*Les Fâcheux*	+	−
7M	*L'École des femmes*	+	−
8M	*La Princesse d'Élide*	−	−
9M	*Tartuffe*	+	−
10M	*Le Misanthrope*	−	−
11M	*Mélicerte*	−	−
12M	*Pastorale comique*	−	−
13M	*Les Femmes savantes*	+	−

an interval between 0.95 and $0.95 + 0.05^5$ while the real value of probability is close to the right border of the interval.

The Probability Algorithm of Recognition

After stoppage of the determinate algorithm there were seven objects to be attributed which had not been attributed to a single one of the classes. Implementation of the probability algorithm of recognition calls for transformation of the original data matrix into a matrix of Euclidean distances between *a priori* classes and non-attributed objects.

The original data matrix $Z = \{\bar{X}i\}$ (see Table 9) was formed using average values for $\bar{X}i$ of each of the five diagnosed parameters.

Since the values of the diagnostic parameters used can vary quite widely, it is necessary to standardize the data matrix so as to bring all parameters under a single scale. Under this standard form the average values of all parameters equal zero, while dispersions are equal to one. The standardized matrix (see Table 10) is shown as $X = \{x_{ij}\}$. In order to build this matrix such values as the average value of each parameter are used:

$$\bar{z}_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij}, \quad j = 1, \ldots, n \tag{5}$$

and the dispersion of parameter z_j:

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N} (z_{ij} - \bar{z}_j)^2, \quad j = 1, \ldots, n \tag{6}$$

After that elements from the matrix X are calculated:

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{\sigma_j}, \quad i = 1, N, \; j = 1, \ldots, n. \tag{7}$$

Table 9. Original data matrix $Z = \{\bar{X}i\}$.

	Parameter				
	X2	X4	X21	X31	X32
Object					
L'Étourdi	2.06	0.79	1.95	1.81	1.43
Sganarelle	1.65	0.44	1.57	1.4	1.04
Dom Garcie de Navarre	2.40	0.58	2.36	2.14	1.55
La Princesse d'Élide	2.27	0.63	2.31	2.06	1.37
Le Misanthrope	1.94	0.48	1.87	1.77	1.41
Mélicerte	1.74	0.41	1.64	1.51	1.10
Pastorale comique	1.60	0.39	1.63	1.32	0.90
Class					
Ω_1 (Corneille)	1.91	0.56	1.90	1.69	1.25
Ω_2 (Quinault)	2.02	0.92	2.01	1.83	1.40
$\bar{x}_i$	**1.95**	**0.58**	**1.92**	**1.73**	**1.27**
σ_i	**0.27**	**0.18**	**0.28**	**0.28**	**0.21**

Note: $\bar{x}_i$, average arithmetic values; σ_i, standard deviations.

Table 10. Standardized data matrix $X = \{x_{ij}\}$.

	Parameter				
	X2	X4	X21	X31	X32
Object					
L'Étourdi	0.39	1.18	0.12	0.30	0.73
Sganarelle	−1.13	−0.77	−1.22	−1.17	−1.08
Dom Garcie de Navarre	1.65	0.01	1.57	1.49	1.29
La Princesse d'Élide	1.17	0.29	1.39	1.20	0.45
Le Misanthrope	−0.05	−0.55	−0.16	0.16	0.64
Mélicerte	−0.80	−0.94	−0.97	−0.77	−0.80
Pastorale comique	−1.32	−1.05	−1.01	−1.45	−1.73
Class					
Ω_1 (Corneille)	−0.17	−0.10	−0.05	−0.13	−0.10
Ω_2 (Quinault)	0.24	1.91	0.33	0.37	0.59

Using the function of Euclidean distance (Formula 8), a matrix of Euclidean distances between the *a priori* classes and the non-attributed objects was made on the base of the standardized data matrix received (see Table 11).

$$d(a,b) = \left[\sum_{i=1}^{n} \left(x_{aj} - x_{bj}\right)^2\right]^{0,5}, \tag{8}$$

where n is the dimension of the Euclidean distance and a and b are two points in the space E_n with co-ordinates $a(x_{a1}, x_{a2},\ldots, x_{an})$, $b(x_{b1}, x_{b2},\ldots, x_{bn})$.

The likelihood of the jth object belonging to the ith class (see Table 12) is calculated on the basis of the matrix of Euclidean distances between the objects and the *a priori* classes under the following formula:

$$P_{ji} = \frac{1}{d_{ji}}\left(\sum_{k}\frac{1}{d_{jk}}\right)^{-1}. \tag{9}$$

where d_{ji} is the distance between the jth object and the ith class, and d_{jk} is the distance between the jth object and the other classes of classification.

Table 11. Matrix of Euclidean distances between classes and non-attributed objects $d(X_i, \Omega_i)$.

	Class	
	Ω_1 (Corneille)	Ω_2 (Quinault)
Object		
L'Étourdi	1.70	0.79
Sganarelle	2.18	4.08
Dom Garcie de Navarre	3.24	2.97
La Princesse d'Élide	2.47	2.30
Le Misanthrope	0.93	2.53
Mélicerte	1.69	3.76
Pastorale comique	2.75	4.66

Table 12. Matrix of the likelihood of objects belonging to *a priori* classes $P(X_i \in \Omega_j)$.

	Class	
	Ω_1 (Corneille)	Ω_2 (Quinault)
Object		
L'Étourdi	0.32	0.68
Sganarelle	0.65	0.35
Dom Garcie de Navarre	0.48	0.52
La Princesse d'Élide	0.48	0.52
Le Misanthrope	0.73	0.27
Mélicerte	0.69	0.31
Pastorale comique	0.63	0.37

When there are two *a priori* classes, the deciding rule can be formulated in the following way:

$$X_i \in \Omega_j \,.\, P\left(X_i \in \Omega_j\right) \geq 0.5. \tag{10}$$

Under the given deciding rule the threshold value of probability that the object belongs to one of the *a priori* classes should be equal to or greater than 0.5. The plays *L'Étourdi*, *Dom Garcie de Navarre* and *La Princesse d'Élide* were attributed to Quinault, and the plays *Sganarelle*, *Le Misanthrope*, *Mélicerte*, and *Pastorale comique* were attributed to Corneille as a result of the work of the probability algorithm of recognition.

Appraisal of the Quality of the Classification

The final stage of the procedure of recognition was to appraise the quality of the classification received, based on revealing the principles of each class.

In order to consider the procedure of recognition complete it is necessary, first of all, to check the conditions of conformity of the given separation into classes to the conceptual understanding of a cluster and, secondly, to the condition under which the cluster received can be called thickened. Cluster and accumulation are two qualitative gradations which make it possible to determine how closely objects are located within a class as compared to the location of objects within the whole original aggregate. Classes formed in the recognition process will correspond to the conceptual notion of a cluster in the case that the square of the average distance between objects of the class will be less than the square of the average distance between objects of the original aggregate. In other words $\bar{d}^2(\Omega_n) < \bar{d}^2(\Omega_N)$. Accumulation takes place in this case when the maximum square of distance between objects of the given class is less than the average square of distance between objects of the original aggregate, in other words $d_{\max}{}^2(\Omega_n) < \bar{d}^2(\Omega_N)$. At this stage conclusions made using the results of the probability algorithm of recognition were proven.

The next stage in appraising the quality of the classification was to compare the homogeneity of *a priori* and *a posteriori* classes. The average distance between objects $\bar{d}(\Omega_N)$ was used as a criterion of homogeneity. An appraisal of the homogeneity of the make-up of the *a priori* classes completely confirmed the results received from the determinate and probability algorithm of recognition for the class Ω_1 (Corneille), to which objects were attributed with various levels of probability. As for Ω_2 (Quinault), based on the results of the analysis of the quality of classification, of three objects attributed to this class, only one object, the play *L'Étourdi*, can be attributed to Quinault with a probability of 0.68. The other two objects form a separate *a posteriori* class Ω_3, as objects which do not belong to any one of the *a priori* classes, but which make up one class that fulfils the conditions of a cluster and accumulation.

Thus, the given research proved the hypothesis (H_a^2): the texts of the plays ascribed to Molière are probably works by Corneille, Quinault, and an unknown author.

We show the data received in a matrix of solutions, where 0 is a refusal of recognition, 1 is a positive solution for the determinate algorithm of

Table 13. Matrix of solutions.

Code	Work	Ω_1 (Corneille)	Ω_2 (Quinault)	Ω_3
1M	*L'Étourdi*	0	1*	0
2M	*Le Dépit amoureux*	1	0	0
3M	*Sganarelle*	1*	0	0
4M	*Dom Garcie de Navarre*	0	0	1*
5M	*L'École des maris*	1	0	0
6M	*Les Fâcheux*	1	0	0
7M	*L'École des femmes*	1	0	0
8M	*La Princesse d'Élide*	0	0	1*
9M	*Tartuffe*	1	0	0
10M	*Le Misanthrope*	1*	0	0
11M	*Mélicerte*	1*	0	0
12M	*Pastorale comique*	1*	0	0
13M	*Les Femmes savantes*	1	0	0

attribution and 1* is a positive solution for the probability algorithm of attribution (see Table 13).

CONCLUSIONS

A study of documental–historical facts and data using philological analysis allowed us to form a complex attribution hypothesis which suggests that some of the plays ascribed to Molière were written by Corneille, Quinault, and unknown authors. The attribution done proved the high informational capacity and style-differentiating capabilities of the parameters at a syntactical level. The authorship of six plays out of 13 objects analysed was determined as a result of the work of the determinate algorithm: *Le dépit amoureux*, *L'École des maris*, *Les Fâcheux*, *L'École des femmes*, *Tartuffe*, and *Les Femmes savantes* were attributed to Corneille with a probability of more than 0.95. Quinault was attributed authorship of the play *L'Étourdi* as a result of the work of the probability algorithm of recognition, while Corneille was attributed with authorship of the plays *Sganarelle*, *Le Misanthrope*, *Mélicerte*, and *Pastorale comique* with various degrees of probability (from 0.63 to 0.73). The results of the probability algorithm were corrected in the process of the procedure of appraising the quality of classification. Two plays, *Dom Garcie de Navarre* and *La Princesse d'Élide*, made up the *a posteriori*

class. Overall, the variant of the alternative hypothesis (H_a^2) was corroborated: the texts of plays ascribed to Molière are the works of Corneille, Quinault, and one unknown author.

REFERENCES

Boissier, D. (2004). *L' Affaire Molière. La grande superchérie littéraire*. Paris: Jean-Cyrille Godefroy.

Bongard, M. M. (1967). *Проблемы узнавания*. Moscow: Nauka.

Campbell, L. (1867). *The Sophistes and Politicus of Plato*. Oxford: Clarendon Press.

Chepiga, V. (2007). Les archives de Romain Gary. *Genesis*, *29*, 141–143.

Doerffel, K. (1990). *Statistik in der analytischen Chemie*. Leipzig: Deutscher Verlag.

Forestier, G. (2003). Le dossier Corneille-Molière ou: D'un vrai canular a une fausse découverte scientifique. Le Centre de Recherche sur l'Histoire du Théâtre. Retrieved from http://www.crht.org/ressources/dossiers/corneille-Molière/

Fucks, B. (1975). По всем правилам искусства: точные методы в исследованиях литературы, музыки и изобразительного искусства. Iskusstvo i EVM. M. 134–356.

Khetso, G., Gustavsson, S., & Bekman, B. (1989). Кто написал «Тихий Дон»?:(проблема авторства «Тихого Дона»). Moscow: Kniga.

Labbé, C., & Labbé, D. (2001). Inter-textual distance and authorship attribution Corneille and Molière. *Journal of Quantitative Lingustics*, *8*(3), 213–231.

Labbé, D. (2003). *Corneille dans l'ombre de Molière. Histoire d'une recherche*. Paris and Brussels: Les Impressions Nouvelles.

Louÿs, P. (1919). L'auteur d'Amphitryon. *Le Temps*. October 16.

Lutoslawski, W. (1897). *The Origin and Growth of Plato's Logic with an Account of Plato's Style and of the Chronology of his Writings*. London: Longman.

Martynenko, G. Y. (1988). *Основы стилеметрии*. Leningrad: Leningrad University.

Marusenko, M. A. (1990). *Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов*. Leningrad: Leningrad University.

Marusenko, M. A., Bessonov, B. L., & Bogdanova, L. M. (2001). *В поисках потерянного автора: этюды атрибуции*. Saint Petersburg: Faculty of Philology, Saint Petersburg University.

Poulaille, H. (1957). *Corneille sous le masque de Molière*. Paris: Grasset.

Sevbo, I. P. (1981). *Графическое представление синтаксических структур и стилистическая диагностика*. Kiev: Naukova dumka.

Vašák, P. (1980). Metody urcovani autorstvi. Prague: Academia.

Voltaire (1739). Vie de Molière. Amsterdam: Catuffe.

Vidal, P. (2001). *Molière – Corneille. Les mensonges d'une légende*. Paris: Lafon.

Wouters, H., & de Ville de Goyet, C. (1990). *Molière ou l'auteur imaginaire?* Brussels: Complèxe.