COGNITIVE PHILOLOGY

No 1 (2008)

Philology and Information Theory

Paolo Canettieri[1], Vittorio Loreto[2],
Marta Rovetta, Giovanna Santini[3]

[1]Sapienza Università di Roma, Dipartimento di Studi Europei e Interculturali Piazzale Aldo Moro 5, 00185 Roma – Italia (paolo.canettieri@uniroma1.it). [2]Sapienza Università di Roma, Dipartimento di Fisica, Piazzale Aldo Moro 5, 00185 Roma – Italia. [3] Università di Arezzo, Università di Cassino, Università di Viterbo.

ABSTRACT: Methods borrowed from Information Theory are applied to the traditional textual criticism. A critic to the raw cladistic methods and an interpretation of the dichotomy-phenomenon are furnished. The same methods are applied to 13th century italian poetry to determine authorship attributions and to verify common accepted literary taxonomy.

1. Introduction

Philology is a human science primarily applied to literary texts and traditionally divided into lower and higher criticism. Lower criticism tries to reconstruct the author's original text and higher criticism is the study of the authorship, style, and provenance of texts. The employment of methods borrowed from information theory makes it possible to bring together methodologically some of the sectors of the two fields. The outcome of the experiments in both text criticism and text attribution has been encouraging. In the former, the tests performed on three different traditions have provided results very similar to those obtained by traditional methods in a great amount of time. The experiments carried out both at the levels of 13th century Italian poets and schools have shown that it is possible to draw texts closer to one another. Furthermore, the method we have used allows establishing the attribution of anonymous writings: in particular, the attribution of the poem *Il Fiore* to Dante Alighieri is probably to be excluded in favour of his master Brunetto Latini.

We'll now present a short and oversimplified summary of our information-theory oriented approach. For a more formal treatment we refer to Benedetto *et al*., 2002, 2003 and Baronchelli *et al*., 2005. Born in the context of electric communications, information theory has acquired, since the seminal paper of Shannon (1948), a leading role in many other fields as computer science, cryptography, biology and physics (Zurek, 1990).

In information theory the word information acquires a very precise meaning, namely that of the so-called *entropy* of the string, a measure of the *surprise* the source emitting the sequences can reserve to us. Suppose the surprise one feels upon learning that an event *E* has occurred depends only on the probability of *E*. If the event occurs with probability 1 (sure!) our surprise in its occurring will be zero. On the other hand if the probability of occurrence of the event *E* is quite small our surprise will be proportionally large.

One could now ask how is it possible to extend the definition of the entropy for a generic string of character without any reference to its source. This is the typical case when one has a text written for which the source or its statistical properties could be typically unknown. Among the many equivalent definitions of entropy the best for this case is the so-called Chaitin – Kolmogorov complexity or algorithmic complexity: the algorithmic complexity of a string of characters is given by the length (in bits) of the smallest program which produces as output the string. A string is said *complex* if its complexity is proportional to its length. This definition is really abstract, in particular it is impossible, even in principle, to find such a program. Since this definition tells nothing about the time the best program should take to reproduce the sequence, one can never be sure that somewhere else it does not exist another shorter program that will eventually produce the string as output in a larger (eventually infinite) time (Shannon, 1948; Khinchin, 1957).

One has to recall now that there are algorithms explicitly conceived to approach the theoretical limit of the optimal coding. These are the file compressors or zippers. It is then intuitive that a typical zipper, besides trying to reduce the space occupied on a memory storage device, can be considered as an entropy meter. Better will be the compression algorithm, closer will be the length of the zipped file to the minimal entropic limit and better will be the estimate of the entropy provided by the zipper. It is indeed well known that compression algorithms provide a powerful tool for the measure of the entropy and more in general for the estimation of more sophisticated measures of complexity.

A great improvement in the field of data compression has been represented by the so-called LZ77 algorithm (Lempel, Ziv, 1977). This algorithm zips a file by exploiting the existence of repeated sub-sequences of characters in it. Its compression efficiency becomes optimal as the length of the file goes to infinity (Wyner, Ziv, 1994). It is interesting to briefly recall how it works. The LZ77 algorithm finds duplicated strings in the input data. The second occurrence of a string is replaced by a pointer to the previous string given by two numbers: a distance, representing how far back into the window the sequence starts, and a length, representing the number of characters for which the sequence is identical. For example, in the compression of an English text, an occurrence of the sequence *the* will be represented by the pair (d,3), where d is the distance between the occurrences we are considering and the previous one. It is important to mention as the zipper does not recognize the word *the* as a *dictionary word* but only as a specific sequence of characters without any reference to the words belonging to a specific dictionary. The sequence will be then encoded with a number of bits necessary to encode d and 3. Roughly speaking the average distance between two consecutive *the* in an English text is of the order of 10 characters. Therefore the sequence *the* will be encoded with less then 1 byte instead of 3 bytes.

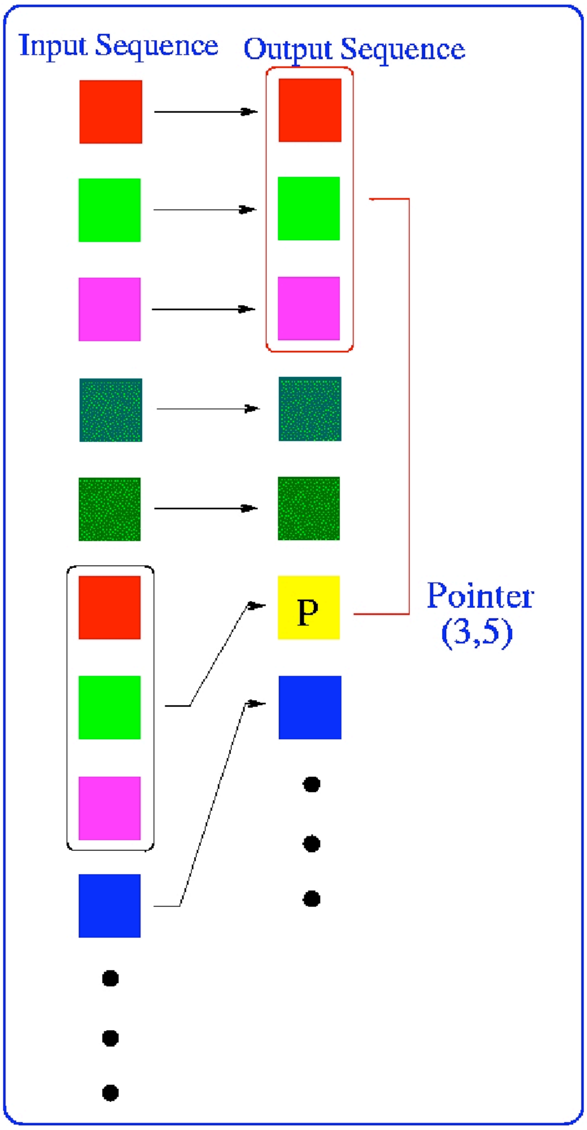


Figure 1. The LZ77 algorithm searches in the look-ahead buffer for the longest substring (in this case substring of colors) already occurred and replaces it with a pointer represented by two numbers: of the matching length and distance.

It is interesting to recall the notion of relative entropy (or Kullback-Leibler divergence: Cover, Thomas, 1991) which is a measure of the statistical remoteness between two distributions. A linguistic example will help to clarify the situation: transmitting an Italian text with a Morse code optimized for English will result in the need of transmitting an extra number of bits with respect to another coding optimized for

Italian: the difference is a measure of the relative entropy between, in this case, Italian and English (supposing the two texts are each one archetypal representations of their Language, which is not).

We should remark that the relative entropy is not a distance (metric) in the mathematical sense: it is neither symmetric, nor does it satisfy the triangle inequality. For our purpose it is crucial to define a true metric that measures the actual distance between sequences. Our proposal is based on the computation of the relative entropy. There exist several ways to measure the relative entropy. One possibility is of course to follow the recipe described in the previous example: using the optimal coding for a given source to encode the messages of another source.

Here we follow the approach recently proposed (Benedetto et al., 2002), which is similar to the approach by Merhav and Ziv (1993). The aim is that of defining a notion of remoteness between two text, from now on identified as A and B, exploiting the properties of the relative entropy. In particular we start reading sequentially file B and search in the look-ahead buffer of B for the longest sub-sequence already occurred only in the A part. This means that we do not allow for searching matches inside B itself. As in the usual LZ77, every matching found is substituted with a pointer indicating where, in A, the matching subsequence appears and its length. This method allows us to measure (or at least to estimate) the cross-entropy between B and A, and consequently their relative entropy.

Now we address the problem of defining a distance between two generic sequences A and B. A distance D is an application that must satisfy three requirements: positivity, symmetry and triangular inequality. As it is evident the relative entropy does not satisfy the last two properties while it is never negative. Nevertheless, in order to obtain a real mathematical distance one can define a symmetric quantity by using the symmetrized relative entropy which can be made satisfying the triangular inequality with a suitable prescription (see Baronchelli *et al.*, 2005 for details). It is important to quote that there exist similar approaches to define a distance between pairs of sequences (Otu and Sayood, 2003, Li *et al.*, 2004, Kaltchenko, 2004).

Starting from the distance matrix one can construct a tree representation of the corpus. Our method is mutuated by the phylogenetic analysis of biological sequences (Cavalli-Sforza and Edwards, 1967, Felsenstein, 1984) and takes as input the distance matrix, i.e. a matrix whose elements are the distances between pairs of texts (sequences) and produces as output a tree whose leaves are the elements of the corpus. With these trees a classification is achieved by observing clusters that are supposed to be formed by similar elements.

2. Phylogenetic trees and stems

Behind the philological activity lies a primary cognitive requirement of man which consists in specifying the sense or literal meaning of what has been said by another man or group of men, generally of greater authority. In this sense, the mental outline controlling such an activity can therefore be synthesized as: /A has not said x, A has said y/. This requirement derives, essentially, from the loss of information inherent in any type of data transmission. In the pre-technological age, information was transmitted in an exclusively oral form and useful mental techniques have been necessary to reconstruct the original dictation, a practice certainly supported by the hierarchical structure of society.

Before the invention of printing, texts was copied by hand and variants were typically introduced by the scribe. Textual criticism examines the extant copies to sort through the variants and to establish a ‛critical text’ as close as possible to the original. The difficulty is that it is not always apparent which variant is original and which one is innovative.

There are two different approaches to textual criticism: copy text editing and ecdotics. In the former, the textual critic selects a base text from a manuscript and makes emendations in places where it appears wrong. In the latter, ecdotic critic examines the variants in order to find patterns of error, requiring to reconstruct the history of the text. According to the principle that “a community of error implies a unity of origin”, the critic determines the relations among the extant manuscripts, so as to place them in a family tree (*stemma codicum*).

Here we proposed an integrated method for "stemma" reconstruction which combines the traditional ecdotics approach with an information theory oriented methodology. In the information theory oriented approach the key element is the definition of a notion of "distance" between pairs of manuscripts and its computation is based on data compression techniques. All the pair-wise distances among all the manuscripts in the corpus form the so-called distance matrix which is used to construct a phylogenetic-like tree by minimizing the net disagreement between the matrix pairwise distances and the distances measured on the tree. This phylogenetic-like tree is compared with the others constructed with the traditional ecdotics method and a consensus tree is obtained.

As is well known, tree-like structures are designed to represent phylogenetic relations among genes or organisms and have been used several times in humanistic studies to classify the genealogic relations among the cultures of different populations (Shevoroshkin, 1989), linguistic families (Greenberg 1957; Stevick 1963; Maher 1966; Bender, 1976; Picardi 1977; Koerner, 1981 e 1983; Flight 1988; Bateman *et al.*, 1990; Hoenigswald 1990; Shevoroshkin and Woodford, 1991; Cavalli Sforza *et al.*, 1994; Gray and Jordan, 2000; Searls, 2003) and manuscript witnesses of a text (Platnick and Cameron, 1977;

Timpanaro 1985; Antonelli, 1985; Cameron, 1987; O'Hara, 1996; Montanari 2003). As regards textual criticism, the software developed for use in biology has been first applied by the *Canterbury Tales Project* to determine the relationships among the 84 surviving manuscripts and four early printed editions of the *Canterbury Tales* (O'Hara and Robinson, 1993; Barbrook *et al*., 1998; Robinson *et al*., 2003; Spencer *et al*., 2003). The texts of different manuscripts are entered into a computer, which records all the differences among them. The manuscripts are then grouped together according to their shared characteristics.

The trees of the biologists are generally formed on the basis of mutations, which are therefore very similar to the variants in ecdotics (Spencer and Howe, 2001 and 2002; Spencer *et al*., 2004). They present properties which are similar to those found in the evolution of texts, as they admit lost elements. Despite the identification of typical analogies existing among the various trees, scholars in the single disciplines have not adopted a uniform terminology. We feel that the terminology realized by the mathematicians who have studied the theory of the graphs is the most appropriate and precise (Bollobás, 1998; Douglas, 2001). 'Graphs' are mathematical structures formed by 'nodes' (single elements identified by a label) and edges or links connecting pairs of nodes. The trees are a particular family of graphs which can be defined as a connected and acyclic oriented graph. In a tree there exists a privileged node called root. The nodes connected to the root are called children and the root is their ancestor. In the trees the children are not interconnected by any edge and there is only one edge linking them to the root. However, various edges can be traced for each child-node of the root so that it may be connected to new nodes that we call children of this node and which, in turn, will be called ancestors. The terminology and the initial ties are extended to the new descendants and to the descendants' descendants, and so on recursively. Therefore each node has only one ancestor but in principle many children. The nodes without children are called leaves.

Any node A is defined as ancestor of a node B if there exists a node C which is the child of A and ancestor of B. This is a recursive definition. Since trees are mathematical structures defined recursively, it is often easier to express definitions, properties or algorithms on trees recursively. Broadly speaking, fathers are also called ancestors and the nodes of the same ancestor are defined as being dominated by that ancestor. All the nodes of a tree are dominated by the root.

In a graph the 'path' between two nodes is a sequence of edges that have to be crossed in order to reach one of the two nodes starting from the other. The nodes that are linked at the root by means of a path constituted by the same number of edges forms a set which is called 'level' (the equivalent of 'stemma level' in ecdotics). Each level is univocally identified by this number and is separated from the other levels: the greater the number defining a level, the lower is the level itself; the smaller the number, the higher is this level. The lowest level is always formed by leaf-nodes, the highest level from the root.

In genetics trees can be rooted or unrooted. The creation of these trees is based on different principles and the results depend on the method that has been chosen. The branching structure is called topology of the tree. The topology of trees with roots is also represented by setting in brackets the sequences that derive from the same node. For example the first tree of fig. 2 could also be represented grouping (A,B) and (C,D) first, and then these two new taxa (namely any taxonomic unit, e.g. species, populations, sequences, morphological features) as follows: ((A,B),(C,D)).

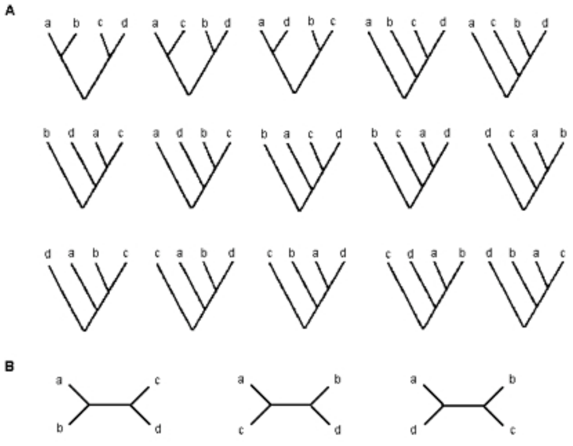


Figure 2. Possible rooted (A) and unrooted (B) trees with 4 elements.

Various trees are possible for a given number of species (m). For example if m = 4 there are 15 possible trees with roots and 3 without roots, as seen in fig. 2. The number of trees grows very rapidly with the number of taxonomic units (Edwards and Cavalli-Sforza, 1967): if m = 10 we will have more than 34 million possible trees, of which only one will be the true one. This means that it is practically impossible to assess the goodness of all the phylogenetic trees.

Phylogenetic trees generally regard species or populations and are constructed by comparing genes. One of the simplest methods, although it is not without errors, with particular regard to the topology of the tree, is the one based on the distance matrix.

A matrix of (developmental) distances is constructed among the sequences, for instance according to the mutations encountered. For example the Fitch-Margoliash (Fitch, Margoliash 1967) method:

	1) Man	2) Chimpanzee	3) Gorilla	4) Orangutan
2) Chimpanzee	0.095			
3) Gorilla	0.113	0.118		
4) Orangutan	0.183	0.201	0.195	
5) Gibbon	0.212	0.225	0.225	0.222

We examine the table and look for the minor distance. The two sequences (1,2) are grouped together and considered as a new sequence. The length of the branches stretching from the bifurcation to nodes 1 and 2 is assumed to be equal (thus = 0.095/2). The table of the distances where 1 and 2 are grouped together is recalculated. Each distance will be calculated as average of the distances from 1 and from 2.

	u) (1,2)	3) Gorilla	4) Orangutan
3) Gorilla	(0.113+0.118)/2 = 0.115		
4) Orangutan	(0.183+0.201)/2 = 0.192	0.195	
5) Gibbon	(0.212+0.225)/2 = 0.218	0.225	0.222

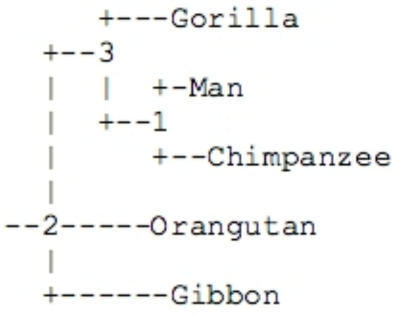
At this point the minor distance is between (1,2) and 3. The length of the branch leading to 3 will be half the distance between (1,2) and 3, while the length of the branch leading to the bifurcation between 1 and 2 will be such that - if added to the length of branch 1 - it will be the same as the branch leading to 2. This procedure ensures that the distances in the reconstructed tree are close to the ones that have been observed. (1,2) and 3 are grouped in ((1,2),3) and the matrix of the distances is recalculated:

	v) ((1,2),3)	4) Orangutan
4) Orangutan	(0.183+0.201+0.195)/3= 0.193	
5) Gibbon	(0.212+0.225+0.225)/3= 0.221	0.222

At this point the minor distance is between ((1,2),3) and 4. ((1,2),3) and 4 are grouped in (((1,2),3),4) and the distance matrix is recalculated:

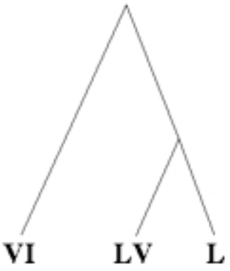
	z)(((1,2),3),4)
5) Gibbon	(0.212+0.225+0.225+0.222)/4 = 0.221

Finally, the following tree is obtained:



```
   +---Gorilla
+--3
|  |  +-Man
|  +--1
|     +--Chimpanzee
|
--2-----Orangutan
|
+------Gibbon
```

The application of methods like this to linguistics and textual criticism is now quite diffused, although it is still far from representing the real developmental processes appropriately. The methods are based on the comparison of sequences of characters and involves a grouping of those which appear closer to one another: it is obvious that in this way the trees are always dicotomic. It is instead clear that in both linguistics and ecdotics the vicinity of sequences does not necessarily imply a phylogenetic relation. This concept can be illustrated by two extreme cases: translations in different languages operated on the original or on the archetype would be placed in an extremely distant position one from the other. Otherwise, if we were to use the phylogenetic method to establish relations among interdependent languages, such as classical Latin (L), vulgar Latin (LV) and vulgar Italian (VI), the developmental process being L > LV > VI, L and LV would be sharing the same tree, while VI would be on another branch:



Likewise, if three copies of a text are in a dependence relation (A,B,C), but C presents a *lacuna* that A and B do not have, or has any singular alteration (interpolation, significant error, etc.), the tree produced by the phylogenetic analysis method will draw A and B closer with respect to C: ((A,B),C). We therefore come across a theoretical error which is similar to the one evidenced in the method of Dom Quentin and of his followers, who tried to automatize textual criticism (Quentin 1926; Froger, 1968).

In order to obtain scientifically remarkable results it is instead necessary to apply to the methods of phylogenetic analysis the most relevant achievements of the Lachmann method, based on evident errors and innovations (Maas 1957, Montanari 2003). This method can be summarized as follows: given two copies A and B of the same text, if: A contains an error (e1); B contains the same error; A contains an error e2 (different from e1); B does not contain e2; then: A derives from (<) B. This is an implication formula which can be expressed as follows:

[e1 □ A □ B, e2 □ A] □ A < B.

On the other hand, if: A contains an error (e1); B contains the same error; B contains an error e2 (different from e1); A does not contain e2; then: B derives from A:

[e1 □ A □ B, e2 □ A] □ B < A.

This formula detects the case of the so-called *descripti* (i.e. copies whose sources have been conserved), but it could be generalized. For example if: A contains an error (e1); B contains the same error; A contains an error e2 (different from e1); B does not contain e2; B contains e3 (different from e1 and from e2); A does not contain e3; then A and B are connected by the same node, which represents a lost copy. The formula can be expressed as follows:

[e1 □ A □ B, e2 □ A, e3 □ B] □ (A,B).

In the case of three copies of a lost original, the formula will be:

[e1 □ A □ B □ C, e2 □ A, e3 □ B, e4 □ C] □ (A,B,C).

The variants (v) can also be used for creation of the tree. If we have witnesses A,B,C under the following condition (vs indicates that the variant implies the same textual site):

[e1 □ A □ B □ C, e2 □ B □ C, e3 □ A, e4 □ B, v1 □ A □ B vs v2 □ C],

we can consider v2 as separative variant of C with respect to B, in the same way as we consider e4 as separator of B with respect to C: ((B,C),A).

In case there are 4 copies available and the following conditions take place:

[e1 □ A □ B □ C □ D, e2 □ B □ C □ D, e3 □ A, e4 □ B, e5 □ C, e6 □ D, v1 □ A □ B vs v2□ C □ D] □ (((C,D),B),A).

The 'significant variant' can be used for the creation of the tree, just like 'separative' and 'conjunctive' errors if the level above the place generating the variant is justified by error. Theoretically speaking, this means that a conjunctive variant can associate witnesses even at the higher levels of the stemma and that a variant under the same condition can have a separative value. For instance for a four-level (i.e. with five witnesses) bifurcated stemma:

[e1 □ A □ B □ C □ D □ E, e2 □ B □ C □ D □ E, e3 □ D □ E, e4 □ A, e5 □ B, e6 □ C, e7 □ D, e8 □ E, v1 □ A □ B vs v2 □ C□ D ⊠ E] □ ((((D,E),C),B),A).

The absence of some of the above conditions, in particular the lack of significant errors in one or more witnesses, a very frequent case in real traditions, implies a large number of possible trees and therefore a variety of texts that can be reconstructed mechanically. From a practical point of view, many expedients involving the historico-cultural aspect intervene as support, but in most cases with the consequence that the changes of the hermeneutic or epistemological paradigms correspond to changes in the trees and therefore in the texts reconstructed with their aid.

Another problem posed by the reconstructive method is the tendency to produce bifurcated trees: in this respect, the trees of the philologists and those of the biologists are surprisingly similar (Reeve, 1998; Howe *et al.*, 2001; Howe *et al.*, 2004; Spencer et al., 2004). The percentage of bifurcated textual trees is around 90%. Since Bédier (1928) evidenced this phenomenon, many explanations have been given (Shepard, 1930; Greg, 1931; Whitehead and Pickford, 1951; Castellani, 1957; Kleinlogel, 1968; Fourquet, 1946; Timpanaro, 1985; Reeve, 1986; Grier, 1988; Hall, 1992; Guidi and Trovato, 2004). We feel that this bifurcation of the textual trees depends on the ways in which tradition develops and that this aspect is partly connected with the analogy with the trees of life. We distinguish as follows:

C = child; F = father; i = set of; L = lost; Le = tree leaf; O = original; P = preserved; R = tree root; S = sterile (i.e. O or C never copied); T = tree; W = witness; Ω = archetype; < = descending from.

The 'real' T is given by iF □ iC, i.e. by iW, the 'reconstructed' T by iP. However, it should be pointed out that if an PC descends from a PF in ecdotics we are dealing with *descriptus* and such witness is not generally represented in the reconstructed T, which will be composed by iPC < LF. In the real T, iLe is given by iPSC □ iLSC, in the reconstructed T, iLe is given by iP, with the above warning.

The grade of fertility is given by the ratio between the number of W and F. The grade of fertility in the real tree of Castellani (1957), for example, is 15 F over 53 W, therefore around 3.5. With the same number of W, if the tree had a higher number of levels, fertility would increase.

Each W has from 0 to n probabilities of being copied. Whenever a W meets probability 0 there will be an SC and therefore the increase of the possibilities of extinction of the relative branch. With the same number of W, the wider horizontally is the tradition, the more likely is the extinction of the branches.

The process of decimation is not equally probable since it is much easier for the most ancient codices to disappear, either because they have more time to undergo irreparable material damage or to incur fortuitous events causing loss or decay, or because the evolution of writing can make it very difficult for them to be copied and thus less prolific. The older a W, the less prolific will it be. With the passing of time, prolificity of each single W should, as a rule, diminish, in the same way as the periods of stasis in the copy (for example when there are changes in the reproduction systems: from manuscript to print, etc.) should make sterility increase.

The likelihood of extinction of a branch increases according to the sterility of the witnesses of that branch and to the probability they have of getting lost (for example it is very high if a branch is formed by a single witness copied precociously from the original or from an ancient SC).

Even the density of W distribution on the various branches is not equally probable: the branches in which the W are from the very origins less prolific will become less and less crowded until complete extinction; while the branches in which W are from the start very prolific will be growingly populated and will have greater chance of survival.

The problem of extinction of the branches in the context of textual tradition can be compared to the problem of the "Gambler's Ruin", described in Raup 1991, where it is applied to the extinction of genres in biology. We can compare the number of W in a branch to the number of species in an evolutionary group and to the initial capital of a player, a chronological scale in hundreds of years to that expressed in millions of years of genres and to the temporal scale of the player. Let us imagine that for each interval of a century each W has 50 probabilities out of 100 to survive and consequently to produce new W: as occurs for species and for the player's capital, even the number of W "will undergo a larger or smaller amount of fluctuations following a random itinerary"; it is however certain that the final extinction of the branch is an inevitable tendency and that the higher is the number of W produced the more likely it is for it to survive in the long run. In order to know the probabilities of extinction at each interval, it would be necessary to evaluate if even in the case of the textual tradition the rhythms of copy and loss of W are equivalent, in any case, however, the logic of casual itinerary would remain valid. The model of the "Gambler's Ruin" also explains clearly the phenomenon of asymmetry of the trees (Brambilla Ageno 1975; Weitzmann, 1982, 1985, 1987; Guidi-Trovato 2004): a branch starts only from a W and is able to survive if this W is copied before extinction, if many copies are produced in the early stage then the branch life will be longer, otherwise it will die very soon.

If we consider the application of the same probabilistic model to the problem of extinction of surnames (Galton and Watson, 1875), which tries to explain the reasons for the short-lasting existence of their majority against long-term preservation of a small very diffused part (thus belonging to very extended and branched families), it is easily understood why asymmetry and very high percentage of decimation of the manuscript tradition can be indicated as causes of the phenomenon of dicotomy of the trees in ecdotics.

Other reasons can be identified if we move from real T to reconstructed T, in other words if we perform a bottom-up instead of a top-down analysis. In ecdotics the significant errors (conjunctive and separative) for the reconstruction of T are those that for their very nature are more evident and therefore are even more liable to reparatory interventions, by conjecture or by contamination. It should be added that with high prolificity the likelihood of contamination increases and therefore the ramifications that have reached us are strongly contaminated.

Correction of the errors can contribute to reduce the branches of T (Timpanaro, 1985): given a tripartite branching (A,B,C) deriving from LF which contains e1 and e2, if A corrects e2 by conjecture or by horizontal transmission from a W belonging to another branch (thus deriving from LF2) or from O, then BC will result shared by e2, against A, so that the operator will think he is dealing with a bipartite branching ((B,C),A). Likewise, given the same type of branch (A,B,C), if A introduce an error of B (e3) by horizontal transmission, then AB will result to have e3 in common, against C, thus leading once again to a bipartite T.

Horizontal transmission produces on the textual T a reduction of branching similar to the one determined in the phylogenetic T by hybridizations and horizontal transfers (transition of information from one organism to another): if, for example, a population has a mixed origin, its position in the T will be very near to its genetically closest population and the mixture will not be well represented. In short, a phylogenetic transmission with hybridations does not adapt to the model of a T (Cavalli-Sforza *et al.*, 1994). Likewise, in the presence of contamination, textual tradition cannot be represented by a T but it will be necessary to use a reticular graph. The more reticular is the structure, the more simplified its projection onto the T.

It should be emphasized that, unlike what occurs in ecdotics, the dichotomy of the phylogenetic trees is intrinsic in the object of study and therefore the analogy recorded by the scholars is only epiphenomenally referred to the substance of things (Reeve, 1998).

3. Some applications in textual criticism

Analysis methods inspired to information theory have been frequently used in philological surveying, with particular regard to language evolution, but also textual phylogenetics and text attribution (Benedetto *et al.*, 2002 e 2003; Bennett *et al.*, 2003).

The procedure we have elaborated consists in the constant recource to the digitized sources, in exploiting data-compression techniques to estimate distance matrices to be used in the successive organization in tree structures. In our applications we used the Fitch-Margoliash and the Neighbour-Joining methods of the package PhylIP (Phylogeny Inference Package) which basically construct a tree by minimizing the net disagreement between the matrix pairwise distances and the distances measured on the tree. At present the method does not exclude, but integrates traditional philological analysis. We have conducted experiments on two different traditions:

a. a modern tradition composed of 33 Chain Letters, already submitted to an analysis similar to ours by Bennett, Li, Ma (2003) and studied in depth by Vanarsdale, 1998, from whom we have drawn information relative to the dates of the witnesses (figg. 3-4);

b. an ancient tradition concerning the witnesses of the *Vita Nova* by Dante Alighieri (fig. 5), a work whose stemmatic structure had already been sufficiently clarified by Barbi (1907 e 1932) and Gorni (1996).

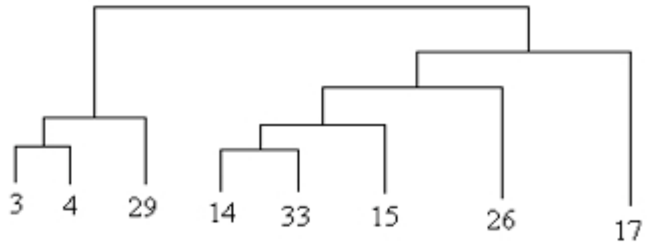


Figure 3. Partial tree, using the philological method of the 33 Chain Letters. Three variants are common to group ((3,4), 29), which has 5 variants with (14, 15, 17, 26, 33) that two other common variants identify as separate group, which is presumably in the upper part of the tree. Within this group one variant could identify a subgroup (15,17,33) and another one a group (15,33): in such case we would have ((15,33), 17), but the two variants alone do not seem to have the necessary weight to justify this hypothesis. The two variants grouping (14,15,26,33) and (14,15,33) respectively have instead more weight. Another common variant joins (14,33) (through 13, by contamination or phylogenetic intervention). Given this situation, in large part confirmed by the authomatics tree, the concordance according to the witnesses of this group with others of another group indicates possible contamination, correction or polygenetic variation, phenomena that can be hypothesized for all witnesses. This is the group resulting to be the most ancient from the date of the texts (from the end of the 60's to the early 80's).

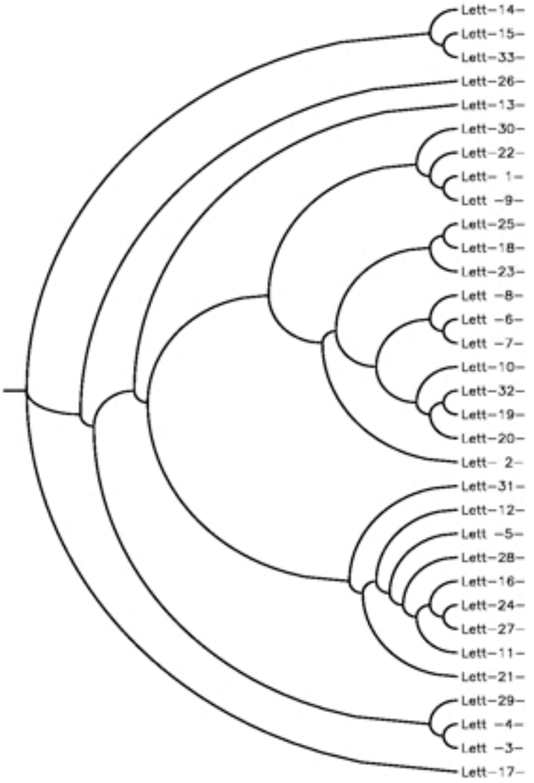


Figure 4. Phylogenetic tree of the 33 Chain Letters already analyzed by Bennett *et al*., (2003). The tree is obtained using the Neighbor-Joining method applied to a distance matrix whose elements are computed in terms of the relative entropy between pairs of texts. Details are reported in (Benedetto et al. 2002 and Baronchelli et al. 2005). Though the tree is unrooted we have chosen the outgroop as the letter 17 since the philological analysis gave us the indication that the letter 17 is the oldest one. Our tree is perfectly consistent with the partial one described in Fig.3 obtained with the standard philological method. The results are also consistent with the ones obtained by Bennett et al. (2003) to which we refer for a detailed description of the corpus.

```
        +---------------------------To
     +--3
     |  |  +------------------------T
     |  +--5
  +--1     +------------------------K
  |  |
  |  |  +---------------------------V
  |  +--2
  |     +---------------------------S
  |
--4---------------------------------M
```

Figure 5. Phylogenetic tree of Dante's *Vita Nova*. The tradition is formed by forty manuscripts, some of which are fragmentary. Among these, the manuscripts representing the entire tradition in the upper levels of the branches of Barbi's genealogical trees and that are used by the editors for the reconstruction of the text are eight: mss. K, T, To, K2, S, V, M, O. Barbi's tree can be summarized as follows: (((K,T),(To,K2)),((S,V),(M,O))). For this study we used the electronic edition of seven of these manuscripts and the same data-compression approach used for the experiments of the chain letters (we were unable to find any digitized version of K2) edited by S. Albonico.

In both cases a substantial convergence has been evidenced between the traditional method and the compression method. The data-compression techniques and the phylogenetic methods on the whole allow to group correctly the different families of manuscripts, and therefore they represent a precious tool for ecdotic investigation: in fact they becomes essential when one is faced with a very broad tradition, the traditional analysis of which would be far too complex and long and thus unacceptable (see Robinson and O' Hara, 1996; Baret *et al*., 2003; Macé *et al*., 2003; Mooney *et al*., 2003; Spencer *et al*., 2002; Lantin *et al*., 2004). The methods experimented so far have proven to be efficient and in particular rapid tools of classification, for a useful taxonomic organization of the witnesses of a text. They are in fact able to indicate the distance between two texts even if this distance is not necessarily a genealogical indication. The final aim of our preliminary work is the creation of an expert system based on the theory of information and able to identify the errors, completely independent with respect to the traditional philological methods.

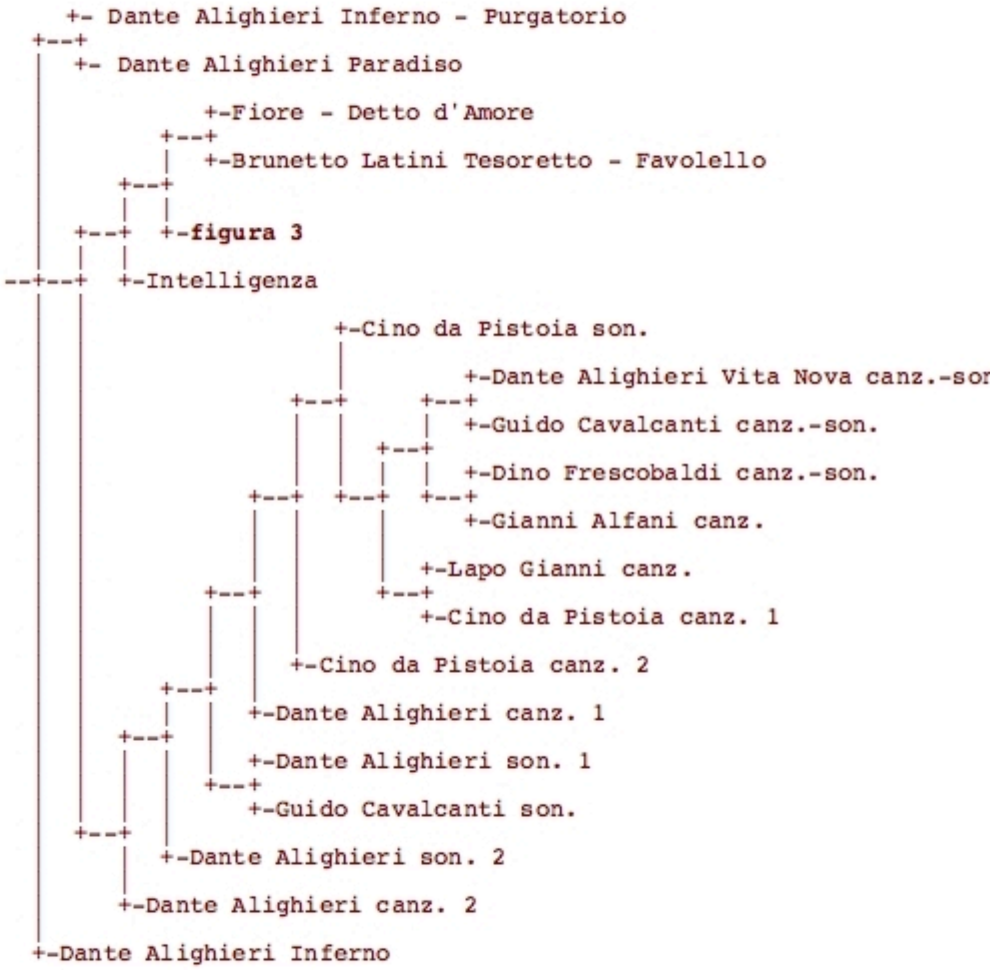
4. Poetry schools, attribution and intertextuality

Within the frame of attribution there is generally a distinction between 'external' and 'internal' criteria (Contini, 1984). The former concern the witnesses of other authors, as well as historical, cultural or biographical references in the text, and analysis of the sources. On the other hand, the 'internal' criteria are those that the philologist can draw from direct analysis of the language, of the metre or style of a text. The latter criteria are those followed by stylometry, which is the quantitative and statistical analysis of the literary style. For this type of analysis to be applied profitably it is necessary that the texts compared are sufficiently long and that they can be compared from the point of view of genre, language and contents. Furthermore, the stylistic aspects under consideration will have to respond to the requisites specified by Bailey in 1979, i.e. they will need to be outstanding, structural, frequent, easy-to-quantify and relatively free from the conscious control of the author. The methodological proposals that have so far followed one another vary essentially in the choice of the characteristics to be used as units of measure of the style: they vary from average length of words (Mendenhall 1887; Brinegar 1963; Mosteller, Wallace 1964), to average number of syllables per word and frequency of monosyllables (Fucks 1952; Fucks, Lauter 1965; Bruno, 1974; Brainerd, 1974), to length of sentences (Yule, 1938; Williams, 1940; Wake, 1957; Morton, 1965; Sichel, 1974; Kjetsaa, 1979). The stylistic traits which have been considered pertinent are the percentages of the different parts of discourse (nouns, verbs, adjectives, etc.) used in a text (Somers, 1966; Antosch, 1969; Brainerd, 1973 and 1974), the frequency of the so-called 'function words', i.e. the words that are not determined by the context (Ellegard, 1962), as well as the preference assigned by the author to either of the elements of a pair of synonyms (Mosteller, Wallace, 1964; Morton, 1978; Burrows, 1987). Up until now, stylistic investigations have generally been of the lexical type: statistical models have been developed to calculate the richness of an author's vocabulary, the average distance in which new words are generated in a text, the percentage of presence of *hapax legomena* and *hapax dislegomena* (Holmes, 1994). Finally, great emphasis has been given in France to the controversy aroused by the method of measurement of 'intertextual distance' proposed by Labb⊠ and Labb⊠ (2001). The two scholars performed complete lemmatization of the texts to be examined and then analyzed the distance between the dictionaries obtained: the shorter the distance, the greater the likelihood that the two texts could be assigned to the same author, or belonged to the same literary genre, or had been written in the same period or were dealing with the same subject.

The classification power of the method we have adopted can be verified on corpora related to the works of a specific period. The system has proven efficient for the classification of texts according to hierarchical criteria: in particular, it can bring closer texts belonging to the same author, associate different authors according to the poetry school, genre or socio-cultural *milieu* they belong to. The method has been applied to the texts of 13th century Italian poetry, also including in the corpus Dante Alighieri's *Divina Commedia*.

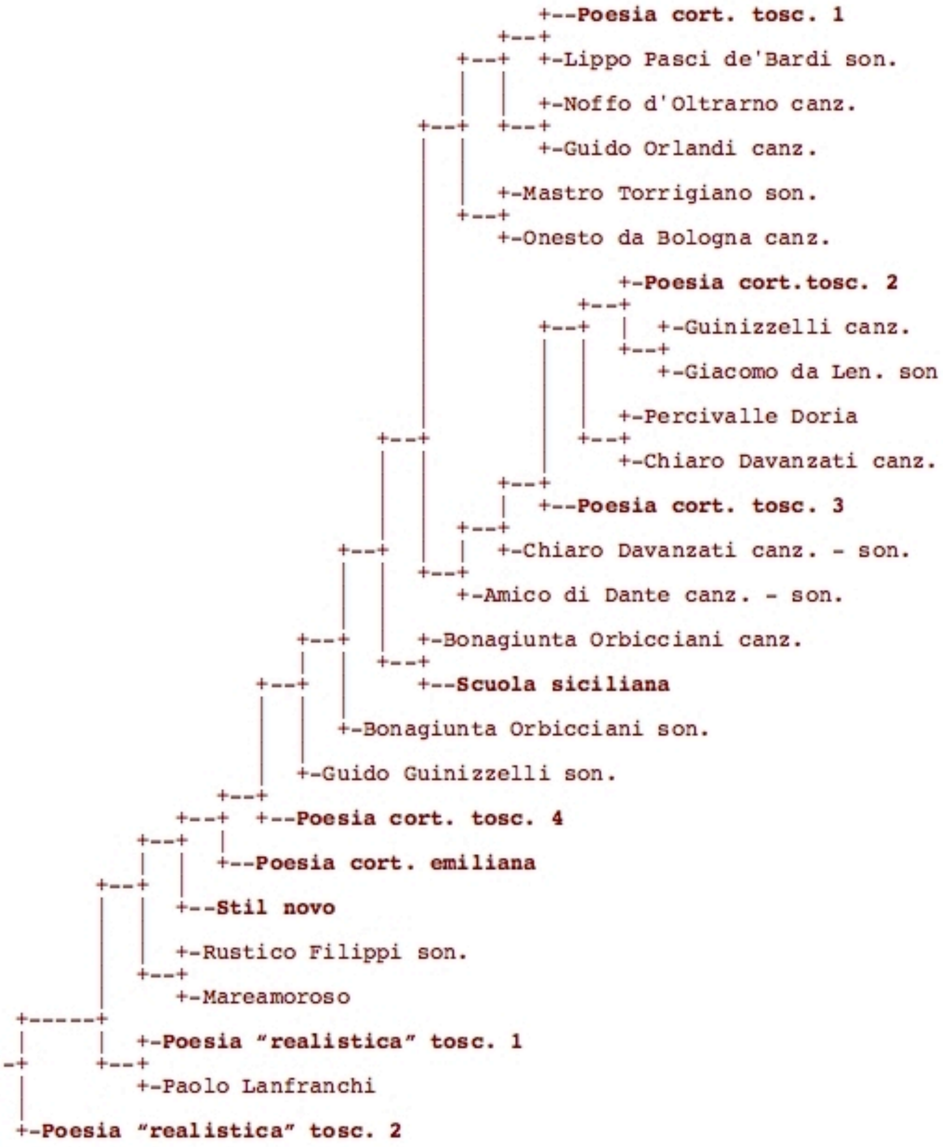
All the editors' interventions were removed from the corpus of texts that we used, and graphical rendering of metric and phonetic phenomena were standardized, respecting however the linguistic identity of the documents. All divisions, modern numbering of the texts, punctuation, integrations and emendations marks made by the critical editors have been eliminated. In order to facilitate the identification of equivalent lemmas, it was decided that phonosyntactic gemination should not be represented graphically, and that whenever possible synalepha was to replace the phenomena of apocope, apheresis, or elision not envisioned by contemporary Italian. All the electronic texts were subdivided automatically into files 10,000 characters long (i.e. a size of 10Kb), in order to have several texts for each work to be used for consistency and robustness checks. Texts shorter than 10Kb were left untouched. With the corpus of electronic ready we performed several experiments aimed both at the authorship attribution and to the construction of a classification tree of authors and schools. Texts were analyzed with the data-compression technique introduced in Benedetto *et al*. (2002) which allows to construct a distance matrix, i.e. a symmetrical matrix whose elements represent the distance between a pair of texts (for further details concerning the notion of distance and its definition we refer to Baronchelli *et al*., 2005). The distance matrix is then used in the framework of well-established phylogenetic methods to construct trees. Here we only present a synthesis of the results and we invite the interested reader to contact the authors for a more complete picture.

The groupings represented by the trees we obtained are generally consistent with those traditionally indicated in literary compendia (figg. 2-3). The program recognizes the poetry schools very clearly, and brings together the works of great as well as those of minor and less productive authors. The tree identifies the Sicilian School, the Dolce Stil Novo and the Tuscan courtly poetry. In the group represented by the poets of the Dolce Stil Novo, all the poems of the *Vita Nova* are connected by the same node, which is linked to that of Guido Cavalcanti's works (fig. 3). The *Commedia* is collocated in a different branch separated from the lyrical production (fig. 2).



```
     +- Dante Alighieri Inferno - Purgatorio
  +--+
  |  +- Dante Alighieri Paradiso
  |
  |           +-Fiore - Detto d'Amore
  |        +--+
  |        |  +-Brunetto Latini Tesoretto - Favolello
  |     +--+
  |     |  |
  |  +--+  +-figura 3
  |  |  |
--+--+  +-Intelligenza
  |  |
  |  |                 +-Cino da Pistoia son.
  |  |                 |
  |  |                 |        +-Dante Alighieri Vita Nova canz.-sor
  |  |              +--+     +--+
  |  |              |  |     |  +-Guido Cavalcanti canz.-son.
  |  |              |  |  +--+
  |  |              |  |  |  |  +-Dino Frescobaldi canz.-son.
  |  |           +--+  +--+  +--+
  |  |           |  |     |     +-Gianni Alfani canz.
  |  |           |  |     |
  |  |           |  |     |  +-Lapo Gianni canz.
  |  |        +--+  |     +--+
  |  |        |  |  |        +-Cino da Pistoia canz. 1
  |  |        |  |  |
  |  |        |  |  +-Cino da Pistoia canz. 2
  |  |     +--+  |
  |  |     |  |  +-Dante Alighieri canz. 1
  |  |  +--+  |
  |  |  |  |  |  +-Dante Alighieri son. 1
  |  |  |  |  +--+
  |  |  |  |     +-Guido Cavalcanti son.
  |  +--+  |
  |     |  +-Dante Alighieri son. 2
  |     |
  |     +-Dante Alighieri canz. 2
  |
  +-Dante Alighieri Inferno
```

Figure 2. Summarized and simplified phylogenetic tree of the poetry works of 13th century (the length of the branches does not represent the distances): Dante Alighieri Inferno = 17 files; Dante Alighieri Purgatorio = 17 f.; Dante Alighieri Paradiso = 17 f.; Fiore = 12 f.; Detto d'Amore = 2 f.; Brunetto Latini Favolello = 1 f.; Brunetto Latini Tesoretto = 7 f.; Intelligenza = 10 f.; Cino da Pistoia son. = 1 f.; Dante Alighieri Vita Nova canz. - son. = 4 f.; Guido Cavalcanti canz. - son. = 3 f.; Dino Frescobaldi canz. - son. = 2 f.; Gianni Alfani canz. = 1 f.; Lapo Gianni canz. = 2 f.; Cino da Pistoia canz. 1 = 2 f.; Cino da Pistoia canz. 2 = 2 f.; Dante Alighieri canz. 1 = 2 f.; Dante Alighieri son. 1 = 1 f.; Guido Cavalcanti son. = 1 f.; Dante Alighieri son. 2 = 1 f.; Dante Alighieri canz. 2 = 2 f.



```
                                        +--Poesia cort. tosc. 1
                                     +--+
                                  +--+  +-Lippo Pasci de'Bardi son.
                                  |  |
                                  |  |  +-Noffo d'Oltrarno canz.
                               +--+  +--+
                               |  |     +-Guido Orlandi canz.
                               |  |
                               |  |  +-Mastro Torrigiano son.
                               |  +--+
                               |     +-Onesto da Bologna canz.
                               |              +-Poesia cort.tosc. 2
                               |           +--+
                               |        +--+  |  +-Guinizzelli canz.
                               |        |  |  +--+
                               |        |  |     +-Giacomo da Len. son
                               |        |  |
                               |        |  |  +-Percivalle Doria
                            +--+        |  +--+
                            |  |        |     +-Chiaro Davanzati canz.
                            |  |     +--+
                            |  |     |  +--Poesia cort. tosc. 3
                            |  |  +--+
                         +--+  |  |  +-Chiaro Davanzati canz. - son.
                         |  |  +--+
                         |  |     +-Amico di Dante canz. - son.
                         |  |
                      +--+  |  +-Bonagiunta Orbicciani canz.
                      |  |  +--+
                   +--+  |     +--Scuola siciliana
                   |  |  |
                   |  |  +-Bonagiunta Orbicciani son.
                   |  |
                   |  +-Guido Guinizzelli son.
                +--+
             +--+  +--Poesia cort. tosc. 4
          +--+  |
          |  |  +--Poesia cort. emiliana
       +--+  |
       |  |  +--Stil novo
       |  |
       |  |  +-Rustico Filippi son.
       |  +--+
       |     +-Mareamoroso
 +-----+
 |     |  +-Poesia “realistica” tosc. 1
-+     +--+
 |        +-Paolo Lanfranchi
 |
 +-Poesia “realistica” tosc. 2
```

Figure 3. Detail of a branch of the phylogenetic tree illustrated on fig. 2: **Poesia cort. tosc. 1** = Panuccio dal Bagno canz. - son. (4 files) and Dante da Maiano canz. - son. (5 f.); **Poesia cort. tosc. 2** = Carnino Ghiberti, Bondie Dietaiuti, Neri de' Visdomini, Maestro Francesco, Inghilfredi da Lucca (2 f.), Bonagiunta Orbicciani canz. (2 f.), Pucciandone Martelli; **Poesia cort. tosc. 3** = Chiaro Davanzati canz. - son. (8 f.), Rustico Filippi, Noffo Bonaguidi, Guittone d'Arezzo son.; **Poesia cort. tosc. 4** = Guittone d'Arezzo canz. - son. (24 f.), Bacciarone di messer Baccone, Meo Abbracciavacca, Monte Andrea canz. - son. (11 f.), Jacopo da Leona; **Scuola siciliana** = Re Enzo, Stefano Protonotaro, Pier delle Vigne, Rinaldo d'Aquino, Federico, Ruggieri d'Amici, Giacomino Pugliese (2 f.), Ruggierone da Palermo, Tommaso di Sasso, Giacomo da Lentini canz. (3 f.), Mazzeo di Ricco, Jacopo Mostacci, Guido delle Colonne; **Poesia cort. emiliana** = Tommaso da Faenza and Onesto da Bologna; **Stil novo** = Guido Orlandi, Dino Compagni; **Poesia “realistica tosc.” 1** = Cecco Angiolieri son. (6 f.), Meo dei Tolomei (2 f.); **Poesia “realistica tosc.” 2** = Cenne da la Chitarra, Folgore da San Gimignano (2 f.), Muscia da Siena, Cecco Angiolieri son. dubbi (2 f.). Moreover: Chiaro Davanzati caz. - son. = 6 f.; Amico di Dante canz. - son. = 3 f.; Rustico Filippi son. (2 f.); Mareamoroso (2 f.); when the file number is not specified it means that there’s only 1 file.

Furthermore, the program provides precious suggestions for the attribution of anonymous works. In the case of works belonging to a known author, since the percentage of unidentification of files containing texts by the same author (procedure of identification “known author over known author”) is equal to 10%, this will be the percentage of unreliability for the cases in which anonymous works are drawn closer to works written by known authors. Therefore, the proximity of the *Mare Amoroso* to the production of Rustico Filippi is extremely interesting. *Mare amoroso* is the first example of poem in blank verse belonging to the Italian poetic tradition; the text, anonymous, is referred by one manuscript only, which also transmits the *Tesoretto* and the *Favolello* by Brunetto Latini. The first editor (Grion 1868) of the poem, just like its discoverer Trucchi (1846), ascribed the work to Brunetto Latini, essentially due to the presence in the codex of the two poems by ser Brunetto and to a number of recurring stylistic traits in the *Mare amoroso* and in the canzone by Brunetto *S'eo son distretto*. This hypothesis, supported by Bertoni (1901), was rejected by Gaspary (1882), Monaci (1912) and Cian (1901). Two more recent editors, Contini (1960) and Vuolo (1962), owing to the fact that the only witness could be dated back no later than to the beginning of the 14th century, agree on assigning the text to the end of the 13th century, without expressing themselves on the name of the author. The Florentine linguistic traits were confirmed by Gorni (1986), who supported the hypothesis of the Florentine provenance with some metrical comments about the origins of blank verse; finally De Laude (1993) again brought up the name of Brunetto Latini after careful analysis of the sources common to *Mare amoroso*, *Tresor* and *Rettorica*. Although unable to make any authorship proposal, we wish to underline that none of the elements so far ascertained, i.e. 13th century dating and Florentine traits of the author, hinder the vicinity that our tree establishes between the anonymous poem and Rustico Filippi, Florentine author of rhymes to whom Brunetto Latini dedicated the *Favolello*.

The most important result is represented by the position in the tree of the poem *Il Fiore*: it is strictly related to the *Detto d'Amore*, its homologue transmitted by the same manuscript, and linked to the node from which the work of Brunetto Latini also derives, while it is considerably distant from the production of Dante Alighieri. *Il Fiore*, a total of 232 sonnets representing in concise form the narrative part of the *Roman de la Rose*, had been attributed to Dante by its first editor Castets (1881) and this attribution, for a long time strongly contested, has then become largely approved thanks to the intervention of Contini who, starting from the contribution of *La questione del Fiore* (1965), up to the edition of the poem (1984), strongly supported Dante's authorship. Contini was to inaugurate a new way of facing the problem of attribution, not based on external criteria, but on the stylistic analysis of the text and on a close comparison between the style of the author of the *Fiore* and that of the works certainly belonging to Dante. The quantity and quality of the matches induced Contini to consider them not as "una semplice somma di indizi, ma […] un organismo mnemonico […] del tutto assimilabile alla memoria che il Dante della *Commedia* ha di se stesso" (Contini, 1976). After years of almost unanimous consent to Contini's thesis, scientific literature has recently raised again the problem of the poem's authorship; in particular, Fasani (1998) reconsidered the hypothesis of ascribing the poem to Brunetto (already in Muner 1968-69 and 1970-71).

The association deriving from our method thus confirms the hypothesis of attribution by that part of scientific literature that assigns *Il Fiore* to Brunetto Latini. Furthermore, since *Il Fiore* is considered a translation of the *Roman de la Rose*, written in ancient French, Brunetto Latini would seem to be a valid candidate, since his most important work, the *Tresor*, is written in this language.

The reason for the good performance of the assignment method of the texts is of a linguistic type, in particular as regards lexicon and morphology. In the majority of cases the dictionaries of strings common to two texts, automatically produced by the software, indicate that the shared graphical sequences are linguistically meaningful: very often, entire lemmas (even in sequence) and especially suffixes and prefixes are involved. An approximate calculation indicates that the segments linguistically insignificant amount to just over 10%. Therefore, the single authors would seem to repeat the same words in uniform manner and to use similar percentages of occurrences of the same grammatical elements.

5. Conclusions

The trees thus allow to organize in taxonomic form the material available, generally confirming the main achievements of scientific literature, but also adding important concise information about the collocation of certain works. The method, from the identification of common strings upwards, can obviously also be applied in the field of intertextuality, making it possible for a single procedure to be applied to the most important fields of philology.

Further developments can be envisioned: on the one hand, as already said, the modes of analysis and classification of the manuscript sources will be improved, using the Lachmann method, on the other hand general information about diachronic distribution of the branches and nodes will be provided, eventually producing a single tree that includes both manuscript tradition and works. Other traditions could also be investigated, ranging from the many sectors of lyric poetry to the many forms of novel or stage prose, to the many anonymous or practical texts to be organised in a general taxonomy. Finally, other disciplines will be able to benefit from the same method, for example all historical and legal disciplines, and in general all those fields in which taxonomic study of the sources and acknowledgement of authorship of the documents are to be considered of primary importance.

References

- R. Antonelli, *Interpretazione e critica del testo*, in *Letteratura italiana*, vol. 4, *L'interpretazione*, A. Asor Rosa (eds.), Torino 1985, pp. 141-243.
- F. Antosch, *The diagnosis of Literary Style with the Verb-Adjective Ratio*, in *Statistics and Style*, L. Dolezel and R. W. Bailey (eds.), New York 1969.
- D'A. S. Avalle, *Principî di critica testuale*, Padova 1972.
- R. W. Bailey, *Authorship Attribution in Forensic Setting*, in *Advances in Computer-aided Literary and Linguistic Research*, D. E. Ager, F. E. Knowles, J. Smith (eds.), Birmingham 1979.
- *La Vita Nuova di Dante Alighieri*, M. Barbi (eds.), Firenze 1932.
- *La Vita Nuova*, M. Barbi (eds.), Milano 1907.
- A. C. Barbrook, N. Blake, P. Robinson, *The phylogeny of the Canterbury Tales*, in «Nature» 394 (1998), p. 839.
- P. Baret, M. Debuisson, A. C. Lantin, C. Macé, *Experimental phylogenetic analysis of a Greek manuscript tradition*, in «Journal of the Washington Academy of Sciences» 89 (2003), pp. 117-124.
- A. Baronchelli, E. Caglioti, V. Loreto, *Artificial sequences and complexity measures*, in «Journal of Statistical Mechanics» (2005), 04002.
- R. Bateman, I. Goddard, R. T. O'Grady, V. A. Funk, R. Mooi, W. J. Kress, P. Cannell, *Speaking of forked tongues: the feasibility of reconciling human phylogeny and the history of language*, in «Current Anthropology» 31 (1990), pp. 1–24.
- J. Bédier, *La Tradition manuscrite du «Lai de l'Ombre». Réflexions sur l'art d'éditer les anciens textes*, in «Romania» 54 (1928), pp. 161-196, 321-358 (later on booklet, Paris 1929).
- M.L. Bender, *Genetic classification of languages: genotype vs. phenotype*, in «Language Sciences» 43 (1976), pp. 4–6.
- D. Benedetto, E. Caglioti, V. Loreto, *Language Trees and Zipping*, in «Physical Review Letters» 88 (2002), 048702.
- D. Benedetto, E. Caglioti, V. Loreto, *Zipping Out Relevant Information*, in «Computing in Science & Engineering», Jan./Febb. 2003, pp. 80-85.
- C. H. Bennett, M. Li, B. Ma, *Chain Letters and Evolutionary Histories*, in «Scientific American» June 2003, pp. 76-81
- G. Bertoni, *Il "mare amoroso"*, in «Fanfulla della domenica» 28 (1901).
- B. Bollobás, *Modern graph theory*, New York 1998.
- B. Brainerd, *On the Distinction Between a Novel and a Romance: A Discriminant Analysis*, in «Computers and Humanities» 7 (1973), 259-270.
- B. Brainerd, *Weighing Evidence in Language and Literature: A Statistical Approach*, Toronto 1974.
- F. Brambilla Ageno, *Ci fu sempre un archetipo?*, in «Lettere italiane» 27 (1975), pp. 308-309.
- C. S. Brinegar, *Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship*, in «Journal of American Statistical Association» 58 (1963), pp. 85-96.
- A. M. Bruno, *Toward a Quantitative Methodology for Stilistic Analyses*, Berkeley 1974.
- J. F. Burrows, *Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style*, in «Journal of the Association for Literary and Linguistic Computing» 2 (1987), pp. 61-70.
- H. D. Cameron, *The upside-down cladogram. Problems in manuscript affiliation*, in *Biological Metaphor and Cladistic Classification. An Interdisciplinary Perspective*, H. M. Hoenigswald (eds.), L. F. Wiener, Philadelphia 1987, pp. 227-242.
- P. Canettieri, V. Loreto, M. Rovetta, G. Santini, *Higher criticism and Information Theory*, in «Rivista di Filologia Cogitiva», http://w3.uniroma1.it/cogfil/ecdotica.html, december 2005.
- A. Castellani, *Bédier avait-il raison? La méthode de Lachmann dans les éditions de textes du Moyen Age* (1957), it. tr. on *Saggi di linguistica italiana e romanza (1946-1976)*, III, Salerno-Roma 1980, pp. 161-200.

- *"Il Fiore", poème italien du XIII^{e} siècle, en CCXXXII sonnets, imité du "Roman de la Rose", par Durante*, F. Castets (eds.), Montpellier 1881.
- L. L. Cavalli Sforza and A. W. Edwards, *Phylogenetic analysis: models and estimation procedures*, in «Evolution» 32 (1967), pp. 550-570 and «American Journal of Human Genetics», 19 (1967), pp. 233-257.
- L. L. Cavalli Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes*, Princeton 1994.
- V. Cian, *Varietà dugentistiche, una probabile parodia letteraria e un saggio di precettistica matrimoniale*, Pisa 1901.
- M. Ciccuto, *Il restauro de "L'Intelligenza" e altri studi dugenteschi*, Pisa 1985.
- G. Contini, *Fiore*, in *Enciclopedia Dantesca*, II, Roma 1970, pp. 895-901.
- G. Contini, *Il Fiore e il Detto d'amore attribuibili a Dante Alighieri*, Milano, 1985.
- G. Contini, *La questione del Fiore*, in «Cultura e scuola» 13-14 (1965), pp. 768-773
- *Poeti del duecento*, G. Contini (eds.), Milano-Napoli 1960.
- T. Cover, J. Thomas, *Elements of Information Theory*, New York 1991.
- S. De Laude, *Per l'attribuzione del "Mare amoroso"*, in *L'attribuzione: teoria e pratica - storia dell'arte, musicologia, letteratura*, Atti del Seminario di Ascona (30 sett. - 5 ott. 1992), O. Besomi and C. Caruso (eds.), Boston-Berlin 1993, pp. 211-223.
- W. B. Douglas, *Introduction to graph theory*, Upper Saddle River 2001.
- A. W. F. Edwards, L. L. Cavalli-Sforza, *Reconstruction of evolutionary trees*, in *Phenetic and Phylogenetic Classification*, V. E. Heywood and J. McNeill (eds.), London 1964, pp. 67-76.
- A. Ellegard, *A Statistical Method for Determining Authorship: The Junius Letters 1769-1772*, in «Gothenburg Studies in English» 13 (1962), pp. 1-115.
- J. Felsenstein, *Distance methods for inferring phylogenies: a justification*, in «Evolution» 38 (1984), pp. 16-24.
- R. Fasani, *Il Fiore e Brunetto Latini*, in «Studi e problemi di critica testuale» 57 (1998), pp. 57-71.
- W. M. Fitch and E. Margoliash, *The construction of phylogenetic trees - a generally applicable method utilizing estimates of the mutation distance obtained from cytochrome c sequences*, in «Science», 155 (1967), pp. 279-284.
- C. Flight, *Bantu trees and some wider ramifications*, in «African Languages and Cultures» 1 (1988), pp. 25–43.
- J. Fourquet, *Le paradoxe de Bédier*, in *Mélanges 1945*, II, *Études littéraires*, Paris 1946, pp. 1-16.
- J. Froger, *La critique des textes et son automatisation*, Paris 1968.
- W. Fucks, *On the Mathematical Analysis of Style*, in «Biometrika» 39 (1952), pp. 122-129.
- W. Fucks, J. Lauter, *Mathematische Analyse des Literarischen Stils*, in *Mathematik und Dichtung*, H. Kreuzer and R. Gunzenhäuser (eds.), München 1965.
- F. Galton, H. W. Watson, *On the probability of the extinction of families*, in «Journal of the Anthropological Society of London» 4 (1875), pp. 138-144.
- A. Gaspary, *La scuola poetica siciliana del secolo XII*, Livorno 1882.
- *Dante Alighieri. Vita Nova*, G. Gorni (eds.), Torino 1996.
- G. Gorni, *Le gloriose pompe (e i fieri ludi) della filologia italiana, oggi*, in «Rivista di letteratura italiana» 4 (1986), pp. 391-412.
- R. D. Gray, F. M. Jordan, *Language trees support the express-train sequence of Austronesian expansion*, in «Nature» 405 (2000), pp. 1052-1055.
- J. H. Greenberg, *Language and evolutionary theory*, in *Essays in Linguistics*, Chicago 1957, pp. 56–65.
- W. W. Greg, *Recent theories of textual criticism*, in «Modern Philology» 28 (1931), pp. 401-4.

- J. Grier, *Lachmann, Bédier and the Bipartite Stemma: Towards a Responsible Application of the Common-Error Method*, in «Revue d'Histoire des Textes» 18 (1988), pp. 263-278.
- G. Grion, *Il Mare amoroso, poemetto in endecasillabi sciolti di Brunetto Latini*, in «Il Propugnatore» 1 (1868), pp. 593-620, 2 (1869), pp. 147-179 and 273-306.
- V. Guidi and P. Trovato, *Sugli stemmi bipartiti. Decimazione, asimmetria e calcolo delle probabilita'*. 1. P. Trovato, *Dagli alberi reali agli stemmi*; 2. V. Guidi, *Manuscript traditions and stemmata: a probabilistic approach*, in «Filologia italiana» 1 (2004), pp. 9-48.
- J. B. Hall, *Why are the stemmata of so many manuscripts traditions bipartites?*, in «Liverpool Classical Monthly» 17 (1992), pp. 31-32.
- H. M. Hoenigswald, *Language families and subgroupings, tree model and wave theory, and reconstruction of protolanguages*, in *Research Guide on Language Change*, E. C. Polome (eds.), Berlin-New York 1990, pp. 441-454.
- *Biological Metaphor and Cladistic Classification. An Interdisciplinary Perspective*, H. M. Hoenigswald and L. F. Wiener (eds.), Philadelphia 1987.
- D. I. Holmes, *Authorship Attribution*, in «Computers and the Humanities» 28 (1994), pp. 87-106.
- C. J. Howe, A. C. Barbrook, L. R. Mooney, *Parallel Problems Encountered during the Construction of Stemma or Phylogenetic Reconstruction*, in *Studies in Stemmatology*, II, *Kinds of Variants*, P. van Reenen, A. den Hollander, M. van Mulken (eds.), Amsterdam 2004, pp. 3-11.
- C. J. Howe, A. C. Barbrook, M. Spencer, P. Robinson, B. Bordalejo, L. R. Mooney, *Manuscript Evolution*, in «Trends in Genetics» 17 (2001), pp. 147-152.
- A. Kaltchenko, *Algorithms for Estimating Distance with Application to Bioinformatics and Linguistics*, (2004) http://xxx.arxiv.cornell.edu/abs/cs.CC/0404039.
- A. I. Khinchin, *Mathematical Foundations of Information Theory*, New York 1957.
- G. Kjetsaa, *And Quiet Flows the Don Through the Computer*, in «Association for Literary and Linguistic Computing Bulletin» 7 (1979), pp. 248-256.
- A. Kleinlogel, *Das Stemmaproblem*, in «Philologus» 102 (1968), pp. 63-82.
- *Linguistics and Evolutionary Theory: Three Essays by August Schleicher, Ernst Haeckel, and William Bleek*, E. F. K. Koerner (eds.), Amsterdam 1983.
- E. F. K. Koerner, *Schleichers Einfluß auf Haeckel: Schlaglichter auf die wechselseitige Abhangigkeit zwischen linguistichen und biologischen Theorien in 19. Jahrhundert*, in «Zeitschrift für vergleichende Sprachforschung» 95 (1981), pp. 1–21.
- C. Labbé and D. Labbé, *Inter-Textual Distance and Authorship Attribution. Corneille and Molière*, in «Journal of Quantitative Linguistics» 8-3 (2001), pp. 213-231.
- A.-C. Lantin, P. Baret, C. Macé, *Phylogenetic analysis of Gregory of Nazianzus' Homily 27*, in *Les poids des mots*, Actes des 7èmes Journées Internationales d'Analyse statistique des Donnés Textuelles (JADT04), G. Purnelle, C. Fairon, A. Dister (eds.), Louvain-la-Neuve, vol. 2, 2004, pp. 700-707.
- A. Lempel and J. Ziv, *A Universal Algorithm for Sequential Data Compression*, in «IEEE Transactions on Information Theory» May 1977, pp. 337-343.
- M. Li, X. Chen, X. Li, B. Ma, P. M. B. Vitanyi, *The similarity metric*, in «IEEE Transactions on Information Theory» 50 (2004), pp. 32-50.
- P. Maas, *Textkritik* (1927), 3d ed. Leipzig 1957.
- C. Macé, T. Schmidt, J.-F. Weiler, *Le classement des manuscrits par la statistique et la phylogénétique: le cas de Grégoire de Nazianze et de Basile le Minime*, in «Revue d'Histoire des Textes» 31 (2003), pp. 241-273.
- J. P. Maher, *More on the history of the comparative method: the tradition of Darwinism in August Schleicher's work*, in «Anthropological Linguistics» 8 (1966), pp. 1–12.
- T.C. Mendenhall, *The Characteristic Curves of Composition*, in «Science» 9 (1887), pp. 237-249.
- N. Merhav and J. Ziv, *Universal Schemes for Sequential Decision from Individual Data Sequences*, in «IEEE Transactions on Information Theory» 39 (1993), pp. 1280-1292.

- E. Monaci, *Crestomazia italiana dei primi secoli con prospetto grammaticale e glossario*, Città di Castello 1912.
- E. Montanari, *La critica del testo secondo Paul Maas. Testo e commento*, Tavarnuzze (Firenze) 2003.
- L. R. Mooney, A. C. Barbrook, C. J. Howe, M. Spencer, *Stemmatic Analysis of Lydgate's 'Kings of England': A Test Case for the Application of Software Developed for Evolutionary Biology to Manuscript Stemmatics*, in «Revue d'Histoire des Textes» 31 (2003), pp. 202-240.
- A. Q. Morton, *Literary Detection*, New York 1978.
- A. Q. Morton, *The Authorship of Greek Prose*, in «Journal of the Royal Statistical Society» A 128 (1965), pp. 169-233.
- F. Mosteller, D. Wallace, *Inference and Disputed Authorship: The Federalist*, Massachusetts 1964.
- M. Muner, *La paternità brunettiana del Fiore e del Detto d’amore*, in «Motivi per la difesa della cultura» 9 (1970-71).
- R. O’Hara, *Trees of History in systematics and philology*, in «Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano» 27/1 (1996), pp. 81-88.
- R. J. O'Hara and P. Robinson, *Computer-assisted methods of stemmatic analysis*, in *The Canterbury Tales Project Occasional Papers*, I, 1993, pp. 53-74.
- H. H. Out and K. Sayood, *A new sequence distance measure for phylogenetic tree construction*, in «Bioinformatics» 19 (2003), pp. 2122-2130.
- A. F. Ozanam, *Documents inédits pour servir a l’histoire littéraire de l’Italie depuis le VIII siècle jusqu’au XIII avec des recherches sur le moyen âge italien*, Paris 1850.
- E. Picardi, *Some problems of classification in linguistics and biology, 1800–1830*, in «Historiographia Linguistica» 4 (1977), pp. 31–57.
- N. I. Platnick and H. D. Cameron, *Cladistic methods in textual, linguistic and phylogenetic analysis*, in «Systematic Zoology» 26 (1977), pp. 380-385.
- H. Quentin, *Essais de critique textuelle (Ecdotique)*, Paris 1926.
- D. M. Raup, *Extinction. Bad Genes or Bad Luck*, New York 1991.
- M. D. Reeve, *Shared innovations, dichotomies, and evolution*, in *Filologia classica e filologia romanza: esperienze ecdotiche a confronto*, A. Ferrari (eds.), Spoleto 1998, pp. 445-505.
- M. D. Reeve, *Stemmatic Method: ‘qualcosa che non funziona?’*, in *The Role of the Book in Medieval Culture*, Proceedings of the Oxford International Symposium (26 September-1October 1982), P. Ganz (eds.), Turnhout 1986, I, pp. 57-70.
- P. M. W. Robinson and R. J. O‘ Hara, *Cladistic Analysis of an Old Norse Manuscript Tradition*, in *Research in Humanities Computing 4*, S. Hockey, N. Ide (eds.), Oxford 1996, pp. 115-137.
- D. B. Searle, *Trees of life and of language*, in «Nature» 426 (nov. 2003), pp. 381-382.
- C. E. Shannon, *A Mathematical Theory of Communication: Part II, The Discrete Channel with Noise*, in «The Bell System Technical Journal» 27 (1948), pp. 623-656.
- W. P. Shepard, *Recent theory of textual criticism*, in «Modern Philology» 26-27 (1930), pp. 129-141.
- *Reconstructing Languages and Cultures*, V. Shevoroshkin (eds.), Berlin 1989.
- V. Shevoroshkin, J. Woodford, *Where linguistics, archeology, and biology meet*, in *Ways of Knowing*, Brockman J. (eds.), New York 1991, pp. 173–197.
- H. S. Sichel, *On a distribution representing sentence-length in written prose*, in «Journal of the Royal Statistical Society» A 137 (1974), pp. 25-34.
- H. H. Somers, *Analyse statistique du style*, Paris 1966.
- M. Spencer, B. Bordalejo, L.-S. Wang, A. C. Barbrook, L. R. Mooney, P. Robinson, T. Warnow, C. J. Howe, *Analyzing the Order of Items in Manuscripts of The Canterbury Tales*, in «Computers and the Humanities» 37 (2003), pp. 97-109.
- M. Spencer, C. Howe, *Estimating distances betweeen manuscripts based on copying errors*, in «Literary and Linguistic Computing» 16 (2001), pp. 467-484.

- M. Spencer, C. J.Howe, *How Accurate Were Scribes? A Mathematical Model*, in «Literary and Linguistic Computing» 17 (2002), pp. 311-322.
- M. Spencer, L. R. Mooney, A. C. Barbrook, B. Bordalejo, C. J. Howe, P. Robinson, *The Effects of Weighting Kinds of Variants*, in *Studies in Stemmatology*, II, *Kinds of Variants*, P. van Reenen, A. den Hollander, M. van Mulken (eds.), Amsterdam 2004, pp. 225-237.
- M. Spencer, K. Wachtel, C. J. Howe, *The Greek Vorlage of the Syra Harclensis: A Comparative Study on Method in Exploring Textual Genealogy*, in «TC: a journal of biblical textual criticism» 7 (2002), http://rosetta.reltech.org/TC/vol07/SWH2002/.
- R. D. Stevick, *The biological model and historical linguistics*, in «Language», 39 (1963), pp. 159–169.
- S. Timpanaro, *La genesi del metodo del Lachmann*, Padova 1985.
- F. Trucchi, *Poesie inedite di dugento autori dall'origine della lingua infino al secolo decimosettimo*, Prato 1846.
- D. W. Vanarsdale, *Chain Letters Evolution*, http://www.silcom.com/~barnowl/chain-letter/evolution.html
- E. Vuolo, *Il Mare amoroso*, Roma 1962.
- W. C. Wake, *Sentence length distributions of Greek authors*, in «Journal of the Royal Statistical Society» A 120 (1957), pp. 331-346.
- M. Weitzmann, *Computer simulation of the development of manuscript traditions*, in «Bulletin of the Association for Literary and Linguistic Computing» 10 (1982), pp. 55-59.
- M. Weitzmann, *The Analisis of Open Traditions*, in «Studies in Bibliography» 38 (1985), pp. 82-120.
- M. Weitzmann, *The evolution of the manuscript tradition*, in «Journal of the Royal Statistical Society» 150 (1987), pp. 287-308.
- F. Whitehead, C. E. Pickford, *The two-branch Stemma*, in «Bulletin bibliographique de la Société Internationale Arthurienne» 3 (1951), pp. 83-90.
- C. B. Williams, *A note on the statistical analysis of sentence-length as a criterion of literary style*, in «Biometrika» 31 (1940), pp. 356-361.
- A. D. Wyner, J. Ziv, *The Sliding-Window Lempel-Ziv Algorithm is Asymptotically Optimal*, in «Proceedings IEEE» 82 (1994), pp. 872-877.
- G. U. Yule, *On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authority*, in «Biometrika» 30 (1938), pp. 363-390.
- *Complexity, Enthropy, and Physics of Information* (1964), W. H. Zurek (eds.), Reading (Massachusetts) 1990.