A Tool for Literary Studies: Intertextual Distance and Tree Classification

Cyril Labbé
Grenoble I University, France

Dominique Labbé
Grenoble II University, France

Abstract

Correspondence:
Cyril Labbé, LSR-IMAG, 681,
rue de la Passerelle, BP. 72,
38402 Saint Martin d'Hères
Cedex, France.
E-mail:
cyril.labbe@imag.fr

How to measure proximities and oppositions in large text corpora? Intertextual distance provides a simple and interesting solution. Its properties make it a good tool for text classification, and especially for tree-analysis which is fully presented and discussed here. In order to measure the quality of this classification, two indices are proposed. The method presented provides an accurate tool for literary studies—as is demonstrated by applying it to two areas of French literature, Racine's tragedies and an authorship attribution experiment.

The degree of proximity between texts and authors is an old question asked in statistics with respect to literary computing. Rather than using a set of 'frequent words' or 'function words', more or less arbitrarily selected, it is deemed better to take into consideration the entire vocabulary of the texts which are compared. A method for that purpose has already been presented and discussed (Labbé and Labbé, 2001; Merriam, 2002, 2003a and 2003b). The properties of the resulting intertextual distance make it a good tool for classifications, especially that of 'tree-analysis' which is presented and discussed in this article. Tree-analysis is well known in taxonomy or in genetics (a bibliography and a software survey can be found on: http://evolution.genetics.washington.edu; see also: Barthélémy and Guénoche, 1988). Furthermore, this method already has had a few applications in the field of literary studies (e.g. Juilliard and Luong, 1997 and 2001; Spencer *et al.*, 2003). It provides an accurate tool for literary studies as seen when applied to two French corpora.

1 Calculation of Intertextual Distance

Given two texts *A* and *B*. If their lengths (in tokens) are equal, the distance between them can be directly calculated by subtracting the two frequencies of each type from each other and by adding together the sum of the results for all the types occurring in both texts. This sum is a Euclidian distance which gives a quantitative answer to our opening question (for the properties of Euclidian distances applied to texts, see Labbé and Labbé, 2001; Labbé and Labbé, 2003). It must be noticed that this distance is derived without any error (presuming the texts contain no spelling mistakes).

If the two texts are not of same length (in tokens), the absolute frequencies can be replaced by relative frequencies. But some questions do arise: does this method neutralize effects of size differences among the texts? Are the properties of the Euclidian distances preserved?

To answer these questions, let us consider:

– N_a and N_b represent sizes of A and B in tokens with $N_a < N_b$;
– F_{ia} and F_{ib} represent the absolute frequency of type i in texts A and B.

What is the probability of i occurring in a sample of N_a tokens drawn out of B? This expected value—or 'mathematical expectancy' (E_{ia})—can be easily calculated:

$$E_{ia(u)} = F_{tb} \times U \ \ with \ \ U = \frac{N_a}{N_b} \qquad (1)$$

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 intertextual distance between A and B:

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

and the relative distance:

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

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 neither jump, nor threshold effect around some values. In the following, unless otherwise mentioned, formula (3) is used.

Remarks. Of course, the spelling of all words has to be carefully checked and, for French corpora, each token must be tagged in order to reduce the effect of the numerous variable endings of words and the high density of homographs in French (Labbé, 1990).

The same result could be obtained by placing the B tokens in a vase, shaking them, and drawing from the vase a large number of samples of N_a tokens. It must be noticed that, in this experiment:

– when a token is drawn from the vase, it is not replaced since this is the only way to be sure that the number of a type i, in a sample of N_a tokens extracted from B, will be always less than or equal to its frequency in B ($E_{ia} = F_{ib}$). Thus, this experiment follows a hypergeometric distribution, not a binomial distribution;
– of course, this last method entails a certain margin of error which decreases as the number of samples increases. But this margin of error will be null only when the ($N_b! - 1$) different samples are all considered. Since that is impossible, a safety margin should either be calculated—through hypergeometric variance—or should be estimated (binomial variance).

Consequently, it is preferable to use formula (3) which gives exactly the same result in a more rapid way and without any margin of error.

In order to facilitate interpretation of the results, the distance values are multiplied by 1000. Expressed in this way, the result of formula (3)—when applied to two texts A and B (the sizes of which are N_a and N_b tokens)—can be defined as: the average number of different tokens which should be counted in all the $(N_a! - 1) \times (N_b! - 1)$ pairs of different samples of 1000 tokens that can be extracted out of these texts (without replacement).

In formula (3), the denominator is equal to $2N_a$ (neglecting rounding-offs): this calculation gives the two texts the same weightings. This incidentally poses the question as to whether the usual drawbacks generated by size difference between the texts under consideration are neutralized. This method, almost always slightly favours longer texts, because the extraction is performed on all the vocabulary—the size of which is a function of text length—and also as a result of the effect of vocabulary specialization (see Hubert and Labbé, 1988). The same problem, of course, occurs when using relative frequencies ... In French texts, it appears that this effect can be overlooked in three circumstances: (1) the size of the smallest texts is not too small (in every case: more than 1000 tokens), (2) the scale proportion between the different texts to be compared is not too large (in every case less than 1:10) and, (3) calculations are performed on texts in which the spellings are normalized, and the words are lemmatized.

Within these limits, the result has a margin of error equal to less than 5% (Labbé and Labbé, 2003). If these limits are taken into account,

intertextual 'distance' can be considered an Euclidian metric, in the same way that, for example, the everyday distance between two objects can be expressed in meters, or between two cities can be in miles.

Given 2 texts A and B, the intertextual distance between them satisfies the following properties:

– $D(A,B) = 0$ if and only if $A = B$ (identity);
– $D(A,B) = D(B,A)$ (symmetry);
– $D(A,B) \leq D(A,C) + D(B,C)$ (triangle inequality).

The word 'distance' is used only when the measure satisfies these conditions. If not, the words 'similarity' and 'dissimilarity' refer to measures which are not strictly Euclidian.

Appendix II displays the results of this calculation when applied to the eleven tragedies of the well-known 17th century French author, Jean Racine (1639–1699: titles and dates of the plays in Appendix I).

In the case of distances between Racine's tragedies, it must be remarked that these works are all in alexandrine verse, and they were written within the strict limits of rigorous rules concerning not only prosody but also the 'unities' of time, space and action, not to mention constraints of the rules of so-called 'decency'. Therefore, it should be expected that the plays would be very similar to each other (as is the actual case with the tragedies of Corneille, Racine's main rival). But in this corpus, distances between plays are relatively great, especially between the couple of plays (*Esther–Athalie*) on the one hand, and the rest on the other. The distance between *Bérénice* and *Esther* (346 different types out of 1000) is far above expected in this kind of corpus (for a single author, plays belonging to the same genre are generally separated by distances between 160 and 250‰). This large diversity is the main characteristic of Racine's work when compared to all other authors of 17th century French theatre. This example shows one of the numerous opportunities provided by a standardized measurement like the intertextual distance: generalized comparisons become possible between a very large number of authors, texts, and corpora.

As it can be seen in Appendix II, even with a small corpus like Racine's tragedies, the matrix is not easy to manipulate, and its information is difficult to analyze and hence, simplification is needed. Firstly, some elements are grouped together in order to reduce the matrix size through a 'good' classification method. Secondly, the result is graphically represented.

A good method of classification should have the following characteristics; initially, it should follow an unsupervised method (human interventions entail errors or bias) and without preliminary training (selection of some texts in order to train the algorithm partly induce final results). In addition, it must accurately represent the maximum of the matrix information with the least deformation (this accuracy must be able to be measured). Within these constraints, the tree-classification method seems the most effectively adapted tool for textual data processing that currently exists in order to automatically classify data and to represent them graphically.

2 Tree-classification of Textual Distances

To understand how a classification tree is built and how to interpret it, an example drawn from Racine's work is first presented. Between his eleven tragedies, the largest distance separates *Bérénice* and *Esther*. *Britannicus* is the closest play to *Bérénice*; *Athalie* is the closest to *Esther*. Thus two opposite or contrasting sets can be created and graphically represented by a 'tree' (Table 1 and Fig. 1).

Figure 1 shows a 'minimal tree' (it is impossible to draw a tree with less than four points—why this is so will be explained below). A, B, C and D are the 'vertices', 'end points' or terminal 'leaves'. A and B

Table 1 Two sets of contrasting plays in Racine's work (measured in different types out of 1000)

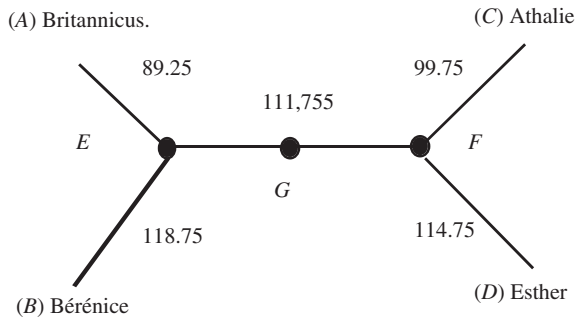|	Britannicus (*A*)	Bérénice (*B*)	Esther (*C*)	Athalie (*D*)
Britannicus (*A*)	0	208	314	302
Bérénice (*B*)	208	0	346	329
Esther (*C*)	314	346	0	214
Athalie (*D*)	302	329	214	0

Fig. 1 Tree representation of Table 1

are adjacent, so as *C* and *D*: they form two sets of 'neighbours' and these two contrasting groups are 'opposite'. Points *E* and *F* materialize these associations and oppositions: they are 'nodes' of the tree. These two nodes are created by the algorithm which establishes their relative positions by calculating lengths of the 'edges'[1] (or 'branches') *AE, BE, CF* and *DF* which link the leaves to the central 'trunk' (*EF*). Each point of the tree is linked to another by a 'path' formed by at least one 'edge'. For example, *AEFC* is the path to follow in order to go from *Britannicus* to *Athalie*: the longer the path, the farther away are the two texts. *G* is the tree 'root'; in Fig. 1 it is placed in the middle of the trunk because points *E* and *F* are simultaneously created by the algorithm. Usually, the tree root is the last node to be created.

It can also be observed that this tree is 'connected': each text (vertex) is linked by at least one edge to all the others. It contains 'no circuit or loop': there is only one path to link one point to another. This graph is 'valued': the path lengths are positive and in fact, proportional to the original values in the corresponding cell of the distance matrix (Table 1). For example, it can be seen in Table 1 and Fig. 1 that the intertextual distance between *Bérénice* and *Britannicus* is equal to the length of the path that links points *A* and *B* ($89.25 + 118.75 = 208$).

In Fig. 1, neighbouring and opposing relations provide 6 different paths: *AEB, AEFC, AEFD, BEFC, BEFD* and *CFD*. It should be noted that there is no specific direction in which these paths should be taken (*AEB = BEA*, etc.)

Among various properties, it can be seen that a minimal tree exists when (and only when) the relation between the four points (*a, b, c, d*) is:

$$D(a, b) \leq \text{Min}\{D(a, c), D(a, d), D(b, c), D(b, d)\} \tag{4}$$

It must be noticed that this relation always exists when the four points are located on the same plane. When they are not, tree is an economical way for representing on a plane the multi-dimensional relations. Sometimes, it entails a 'loss'—or compromized approximations—as seen below:

This 'four point condition' can be applied to larger populations. Given a series of *n* texts (with $n > 4$), *k* (2 or more) out of these *n* texts can be 'grouped' together, provided that all distances between these *k* texts, considered by pairs, are less or equal to all the distances separating these *k* texts from the $n - k$ others in the series (all the possible different pairs must be taken into consideration). Consequently, these *k* texts are grouped together and linked through a same node to the other parts of the tree.

Of course, this condition exists in the particular case of two or more equal edges. Such a situation is easy to imagine: in any large city, one can find suburbs located at the four cardinal points, nearly the same distance from the centre. The corresponding topological tree will present a cross-shape, or a star-shape if the number of points is greater than four (and if they are located on more than two axes). For literary studies this result is not inconclusive. The texts concerned can be said to be more or less 'equidistant': it provides non-trivial information in many cases.

A number of rules must be kept in mind when analysing a tree. The absolute position of leaves and nodes does not matter. For example, Fig. 2 displays some examples from the large number of equivalent trees that can be drawn in order to represent the relationships between these four plays by Racine. *The absolute positions in space are not significant. Only proximities and contrasting oppositions, and lengths of paths linking leaves and nodes of a tree are considered.*

As it can be seen in Table 2, the lengths of these paths (or 'tree-distances') can differ slightly from the original values.
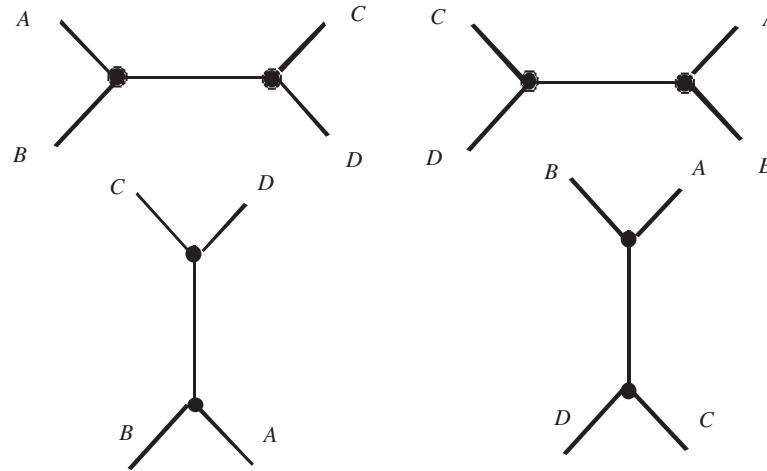
Fig. 2 Some example of equivalent trees

Table 2 Distances between four characteristic tragedies of Racine

| Paths linking couples of texts	A Intertextual distance (%)	B Tree's paths lengths	C |A − B|	1 − C/A
AB Bérénice–Britannicus	208	89.25 + 118.75 = 208	0	1.000
CD Esther–Athalie	214	99.75 + 114.25 = 214	0	1.000
AC Bérénice–Esther	346	118.75 + 111.75 + 14.25 = 344.8	1.2	0.997
AD Bérénice–Athalie	329	118.75 + 111.75 + 99.75 = 330.3	1.3	0.996
BC Britannicus–Esther	314	89.25 + 111.75 + 114.25 = 315.3	1.3	0.996
BD Britannicus–Athalie	302	89.25 + 111.75 + 99.75 = 300.8	1.2	0.996
Sums	1713		5	0.997

If all distances were measured without any adjustment error, the lengths of all these paths should be exactly equal to the corresponding intertextual distances. In reality, they are not. The algorithm used follows Luong's 'grouping method' (Luong, 1988): 'terminal branches' (*AB, CD*) are initially calculated, and final fitting is made on the length of the central trunk. When this technique is applied to a large number of very similar texts, it can entail problems that are discussed at the conclusion of this article.

The last column of Table 2 shows that the lengths measured on the tree are very near to those of the original data. When texts are of different lengths, the intertextual distance margin of error is less than 5% above and below the observed values. In this example, these measurement errors explain the slight differences between original measures and lengths in the tree. It can also be observed that, the greater the dissimilarity between the two texts, the larger the potential error. Among all Racines' tragedies, the two most divergent pairs are: *Bérénice–Esther* and *Bérénice–Athalie*. It can be seen in Table 2 that the mean of deviations between the original data and their tree-representations is less than 0.3%, considerably less than the error margin of calculation. It will be shown next how this information is used to calculate tree quality.

3 Tree-classification on Racine's Tragedies

The internal diversity of Racine's opus appears clearly with tree classification (Fig. 3): the terminal
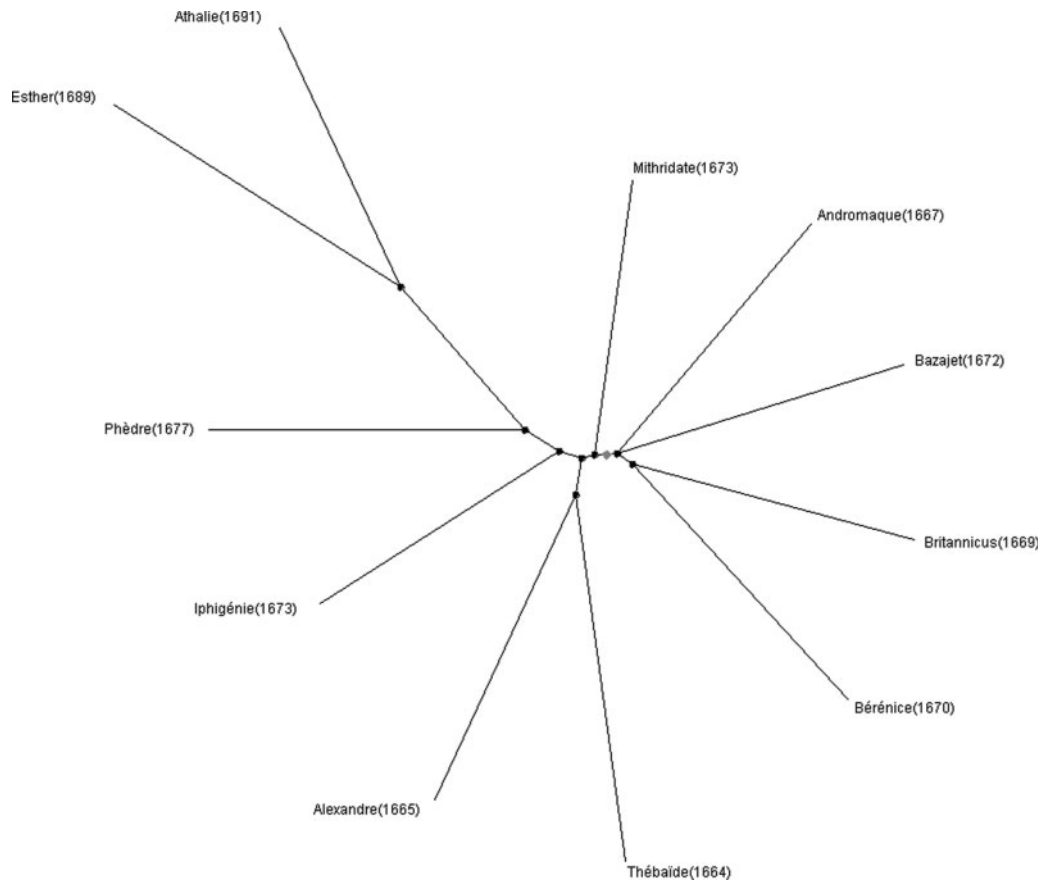
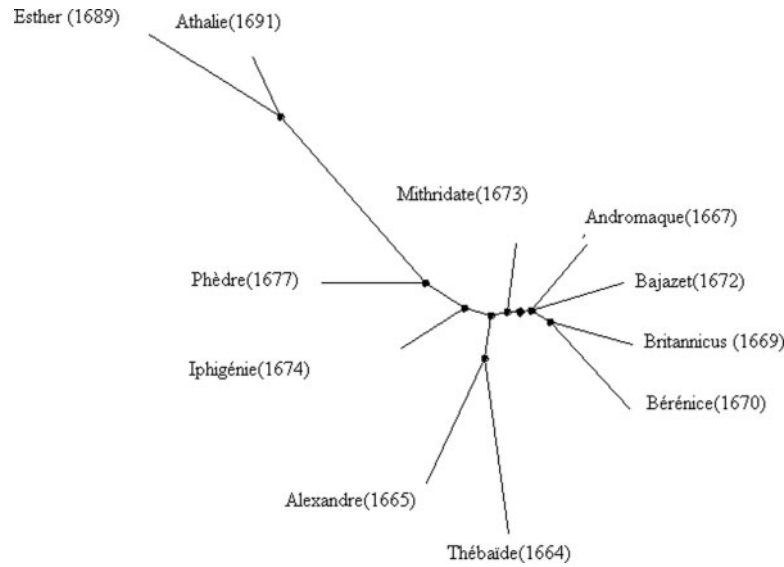**Fig. 3** Tree classification of Racine's tragedies

edges are very long and the central trunk seems to be comparatively short (except beyond Iphigénie to the northwest).

This tree has been drawn, step-by-step in an ascending procedure (Luong 1988, Ruhlman, 2003). Appendix III gives all the steps, following the grouping order. First, the algorithm performs a search on the two or more nearest texts which are opposed in contrast to all the others, in this case, *Esther* and *Athalie*. Since the pairs chosen satisfy the 'four point conditions', they are grouped together and linked to the rest of the plays through a node (n°12). The distance matrix (Appendix II) is 'compressed': the two lines and rows (for *Athalie* and *Esther*), are replaced by just one, corresponding to the node n°12; the original distances are replaced

by the mean distances between this node and each other text. Then the algorithm groups two or more other texts or nodes, once again compresses the matrix, and so on until less than four texts or nodes remain to be grouped. Appendix III shows the various steps and the paths opened on the tree by each node.

Many corpora, characterized by the same genre, author and chronology, present a characteristic 'cross shape' like Fig. 3. This phenomenon can be explained by the fact that, for between seven out of eleven texts, distances are very much alike, with values between 195 and 250‰. In such cases, a simple procedure exists to highlight the marginal fluctuation of a standard chart. The *Y*-axis is cut off just below the lowest value, thus forming the

Fig. 4 Tree classification of Racine's tragedies (Original distances reduced by 150%)

new origin of this axis. Similarly, Luong in unpublished private correspondence has offered a solution. He proposed to subtract from all distances in the original matrix a value chosen below the minimum one (in this case, 193‰: *Iphigénie–Mithridate*). Luong explained that 'this reduction will not alter the tree-topology but it will reduce only all terminal branches and proportionally increase the length of the trunk between the main nodes', thus providing a kind of 'zoom effect' in this crucial zone, and softening the cross shape of the tree. Our experiments have verified Luong's intuition, but it appears that this 'tree reduction' entails a real loss of quality of the tree. Figure 4 presents the optimum trade-off between the loss and gain.

Tree classification separates a few major subsets and clearly isolates the last two plays *Athalie* (1691) and *Esther* (1689). This classification seems actually to confirm the conclusions of traditional literary criticism, especially with regard to the main component that appears to be one of chronological order. As a matter of fact, it has been considered that the evolution of Racine from *Thébaïde* (1664) to *Phèdre* (1677) was quite linear: Racine seemed

to have slowly moved away from Corneille's influence. In fact, tree-classification suggests that Racine's evolution was more complex.

At the start, two 'youth plays': *Thébaïde* and *Alexandre* (1665) are clearly grouped together and isolated in the lower part of the graph. They share the same themes taken from ancient Greece, and their style is relatively austere, very close to that of Corneille's last plays which are contemporaneous with *Thébaïde* and *Alexandre*.

At the opposite end of the spectrum, we find the two 'sacred tragedies': *Esther* and *Athalie*. It is interesting to note that twelve years elapsed between *Phèdre* (1677) and *Esther* (1689), and that Racine is thought to have seriously changed his mind about theatre and religion during this 12-year interval. These two 'sacred tragedies' were ordered by Mme de Maintenon, the second wife of Louis XIVth, from Racine for the moral edification of the occupants of her orphanage. Racine did not want these plays to be publicly performed, and critics consider that they stand apart from his whole work. Figures 3 and 4 clearly confirm this, but they also show that, from the point of view of vocabulary, these plays are the continuation of an

evolution begun after *Mithridate* (1673), many years before.

Four plays, from *Andromaque* (1667) to *Bajazet* (1672), form the right-hand cluster in figs 3 and 4. Two of them are 'Roman'–*Britannicus* (1669) and *Bérénice* (1670)—two others are 'oriental': *Andromaque* (1667) *Bazajet* (1672). The themes remain 'Classical' but they are less complicated and the plot is also simplified. Characters are more 'complete', more sharply delineated. Correspondingly, the richness of vocabulary and stylistic diversity also increase.

Mithridate (1673) seems to be a kind of 'regression' back in the direction of the first two plays. This play seems to mark a turning-point in Racine's work: he hesitates between a return to his former style and themes, and a move in a new direction. He finally chooses the latter. The last plays spread along the northwest branch of the tree. First, two secular ('profanes') plays, *Iphigénie* (1673) and *Phèdre* (1677), in which Racine seems to rediscover Greek mythology, but in a new manner and with a larger vocabulary and stylistic richness than previously, and by emphasizing destiny and the will of the Gods.

In other words, *Iphigénie* and *Phèdre* open a new way forward. This is perhaps the most interesting conclusion of the experiment. Despite the apparent renunciation of Racine and his rallying to the 'moral order' which dominates the last part of the 17th century in France, the last four plays can be seen as being linked together like sisters.

4 Quality Measures

Several questions arise out of these operations. What degree of confidence can be given to the classification and to the chart? Do classification trees always give credible and reliable information with respect to all their component branches? (Spencer *et al.*, 2003).

The last column of Table 2 suggests a way to measure the quality of the whole procedure—from the very beginning (text preparation, measurement of intertextual distances) to the determining of the location of the tree root, which terminates

the procedure. Given a node X linking two texts A and B, the quality of this grouping is measured by the absolute difference between intertextual distance ($D_{(a,b)}$) and the length of the path linking A and B on the tree ($P_{(a,b)}$). In order to normalize this index, the absolute difference is divided by the original value ($D_{(a,b)}$).

$$Quality\ index\,(X_{(a,b)}) = 1 - \frac{|D_{(a,b)} - P_{(a,b)}|}{D_{(a,b)}} \qquad (5)$$

This index varies between 1 (no difference between distance and length of path) to 0 (distance is positive and length of path on the tree is non-existent). The index can be interpreted in two ways or both: as an estimation of errors in the measurement of original data and/or as the proportion of information lost in creating the tree classification. In this second interpretation, the index measures the cost to be paid in order to obtain a readable graph with the help of a somewhat imperfect algorithm.

This index has several advantages. It is calculated on each path, on each node—the mean of all paths involved—and on the entire tree (mean of all node quality). So it gives a complete answer to the degree of confidence that one can place on parts or on the entire tree. As can be seen in Appendix III, for the tree presented in Fig. 3, the global quality index is equal to 0.97 and the quality indices of all the nodes are over 0.95. It must be taken into account that the calculation of intertextual distances entails a maximum error margin of 5%. Thus an index equal to, or over 95%, can be considered as an excellent result. Only six paths must be pointed out due to their quality indexes being under 0.95, although all above 0.90 (Table 3). In other words, each part of the tree retrieves more than 90% of the initial information.

How can this loss of information be explained? Two explanations can be given to account for the differences between the lengths on the tree and their corresponding distances in the matrix.

First, the original values used to calculate the paths are either too approximate, or they contain errors, or they are not Euclidian metrics (some Euclidian distance properties may not have been respected: often these measures are not exactly symmetrical). In such a case, the original matrix contains 'similarity' (or dissimilarity) measures,

Table 3 Problematic paths in Fig. 3

Node	Path	Quality index
17	Thébaïde–Esther	0.918
	Thébaïde–Athalie	0.901
	Alexandre–Athalie	0.939
18	Mithridate–Iphigénie	0.935
20	Britannicus–Alexandre	0.941
	Mithridate–Bérénice	0.941

and it is better to employ the word 'distance' only when referring to the lengths of tree paths which best approximate these 'similarities'.

Second, some degree of 'error' or 'defects' may occur when calculating the tree. Of course, this error does not come from topology theory but from imperfections in procedures and algorithms. On the one hand, original dissimilarities are integers, edges are fractional numbers and, because of roundings, their sums never lead exactly to the same results. The differences between columns A and B in Table 2 or in Appendix III can partly be explained by this. In Fig. 1, the central trunk length is the mean between four paths (*AC, AD, BC, BD*) less the four terminal branches (*AEB, CFD*). On the other hand, when the algorithm classifies a large number of texts, it does not calculate all the paths at the same time. Thus a large amount of computational error rests on some short edges (and the central trunk) which are calculated during the last steps. Table 3 shows it occurring in the two following cases:

- long paths, linking two peripheral texts, which pass through several common edges and a relatively short medial trunk (the final fit is done on this part of the tree). For example, *Thébaide–Esther* and *Alexandre–Athalie*.
- paths between centrally located texts directly linked to the medial trunk as is the case with *Iphigénie-Mithridate* . . .

Here the cross-shape tree is the major factor. When these 'problematic' paths are opened, almost all their lengths are predetermined by previous steps in the procedure (Appendix III) without possible recalculation (as explained in the conclusion below, this is the major improvement that needs to be made with this algorithm).

A great number of experiments leads to a provisional quality scale:

- 0.95 and above: excellent (because the index value falls within the margin of error of the original calculation);
- 0.85–0.95: reliable for the whole tree but particular attention must be given to the 'critical' edges;
- 0.75–0.85: second-rate: the whole tree must be carefully looked at and can be accepted under certain conditions. It is necessary to draw the reader's attention to some edges which are not reliable (the scores of which are under 0.85);
- under 0.75: rejection of the tree. If this score applies to a single edge or a single node only, it would be possible to accept the tree, but it is necessary to alert the reader to the weak spot of this part of the graph.

This quality scale must be used with care, and it is necessary to emphasize that tree quality can decrease as the number of classified texts increases. As a matter of fact, when *n* is large, it occasionally occurs that, for the last steps, the 'four points condition' is not satisfied for some of these nodes. Two solutions are possible: stop drawing the tree or, as is usually done, lower the requirements by degrading the four points condition. But this problem usually occurs with very short edges so that the overall quality index does not decrease markedly.

Furthermore, the quality index can be combined with another procedure to estimate the *confidence* that one can put in an entire tree. Luong has proposed the following solution (Luong, 1988; Ruhlman, 2003). The algorithm starts with the two nearest points (out of a total of *n* points). If their relations to all other texts fulfils the 'four points condition' (see formula 4 above), the maximum number of opposites of this couple is: $(n-2)(n-3)/2$. This case can be generalised. Given a group X of k texts $(y_1, \ldots, y_k)$ within a corpus of n texts $(2 \leq k < n-2)$, X has a number of potential opposites equal to $(n-k)(n-k-1)/2$. It is the maximal possible result or 'theoretical score' (S_{th}). For each potential group of texts, the number of actual opposites gives the 'observed score' (S_{obs}). A 'confidence index' is calculated by dividing S_{obs}

by S_{th}. This index varies between 1 (the texts of X are opposed to all others $n - k$ texts) and 0. It is calculated during each iteration. For the next steps, this index will obviously have fewer possible values and a single failure of the four point condition entails a huge decrease in the index.

Luong's index raises an interesting question: how far must the four point condition be degraded in order to obtain one single group? Of course, if this index is low, less faith can be put in this part of the tree, even if the quality seems to be acceptable.

In any case, it is necessary to publish trees with clear indications as to the following: are the original distances reduced? What are the main values of the quality index? Which are the paths and/or the nodes that are required to degrade the four point condition, and what is the extent of this degrading?

When large corpora are classified, these questions will be of prime importance. This becomes evident, as seen in the context of an experiment carried out with French literature.

5 An Experiment with French Literature

At the beginning of 2002, prof. E. Brunet (University of Nice, France) arranged an experiment in order to evaluate our method of authorship attribution (Labbé and Labbé, 2001). He chose a series of texts which omitted details of authors or titles, and sent them to us. He asked us the precise question: 'which are the texts written by the same author?' In July 2002, a report presenting the results of the experiment was sent to Brunet (Labbé, 2002). Then Brunet revealed the authors' names and titles of the books (see Appendix IV). He chose in fact eleven different authors, two novels within the work of each of these authors and two excerpts (of about 10 000 words length) from each novel. In addition to these forty-four texts, he made up six 'chimerae', sticking together one page out of each novel. These collages or paste-ups were made deliberately: is the algorithm able to recognise the particular nature of these mongrel texts?

The correct answer to Brunet's question was given in the case of the excerpts drawn from the same books. In addition, this answer was given by combining direct examination of the distance matrix with an automatic clustering performed by cluster analysis. The 'success' can be explained by the fact that this traditional classification method was perfectly adapted to data organisation by pairs. However, tree-classification introduces a more sophisticated level of information (Fig. 5).

The mean quality index of this tree is 0.96 and there is no node quality index below 0.92. Of the 1225 paths in this graph, all have a quality index above 0.90, except for the twenty-eight paths the quality indices of which lie within the range [0.90–0.83] (these paths involve some of the 'strange cases' discussed subsequently). Thus this tree can be considered to be 'excellent' but not 'perfect.' What accounts for this slight information loss?

Once again, the tree presents a cross-shape indicating that most of these texts are roughly 'equidistant'. This is the most complicated situation for the algorithm to handle because it has to work with very slight differences. During the three first iterations, the majority of pairs are joined together without difficulty: the Luong confidence index is always equal to one. But, after this simple first task, the quasi-totality of the tree length is set by the fixed location of the first nodes of the tree as can be clearly seen in Fig. 5 (nodes are often very near the centre of the tree). Then a large number of paths need to be recalculated within a very small margin (central trunk and some adjacent edges). More than five hundred paths go along this very short trunk and its few adjacent edges, and these edges are the last part of the tree to be calculated using the mean of what is left when nearly all the lengths are already fixed. In the present case, the algorithm can no longer satisfy the 'four point conditions' for all these remaining calculations. Generally, one or two tree-distances—among the large number of those which link each of these last 'pre-groups' to all other nodes—are insufficiently long.

All the 'chimerae' created by E. Brunet in order to 'fool' the algorithm are correctly grouped at the top of the tree, clustered together and attached to the central trunk very near the root. To appreciate
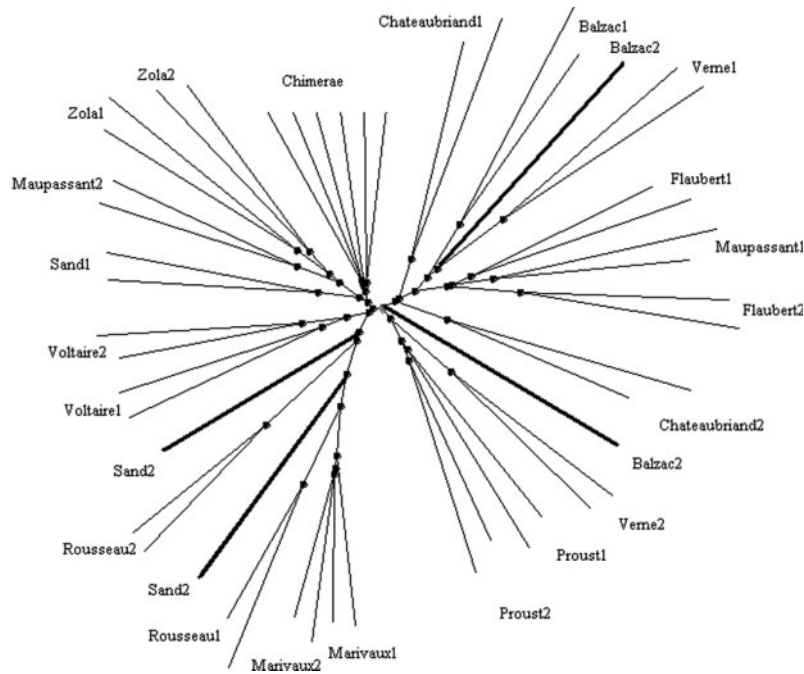
Fig. 5 Tree-classification of the 'Brunet corpus' (without reduction)

the accuracy of classification of the 44 other texts, two levels should be considered.

Firstly, is the classification able to recognise the pairs of excerpts drawn from the same book? Within these twenty-two pairs, twenty are correctly classified. Which are the two odd ones (marked with bold lines in the tree)?

– The two excerpts drawn out of *La Mare au diable* by Georges Sand (1846) are located very close together in the same cluster (South-West) but they are clustered with the two excerpts from l'*Emile* by Jean-Jacques Rousseau (1762). These books treat the same theme, but they are separated by an interval of nearly a century. This is not entirely surprising because it is common knowledge that at the end of her life, Georges Sand adopted a 'Classical' style and used 'Neo-classical' themes about rural life.
– The second anomaly comes from *Le cousin Pons* (Balzac, 1847). The first excerpt is correctly associated with *Les Chouans* (1841) by the same author. The second excerpt is not really

classified: it is attached to the tree root at the end of the classification. In other words, this text seems to be a kind of 'stranger'. In fact, it contains a large number of peculiar words made up by Balzac in order to 'reproduce' the speech of someone speaking French with a German accent.

Secondly, are different books by a same author correctly classified?

Among the eleven authors, only six have all their writings grouped together without ambiguity (Marivaux, Rousseau, Voltaire, Proust, Flaubert, and Zola). This result is not surprising. E. Brunet selected a number of difficult cases. For example, for some authors, he chose very different works: Chateaubriand 1 (*Atala*, 1801) is a novel and Chateaubriand 2 is a biography (*La vie de Rancé*, 1844); nevertheless, they are linked with the same node (through the central trunk). As often as possible, he also selected texts separated by a large interval of time, or known to be very different in style or content. During his lifetime, an author can

considerably change his themes and his style, which together enlarge the intertextual distances within his work. This corpus gives at least three fascinating examples of this phenomenon: G. Sand, J. Vernes, and G. de Maupassant.

For the two first authors, it can be seen that a long period of time elapsed between the two books. For example, G. Sand, who was very 'Romantic' at the beginning of her work was very 'Neo-classical' at the end of it. G. de Maupassant is also an interesting case. His first novel (*Une vie*, 1883) is known to be a kind of remake of *Madame Bovary* (Flaubert, 1857), and it is said that G. Flaubert corrected the manuscript. Logically, *Une vie* is located in the middle of the Flaubert's cluster and very near *Madame Bovary*. Six years (and three novels) later, Maupassant seemed to be fully emancipated from Flaubert's influence and *Pierre et Jean* appears to be very near to E. Zola (at the opposite side of the tree . . .). Thus, it may be deemed that the tree classification in this case was not really erroneous and that, on the contrary, it provides an accurate tool for literary history.

Consequently, as far as a strict authorship attribution is concerned, it is clear that the texts to be compared must belong to the same genre; they also must be contemporaneous and without too many 'anomalies' (like the funny French of the German friend of Balzac's *Cousin Pons*). At least, it must be kept in mind that tree-classification does not provide 'proof' and that these graphs must be carefully analysed. A tree can only suggest certain affinities or, on the other hand, some impossibilities. Furthermore, intertextual distances provide numerically standardized answers and not only 'visual' ones. It must be emphasized that, for this 'Brunet' corpus, intertextual distances alone allowed us correctly to identify more than three-fourths of the couples within the many thousand possibilities, and gave the most probable solutions for all the others without error.

6 Conclusions

A more sophisticated algorithm is planned which should construct trees in two main steps: Firstly,

following the procedure presented in this article, all points and nodes should be situated, and edges should be allotted provisional lengths. Then all these edges should be recalculated in order to distribute the required adjustments along the whole length of the different paths, and not merely on a few sections of them as is presently done. This solution is necessary to adapt the tree-classification fully to large corpora of several hundred texts.

Fundamentally, tree quality relies on the manner by which proximity between texts is measured. Of course, texts must have been carefully checked and word spellings strictly standardized, because it is useless to attempt accurately to measure phenomena whose observations are made inaccurately. Secondly, these measures of proximity must be as near as possible to the usual Euclidian distance characteristics. Within certain limits, it is indeed mostly the case with intertextual distance. Thus combining intertextual distance with tree-classification provides an accurate tool for researchers to help them process an ever-growing body of electronic texts available for literary studies. This procedure will soon be adapted to English texts and corpora.

Note:

- software are online (http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/Labbe)
- the work of Racine and the Brunet corpus are freely available online (http://ota.ahds.ac.uk/2466)

helpful comments and advice about intertextual distance and for his final edition of this text.

References

Above all, tree classification is used in genetics and toxonomy. A bibliographical and a software survey can be found on: http://evolution.genetics.washington.edu.

Barthélémy, J-P. and Guénoche, A. (1988). Les arbres et les représentations de proximité. Paris. Dunod (english translation: *Trees and Proximity Representations*, New York, Wiley, 1991).

Hubert, P. and Labbé, D. (1988). A model of vocabulary Partition. *Literary and Linguistic Computing*, **3**(4): 223–225.

Juilliard, M. and Luong, X. (1997). Words in the Hood. *Literary and Linguistic Computing*, **12**(2): 71–78.

Juilliard, M. and Luong, X. (2001). On consensus between tree-representation of linguistic data. *Literary and Linguistic computing*, **16**(1): 59–76.

Labbé, D. (1990). Normes de saisie et de dépouillement des textes politiques. Grenoble. Cahier du CERAT. http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/Labbe.

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

Labbé, D. (2002). *Qui a écrit quoi*? Grenoble. Cerat. This text is on line: http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/Labbe

Labbé, C. and Labbé, D. (2003). La distance intertextuelle. *Corpus*, **2**(1): 95–118.

Luong, X. (1988). Méthodes d'analyse arborée. Algorithmes, applications. Thèse pour le doctorat ès sciences. Université de Paris V.

Merriam, T. (2002). Intertextual distances between Shakespeare plays, with special reference to Henry V (verse). *Journal of Quantitative Linguistics*, **9**(3): 260–273.

Merriam, T. (2003a). An application of authorship attribution by intertextual distance in English. *Corpus*, **2**(1): 167–182.

Merriam, T. (2003b). Intertextual distances, three authors. *Literary and Linguistic Computing*, **18**(4): 379–388.

Ruhlman, M. (2003). Analyse arborée. Représentation par la méthode des groupements. Grenoble, Polytech' – CERAT, août 2003. http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/Labbe

Spencer, M., Bordalejo, B., Robinson, P., Howe, C. (2003). How reliable is a stemma? An analysis of Chaucer's Miller's Tale. *Literary and Linguistic computing*, **18**(4): 407–422.

Notes

1 Terminology is sometimes confusing. In geometry, the word 'edge' is used for a more than 2-dimension object such as a cube or a tetrahedron. For graphs and trees, 'edge' is used for a segment—a straight line bounded on both sides by node(s) or vertex(ices)—even if graphs and trees are almost always 2-dimensional objects.

Appendix I Racine's work

No	Title	Genre	Date	Length (tokens)	Types
1	La Thébaïde	Tragedy	1664	13,813	1,313
2	Alexandre	Tragedy	1665	13,864	1,372
3	Andromaque	Tragedy	1667	15,076	1,392
4	Les Plaideurs	Comedy	1668	8,041	1,312
5	Britannicus	Tragedy	1669	15,387	1,637
6	Bérénice	Tragedy	1670	13,242	1,346
7	Bajazet	Tragedy	1672	15,297	1,507
8	Mithridate	Tragedy	1673	15,091	1,550
9	Iphigénie	Tragedy	1674	15,782	1,604
10	Phèdre	Tragedy	1677	14,394	1,775
11	Esther	Tragedy	1689	11,147	1,656
12	Athalie	Tragedy	1691	15,492	1,656
Entire work				166,626	4,322

Appendix II Intertextual distances between tragedies by Racine

	Thébaïde	Alexandre	Andromaque	Britannicus	Bérénice	Bazajet	Mithridate	Iphigénie	Phèdre	Esther	Athalie
Thébaïde	0	242	245	260	276	258	242	254	275	317	295
Alexandre	242	0	231	233	260	251	238	241	266	315	295
Andromaque	245	231	0	214	227	202	208	223	245	331	306
Britannicus	260	233	214	0	208	206	218	222	246	314	302
Bérénice	276	260	227	208	0	220	206	226	250	346	329
Bazajet	258	251	202	206	220	0	204	230	244	325	304
Mithridate	242	238	208	218	206	204	0	193	223	313	288
Iphigénie	254	241	223	222	226	230	193	0	216	293	280
Phèdre	275	266	245	246	250	244	224	216	0	286	275
Esther	317	315	331	314	346	325	313	293	286	0	214
Athalie	295	295	306	302	329	304	288	280	275	214	0

Appendix III Calculation steps and quality indices for tree-classification of Racine's tragedies

Node	Paths	Texts distances*	Tree distances*	Path's quality	Node's quality
12	Esther–Athalie	214.5	214.5	1.00	1.00
13	Thébaïde–Alexandre	242.2	242.2	1.00	1.00
14	Phèdre–Esther	285.7	289.8	0.99	
	Phèdre–Athalie	275.2	271.1	0.99	0.99
15	Britannicus–Bérénice	209.4	209.4	1.00	1.00
16	Iphigénie–Phèdre	215.5	219.1	0.98	
	Iphigénie–Esther	293.1	292.4	1.00	
	Iphigénie–Athalie	280.4	273.8	0.98	0.99
17	Thébaïde–Iphigénie	253.9	245.1	0.97	
	Thébaïde–Phèdre	275.3	269.9	0.98	
	Thébaïde–Esther	317.3	343.2	0.92	
	Thébaïde–Athalie	295.1	324.6	0.90	
	Alexandre–Iphigénie	240.9	233.5	0.97	
	Alexandre–Phèdre	265.5	258.4	0.97	
	Alexandre–Esther	315.4	331.7	0.95	
	Alexandre–Athalie	294.9	313.0	0.94	0.95
18	Mithridate–Thebaide	242.1	239.4	0.99	
	Mithridate–Alexandre	237.8	227.9	0.96	
	Mithridate–Iphigénie	192.5	205.0	0.93	
	Mithridate–Phèdre	223.4	229.9	0.97	
	Mithridate–Esther	313.0	303.2	0.97	
	Mithridate–Athalie	287.5	284.5	0.99	0.97
19	Andromaque–Bazajet	202.2	206.2	0.98	
	Andromaque–Britannicus	213.5	209.5	0.98	
	Andromaque–Bérénice	226.6	219.1	0.97	
	Bazajet–Britannicus	205.5	209.4	0.98	
	Bazajet–Bérénice	219.5	219.0	1.00	0.98
20	Andromaque–Thébaïde	244.9	255.0	0.96	
	Andromaque–Alexandre	231.0	243.5	0.95	
	Andromaque–Iphigénie	222.5	220.6	0.99	
	Andromaque–Phèdre	244.5	245.4	1.00	
	Andromaque–Esther	331.0	318.7	0.96	

Appendix III Continued

Node	Paths	Texts distances*	Tree distances*	Path's quality	Node's quality
	Andromaque–Athalie	305.9	300.1	0.98	
	Bazajet–Thébaïde	257.5	254.9	0.99	
	Bazajet–Alexandre	250.7	243.4	0.97	
	Bazajet–Iphigénie	230.3	220.5	0.96	
	Bazajet–Phèdre	243.8	245.3	0.99	
	Bazajet–Esther	324.6	318.6	0.98	
	Bazajet–Athalie	303.8	300.0	0.99	
	Britannicus–Thébaïde	260.3	258.2	0.99	
	Britannicus–Alexandre	232.9	246.6	0.94	
	Bérénice–Thébaïde	276.0	267.8	0.97	
	Bérénice–Alexandre	260.0	256.3	0.99	
	Britannicus–Iphigénie	221.6	223.8	0.99	
	Britannicus–Phèdre	246.3	248.6	0.99	
	Britannicus–Esther	314.2	321.9	0.98	
	Britannicus–Athalie	301.5	303.2	0.99	
	Bérénice–Iphigénie	225.9	233.4	0.97	
	Bérénice–Phèdre	250.0	258.2	0.97	
	Bérénice–Esther	346.2	331.5	0.96	
	Bérénice–Athalie	328.5	312.9	0.95	
	Mithridate–Andromaque	208.4	205.4	0.99	
	Mithridate–Bazajet	203.8	205.3	0.99	
	Mithridate–Britannicus	217.6	208.5	0.96	
	Mithridate–Bérénice	206.1	218.2	0.94	0.97

*number of different type words out of 1000 words.

Appendix IV The Brunet corpus

No	Author	Titles (date)
1	Marivaux 1.1	La vie de Marianne (1731)
2	Marivaux 2.1	Le paysan parvenu (1735)
3	Voltaire 1.1	Zadig (1747)
4	Voltaire 2.1	Candide (1759)
5	Rousseau 1.1	La nouvelle Héloïse (1761)
6	Rousseau 2.1	L'Emile (1762)
7	Chateaubriand 1.1	Atala (1801)
8	Chateaubriand 2.1	La vie de Rancé (1844)
9	Balzac 1.1	Les Chouans (1841)
10	Balzac 2.1	Le cousin Pons (1847)
11	Sand 1.1	Indiana (1832)
12	Sand 2.1	La mare au diable (1846)
13	Flaubert 1.1	Madame Bovary (1857)
14	Flaubert 2.1	Bouvard et Pécuchet (1881)
15	Maupassant 1.1	Une vie (1883)
16	Maupassant 2.1	Pierre et Jean (1888)
17	Zola 1.1	Thérèse Raquin (1867)
18	Zola 2.1	La bête humaine (1890)
19	Verne 1.1	De la terre à la lune (1865)
20	Verne 2.1	Les secrets de Wilhelm Storitz (1905)

Appendix IV Continued

No	Author	Titles (date)
21	Proust 1.1	Du côté de chez Swann (1913)
22	Proust 2.1	Le temps retrouvé (1927)
23	Marivaux 1.2	La vie de Marianne (1731)
24	Marivaux 2.2	Le paysan parvenu (1735)
25	Voltaire 1.2	Zadig (1747)
26	Voltaire 2.2	Candide (1759)
27	Rousseau 1.2	La nouvelle Héloïse (1761)
28	Rousseau 2.2	L'Emile (1762)
29	Chateaubriand 1.2	Atala (1801)
30	Chateaubriand 2.2	La vie de Rancé (1844)
31	Balzac 1.2	Les Chouans (1841)
32	Balzac 2.2	Le cousin Pons (1847)
33	Sand 1.2	Indiana (1832)
34	Sand 2.2	La mare au diable (1846)
35	Flaubert 1.2	Madame Bovary (1857)
36	Flaubert 2.2	Bouvard et Pécuchet (1881)
37	Maupassant 1.2	Une vie (1883)
38	Maupassant 2.2	Pierre et Jean (1888)
39	Zola 1.2	Thérèse Raquin (1867)
40	Zola 2.2	La bête humaine (1890)
41	Verne 1.2	De la terre à la lune (1865)
42	Verne 2.2	Les secrets de Wilhelm Storitz (1905)
43	Proust 1.2	Du côté de chez Swann (1913)
44	Proust 2.2	Le temps retrouvé (1927)
45	Brunet's collages:	Pages no 1 out of all the texts
46		Pages no 10 out of all the texts
47		Pages no 20 out of all the texts
48		Pages no 30 out of all the texts
49		Pages no 40 out of all the texts
50		Pages no 50 out of all the texts