The Evolution of Texts: Confronting stemmatological and Genetical Methods, Proceedings of the International Workshop held in Louvain-la-Neuve on September 1-2, 2004, edited by C. Macé, P. Baret, A. Bozzi, L. Cignoni". *Linguistica Computazionale*, volumes XXIV-XXV. Istituti Editoriali e Poligrafici Internazionali, Pisa-Roma. ISBN: 88-8147-435-2. ISSN: 0392-6907.

TESTING METHODS ON AN ARTIFICIALLY CREATED TEXTUAL TRADITION

Edited by PH. V. BARET, C. MACÉ, P. ROBINSON

Abstract - *Most debates over the validity of methods used for the classification of manuscript traditions are based on theoretical considerations. In order to practically compare different approaches, a workshop around an artificially created textual tradition was organized in Louvain-la-Neuve in September 2004. This paper sums up the main results of this 'method-storming'. All the methods used gave convergent outcomes but none of them was able to reconstruct exactly the real pedigree of the manuscripts. The rationale, advantages and drawbacks of the different methods are discussed in a comparative framework.*

Keywords - *stemmatics*, *testing methods*, *artificial manuscript tradition*, *phylogenetics*

1. INTRODUCTION

At least since Lachmann, editorial work has been considered as a scholarly discipline with scientific rules and methods. Accordingly, it is to be expected that one might test the validity of these rules and methods. One technique for doing so is, among scientists, to organize workshops where the same set of data is analysed by several groups of researchers and the results are compared (see e.g. Almasy *et al.*, 2001; Bovenhuis *et al.*, 1997; see also http://www.gaworkshop.org/). It is also very useful to work on a data set generated for the purpose of the experiment: since it is known what the results should be, it is then easier to determine which methods give better results.

[*] Contributions by C. Peersman, R. Mazza, C. Macé, J. Noret, E. Wattel, M. Van Mulken, P. Robinson, A.-C. Lantin, Ph. V. Baret, P. Canettieri, V. Loreto, H. Windram, M. Spencer, C. Howe, M. Albu, A. Dress.

This is what we have attempted to do with a textual tradition. We first created an artificial textual tradition, gave the data of this tradition to different scholars or groups of scholars, who then met in a workshop to compare their results. Experiments on artificially created traditions have already been carried out in the (recent) past. In the first case, the authors wanted to verify the validity of their computer method (Tombeur *et al*., 1979). In the second (Spencer *et al*., 2004), different phylogenetic approaches (e.g. parsimony, Neighbour-joining, Neighbournet) were applied on the twenty copies of the first eight paragraphs (834 words) of an English translation of the medieval German poem *Parzival* by a team of biologists, in order to test these phylogenetic methods.[1] In the present study, eight teams of biologists, mathematicians and philologists worked on the same artificial manuscript tradition, using different methods.

In this paper, we will first give a short account of how we made the manuscript tradition. Then, we will show the most important results obtained by the different teams. Finally, we will compare these results and try to build some provisional conclusions.

2. THE ARTIFICIAL TRADITION

2.1. *Creating an artificial tradition*

We did not try to reproduce any theoretical antique or medieval situation. Our only purpose was to know in advance the real history of a manuscript tradition and to confront it with several hypothetical reconstructions. We asked 11 people[2] to copy by hand three pages (1015 words) of a French text: Stig Dagerman, *Notre besoin de consolation est impossible à rassasier*, Paris: Actes Sud, 1952 (translated from Swedish by P. Bouquet). Only one of the copyists was not French-speaking, all scribes had a high level of

[1] See also the contribution by Windram *et al.* in this volume. On the real tradition of the original German text, see the contribution by M. Stolz in this volume.

[2] We wish to thank the volunteers: Nancy Castillo, François Gobert, Anne-Catherine Lantin, Renaud Lebrun, Aurélie Macé, Caroline Macé, Catherine Macé, Julie Macé, Laurence Tuerlinckx, Martine Simon and Tine Swaenepoel.

education (university or high school), several were philologists themselves, most of them were female. The text to be copied is written in a literary prose, with a relatively undemanding level of language. The copyists were asked to copy the text as quickly as possible and as spontaneously as possible. Two copyists copied the text twice; that means that for their second copy, they might have been influenced by their previous knowledge of the text. The text was first dictated, from the edition, with its punctuation, to the first copyist, who was not a native speaker. Then the copied text was corrected quickly by the person who had dictated it and who was French-speaking, without looking back at the text of the edition; some mistakes were left uncorrected, just by chance. This text, labelled T2, is the archetype of the whole tradition and we will mostly only refer to T2 in the following analyses.

An important difference between this artificial tradition and any real one is that the interval of time between the archetype and the latest copy is very small (a few weeks). There is no influence of changing time or space on the textual changes since all the copies are contemporary and made in the same cultural context. Accordingly, no historical or external element is relevant for the research.

The sigla of the secondary copies are A, B, C, D, F, J, L, M, S, U, V. The end of U is missing (from the word 583 to the end), but U was copied by V and a now lost manuscript (Ω) before it was mutilated. F was copied from U for the first part, then its copyist looked for another model for the missing part and copied this from C.

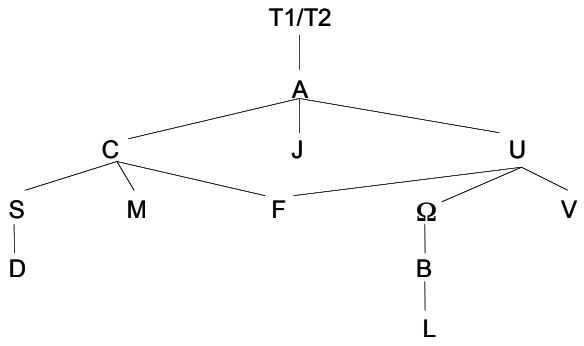

Figure 1 - The true stemma

2.2. *Preparation of the data*

The data concerning the tradition was made available to participating scholars in two forms. The first form was as a collation, recorded in an Excel spreadsheet. This is described in 2.2.1 and 2.2.2 below. The second form was as a set of transcripts, one for each witness, which the scholars themselves could then collate, using the Collate or other software: this is described in 2.2.3 below.

2.2.1. *Recording the collation in a spreadsheet*

The reference text, used to make the collation of the manuscripts, downloaded from the web (http://perso.wanadoo.fr/chabrieres), is slightly different from the 'original' (Publisher: Actes Sud), and from the archetype, the first manuscript copied from the original. This situation illustrates a case, which might not have been infrequent, of an archetype which is not THE perfect text but includes a few 'mistakes', which might have been corrected by its descendant(s). For those who used the computer-assisted collation program Collate, the choice of the base text might be less influential on the data (see 2.2.3).

We gave a number to each 'word' of the reference text.[3] We noted every difference between the reference text and each manuscript. We wrote what the manuscript reads instead of the reference text. We took punctuation into account.

2.2.2. *Encoding the variants*

To encode the variants, we used the variant location[4] (VL) instead of the word as a unit. The variant location is the word or group of words where at least one manuscript does not agree with the reference text.

For each variant location, we gave a (arbitrary) number to every variant reading:

0 is the reading of the reference text,
1 is the first reading we came across,
2 is the second etc.

We use the letter 'x' to code a lacuna (for U in this case) or when a manuscript has a longer omission at the place of another variant location.

According to this encoding, we end up with 119 variant locations (VL). This can be summarized as follows:

Number of words	1015
Number of witnesses	12 (without T1) 13 (with T1)
Number of variant locations (VL)	119
Number of variant readings (VR)	from 1 up to 4 on each VL

[3] By 'word', we mean: group of letters separated by two blank spaces, so, for example: [lois!"] is one word, [m'inspirent] is one word.

[4] We consider a unique variant location a place where the text changed at once. For example, at variant location 98, we assume that the copyist of D omitted at once 53 words (from word 851 to 903 - this is, by the way, typically a 'saut du même au même').

TABLE 1 - Type and frequency of the different types of variants

Code	Type	Count	Frequency (%)
A1	Addition of one word	6	4.51
A2	Addition of two words	0	0
A3	Addition of more than two words	1	0.75
E	Problem of misreading, miswriting...	10	7.52
G	Spelling	15	11.28
I1	Inversion of two consecutive words	0	0
I2	Inversion of two not consecutive words	0	0
I3	Inversion of more than two consecutive words	0	0
I4	Inversion of more than two not consecutive words	1	0.75
L1	Lexical (does not modify the meaning)	5	3.76
L2	Lexical (modifies the meaning)	6	4.51
L3	Lexical (gives the opposite meaning)	3	2.26
M	Morphological	24	18.05
O1	Omission of one word	9	6.77
O2	Omission of two words	3	2.26
O3	Omission of more than two words	4	3.01
P	Punctuation	46	34.59
		133	

It should be noted that the number of specific readings (readings occurring in only one witness) is high. Such variants are considered not kinship-revealing and are therefore neglected by parsimony methods. However philologists consider them important when seeking to determine 'dead-end copies', witnesses which have no extant copy.

TABLE 2 - Number of specific variants in each manuscript

Text	A	B	C	F	J	L	M	S	T2	U	V	Total
Specific variants	4	1	1	5	8	10	10	0	15	0	7	63

The high number of specific variants in T2 is striking, knowing that this manuscript is in fact the head of the tradition. Actually, most of the small peculiarities (mainly concerning the punctuation) in T2, which were easy to identify as such, have been omitted or corrected by the copyist of A, the only direct copy of T2. This makes of course more difficult to determine the root of the tradition.

2.2.3. *Transcripts*

We also generated complete transcripts of each manuscript copy, so that scholars could then collate these transcripts by computer program, in this instance using the program Collate (Robinson, 1994). This has certain advantages. In particular, Collate permits any witness to be used as the collation base, so reducing the possibility of bias from some readings appearing as 'base' while others are 'variants'. Indeed, when used in the 'parallel segmentation' mode, Collate avoids this possibility altogether, as all readings at any one variant location are presented alongside one another, with no presumption of a base text (Robinson, 2004). Collate also permits precise tailoring of each variant: suppressing variant readings which appear uninformative and deciding on the composition of each variant location (for example: the choice whether a variant location several words long should be treated as a single set of phrase variants or as a sequence of single-word variants).

Another team of researchers used a 'multiple alignment' method on the basis of the transcripts.[5]

[5] For details on the multiple alignment methods, see Spencer and Howe (2004).

3. PROCESSING THE DATA

The methods used by the eight teams will be organized in three main groups within this part.

TABLE 3 - Type of data and of methods used by the teams of researchers

	Team	Data	Methods
Stemmatics			
3.1.1.	(1) Peersman - Mazza - Macé	collations + matrix	stemmatics
3.1.2.	(2) Noret	collations	stemmatics
3.1.3.	(3) V. Mulken - Wattel	collations + matrix	weighted support
Phylogenetics - trees & networks			
3.2.1.1.	(4) Robinson / (5) Baret – Lantin	transcripts + Collate / matrix	neighbour joining
3.2.1.2.	(4) Robinson / (5) Baret - Lantin	transcripts + Collate / matrix	parsimony
3.2.1.3.	(6) Canettieri - Loreto	matrix + data compression	neighbour joining
3.2.2.1.	(7) Dress - Albu	matrix	networks
3.2.2.2.	(8) Windram - Howe - Spencer	transcripts + multiple alignment	networks

3.1. *Stemmatics*

3.1.1. '*Classical*' *Stemmatic Method*: *C. Peersman*, *R. Mazza*, *C. Macé*

The 'classical' method for building a stemma is based on the concept of the 'shared mistake' (Maas 1958, West 1973). Instead of 'mistake', we prefer to talk about 'source reading' (primary reading) and 'resulting reading' (secondary reading). It is often difficult to determine the direction of variation (which reading is primary and which is secondary), since, in many cases, the

readings found on a single variant location are all 'correct' (according to both grammatical and orthographical rules, and according to the semantic coherence of the text).

A reduced set of data was used[6], taking only into account the variant locations where it is possible to identify 'source reading(s)' and 'resulting reading(s)'. So, the number of variant locations we have kept is 52. For each variant location, we indicated the primary reading with a 0 and the secondary reading(s) with a 1 (or more).

Here is a list of groups of manuscripts sharing secondary readings:

TABLE 4 - Manuscripts sharing secondary reading

Nr of Mss	1 common sec. read.	2 common sec. read.
11	ABCFJLMST2UV	
10		ABCDFJMSUV
9	ACDFJMSUV	
8		ACDFJMST2
5	ACJMU	BFLUV
4	ACJM	CDMS
3	AJT2	
2	AJ	BL
2	FM	

From this table, we can draw several conclusions for the stemma. Let us start with the 'isolating' mistakes (see Table 2): it is clear that there cannot be extant copies of J, L and M in this tradition, since they contain many mistakes which cannot be found in any other extant manuscript.

[6] R. Mazza treated the data using a visualisation method, which he is presenting in another paper in these Proceedings. Some of his results were used by C. Macé and C. Peersman to build a 'stemma'.

The limited amount of data does not allow us to go beyond the following conclusions:

1) the position of T2 is strange: it might be placed on the top of the stemma or be descended from a different archetype;
2) all the other manuscripts derive from the same archetype and can be divided into three branches: A + J / B + L, F, U, V / C, D, M, S.

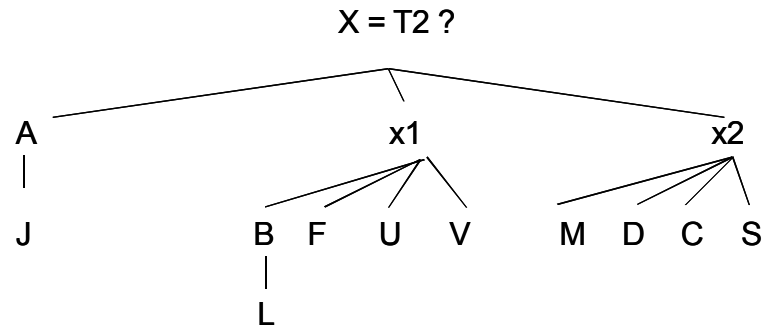

Figure 2 - A first stemma from the classic approach

3.1.2. *Another contribution by a classicist*: *J. Noret*

In a classical approach, a stemma is built from the bottom to the top, using the common mistakes to determine sub-archetypes (x, y, z, z' in this case), in which these mistakes had occurred before being transmitted to their descendants. Note that this stemma is very close to reality (and the shift of exemplar was clearly identified), except for the fact that the extant ancestors are not at the right place.

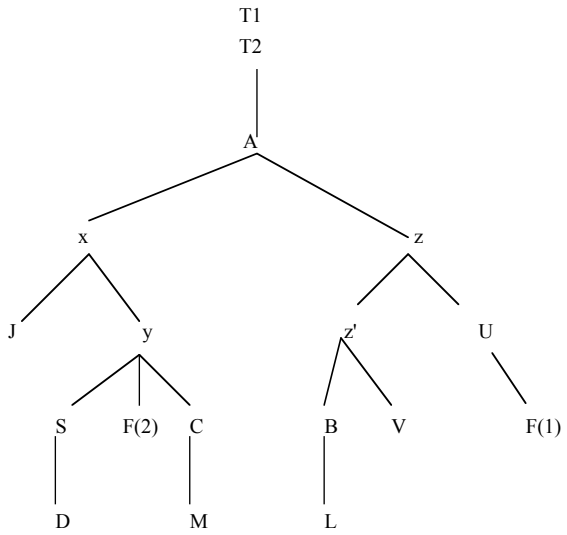

Figure 3 - Stemma obtained by 'classical' approach

3.1.3. *The Weighted Support Method*: *E. Wattel and M. van Mulken*

The tree in Figure 4 is drawn on the basis of the Weighted Support Method. This method distinguishes three stemmatological stages. Typical of this approach is that the decision about the correctness of readings is postponed until after the establishment of the fundamental kinship revealing relations between the manuscripts. At the first stage, the underlying, unrooted structure is determined from the complete list of variants. All extant manuscripts are, for the moment, considered end nodes and all edges reflect the number of supposedly non-transmitted manuscripts. In this phase, the kinship relationships between the manuscripts are established - possible cases of intermediacy are only determined at the second level.

All variants have been attributed a weight, depending on the type of variance which underlies the variant. For instance, semantic differences have been attributed a weight of 2 or 1, typographic or orthographic differences have been accorded 0.1 (cf. Den Hollander, 1997). In order to avoid interpretive mistakes at a very

early stage, the Weighted Support Method fundamentally includes all available information into the stemma building process. Minor variants are therefore attributed small weights, but are nevertheless admitted in the list of variants (Wattel and Van Mulken, 1996b).

To be able to process the list of variants, all variants are decomposed into quadruples. A quadruple is the smallest genealogically significant unit in a variant. A quadruple is a type2-variant, which is a combination of a group of variants, sharing a reading, opposed to another group of manuscripts sharing another reading (e.g. ab/cd). All variants, such as acdjmst1t2v/bflu in the present tradition, are decomposed into quadruples (e.g. ac/bf, ad/bf, aj/bf, am/bf, as/bf, etc.). These quadruples serve as basic information for the unrooted structure.

The common denominator ('best fit') of the information contained in the variants is now used to recompose a stemma. This is an iterative process, in which the sum of the variants which contradict our stemma is evaluated, and the process stops when the evaluated sum is smallest (Wattel, 2004). It implies that all cases of contradictory information must be accounted for separately, once the final tree is established.[7]

At the second stage, it is verified whether the provisional end nodes or intermediate nodes should be contracted (search for intermediacy). If a manuscript has served as a layer to another manuscript, then this manuscript is considered intermediate. The edges in Figure 4 all contain a number, and this number indicates a branch length, and the shorter a branch length, the more an edge is candidate for contraction. According to Figure 4, the edges departing from s, d, c, and f are candidates for contraction, and the same is true for the edges departing from the intermediate nodes @D @C and @A. At this stage, it is safe to contract the edges departing from @D, @C and @A.[8]

[7] Most of the arguments used to discard disturbing variants are *ad hoc* and *a posteriori*, and are not suited for an objective construction. We have to keep in mind that no information in the variants is completely and absolutely reliable. The best we can hope for is a stemma with only a small amount of contradictions, at the variant level, with the global stemma.

[8] The decision with regard to the intermediacy of an extant manuscript can only be taken after careful consideration of the evidence contained in the readings of the

The orientation of the stemma is the object of the third stage. The suspension point is the point that represents the (hypothetical or extant) manuscript closest to the archetype. At this level, it is determined at what point the pedigree must be rooted. Here, the value of the readings is taken into account, and the direction of the transmission is established. The philologist decides upon the authenticity of the readings that some of the variants may contain: not all variant readings allow to make such careful decisions.

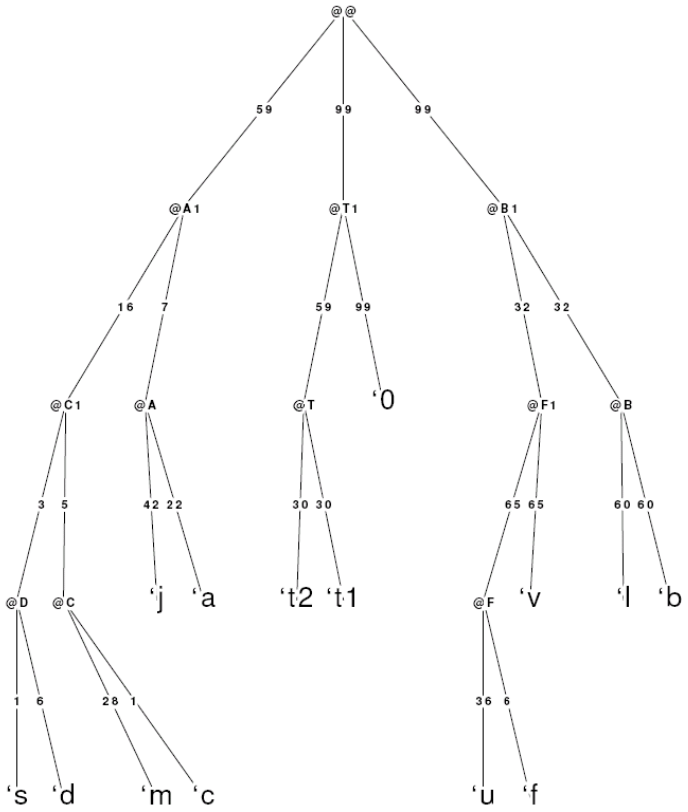

Figure 4 - Stemma based on the Weighted Support Method

variants: if the list of variants contains no information which obliges the philologist to maintain an intermediate node, then an extant manuscript can be considered intermediate (this implies that the list of variants contains no singular readings isolating the candidate intermediate from the rest of the tradition).

3.2. *Phylogenetics*

Phylogenetic analyses are widely used by evolutionary biologists to identify the pattern of evolutionary relationships among organisms. The inference strategy is based on modelling either the distances between objects in order of clustering them by levels of resemblance or the cost transition between the different steps of the possible histories to identify the most probable history. Phylogeny is a combination of a study of the relationships between objects (clustering) and a historical reconstruction (tree construction). In both cases, the methods are constrained by the calculation power of computers on the technical side and by the limit of representation of a complex (multidimensional) reality. In this case, the reality is a m x n array, where m is the number of taxa/objects and n is the number of characters/variants, but this cannot be apprehended by the human eye and must be converted to a reduced form in a two- or three-dimensional space. The usual modes of representation are trees or network. The tree-like representations have the advantage of being consistent with the rationale of descent by modification but they hide part of the complexity and some of the uncertainties.

In practice, the artificial tradition was tackled by three different phylogenetic methods: neighbour joining, parsimony and network.

3.2.1. *Tree based methods*

3.2.1.1. *Neighbour joining*: *P. Robinson*, *A.-C. Lantin*, *Ph. Baret*

In a neighbour joining (NJ) approach, the tree diagram is inferred from a distance matrix. At each stage of a sequential clustering, the principle of this method is to pair objects in order to minimize the total branch length of the tree. The number of objects to arrange is reduced by one at each step (Saitou and Nei, 1987). The PAUP implementation of NJ was used in the present applications (Swofford, 1998).

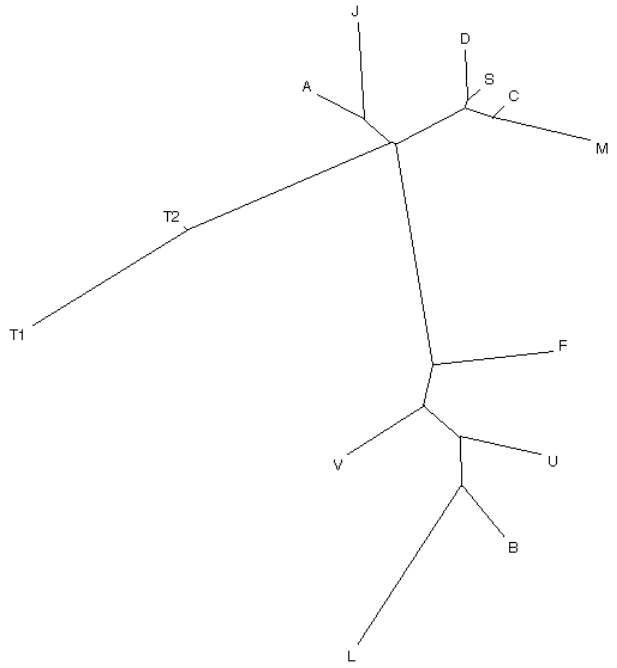

Figure 5 - Neighbour joining tree with branch lengths

Another Neighbour joining tree was generated by another team (Figure 6). A few small differences appear between the two diagrams (mainly the position of U), due to the different types of data used by the two teams.

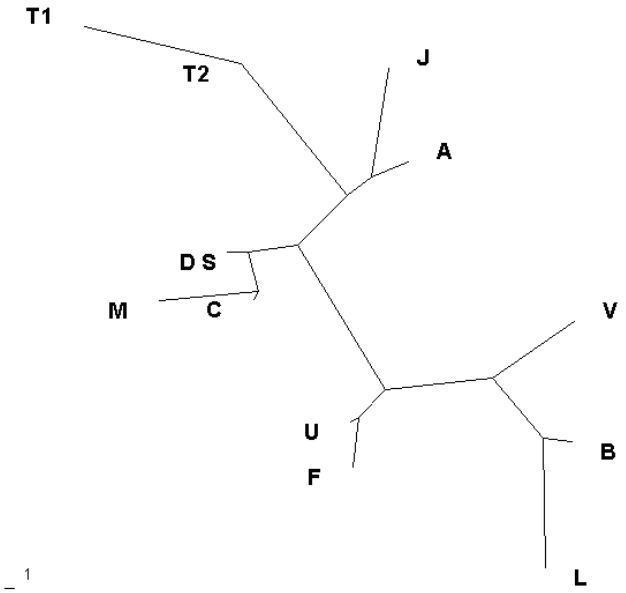

Figure 6 - Neighbour joining tree with branch lengths

The advantage of the method is that it provides a single tree with branch lengths very quickly. The process of construction is sequential and cumulative: at each step a decision is taken and the subsequent steps do not affect the previous decisions. In consequence, a local optimum at one step may condition the following steps and lead to a suboptimal global solution. In this case, the inferred tree may violate the principle of minimum evolution or maximum likelihood topology.

The distances are recalculated all along the process and branch lengths are shown on the final representation.

3.2.1.2. *Parsimony*: *P. Robinson*, *A.-C. Lantin*, *Ph. Baret*

The methods based on a parsimony criterion generate first all the possible topologies linking the different objects and then select the most parsimonious trees which are those that require the least number of changes at the variants locations. As the number of possible trees increases exponentially with the number of objects,

this approach requires intensive computation facilities. In addition, such methods often produce several equally parsimonious trees. The only way to make sense of this collection of equivalent trees is to build a consensus tree which sums up the common patterns of all the most parsimonious trees.[9] By construction, all parsimony or distance-based trees present bifurcations but in a consensus representation, the conflicting branching are summarized as a multifurcating branching pattern (see group C, D, M, S in Figure 7).

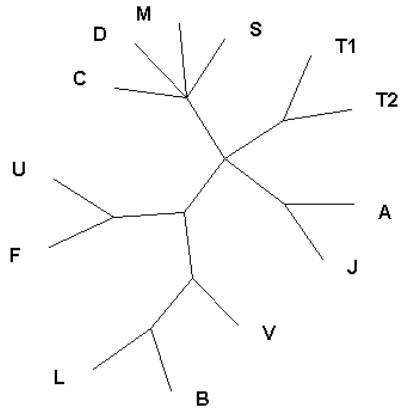

Figure 7 - Consensus unrooted tree resulting from a parsimony analysis (representation without branch lengths)

3.2.1.3. *Data Compression*: *P. Canettieri and V. Loreto*

Canettieri and Loreto propose a hybrid method for stemma reconstruction which combines the traditional ecdotics approach with an information theory oriented methodology (Benedetto *et al.*, 2002). 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 (Figure 8).

[9] Up to 400,000 in an application to a Gregory of Nazianzus' Homily (Lantin *et al.*, 2004).

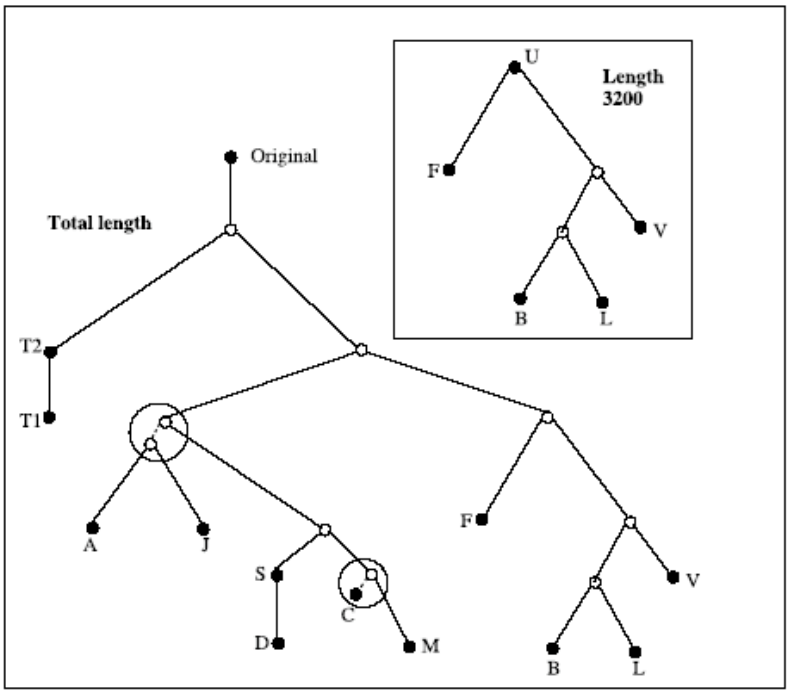

Figure 8 - Stemma resulting only from the data compression approach

Though the phylogenetic method (in this case Fitch-Margoliash: cf. Felsenstein, 1989) always produces binary trees, a circle indicates the regions of the tree where the true stemma could have had a trifurcation.

All the documents have roughly the same length (i.e. the same number of characters) except for the document U. Since the method compares the documents on the basis of the whole information they contain, the document U is treated separately.

3.2.2. *Network based methods*: *H. Windram*, *M. Spencer*, *C. Howe/M. Albu*, *A. Dress*

Network based methods provide an alternative to tree-based methods as they allow the simultaneous representation of multiple conflicting trees. This representation suggests possible reticulation and hybridization events.

Two methods were applied to the manuscript tradition: SplitsTree which is the archetype of the network methods and NeighbourNet, a method integrating both a network logic and an agglomerative algorithm (NJ).

3.2.2.1. *SplitsTree*

Tree-like networks allow for different and conflicting phylogenies. A split is any partition of two non-empty sets defined by an edge in a phylogenetic tree. Split decomposition implemented in SplitsTree is a method for obtaining weakly compatible splits.[10] It requires the prior calculation of a distance matrix. For ideal data - i.e. when the evolution at each variant location follows the overall pattern and this pattern is tree like -, the result of a SplitsTree analysis is a classic tree but in case of conflicting signals a network representation emerges.

For example, in a four taxa system (a, b, c, d), there are three possible unrooted topologies, ((a, b),(c, d)), ((a, c),(b, d)), and ((a, d), (b, c)), each corresponding to one of the possible splits that has two taxa in each set. For ideal tree-like data, only one of these splits will be supported by the distribution of differences at the variant locations. But due to the independent evolution of variant locations, deviation from the theoretical evolution process or lack of information, the data are not always so clearcut. Tree-building methods retain only the best-supported split and in consequence will hide part of the uncertainties. In contrast, split decomposition

[10] Weakly compatible splits are based on a less restricted system of splits when, for any three splits S1, S2, S3 and all Ai in Si (i=1,2,3), at least one of the four intersections A1 ∩ A2 ∩ A3, A1 ∩ A'2 ∩ A'3, A'1 ∩ A2 ∩ A'3, or A'1 ∩ A'2 ∩ A3 is empty (for details, see Bandelt and Dress, 1992; and Huson, 1998).

rejects only the least-supported split in the four-taxon case allowing for a reticulate representation.

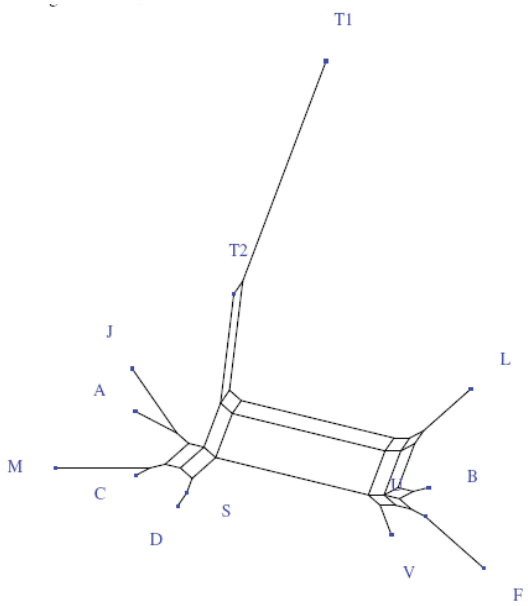

Figure 9 - A Splitstree analysis (excluding U)

Note that, on this graph, U was excluded because the fact that it is deficient could disturb the results. Interpretation of this kind of representation is not easy. It points to a repartition in two main groups (LBVF) and (MCDS AJ). AJ constitutes a sub-partition of the second group.

3.2.2.2. *NeighbourNet*

NeighbourNet is a method for constructing networks using the agglomeration principle. In comparison with the classic NJ, the agglomeration of pairs of objects is postponed until a later stage of the grouping process. In the iterative process, each pair of nodes is paired a second time before the distance matrix is reduced. The character matrix was converted to Hamming distances, counting gaps as differences but discounting missing data. Since the rate of

differences was low, there was no need to correct for multiple hits at the same location. The NeighbourNet method (Bryant and Moulton, 2002) is useful when we do not know much about the evolutionary processes that generated a data set, and when the phylogeny may not have been purely tree-like.

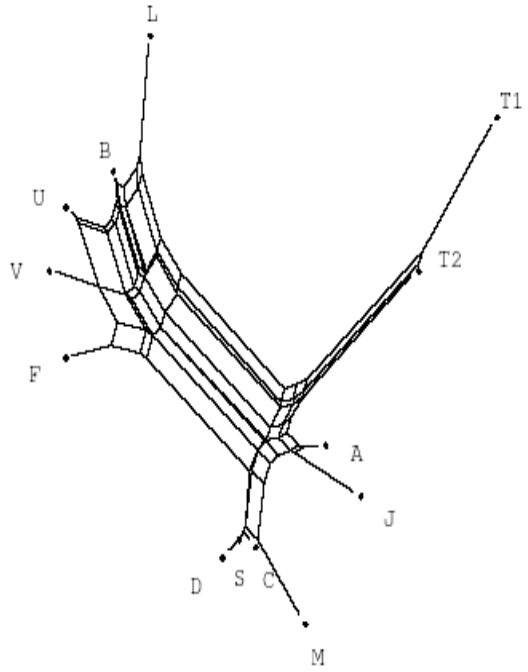

Figure 10 - A NeighbourNet approach

Overall, there appear to be two major groups of witnesses (B, F, L, U, V) and (A, C, D, J, M, S, T1, T2). Nevertheless, there is some support for a number of other ways of splitting the data. The corrected form T2 is closer to the other witnesses than T1.

The structure is very close to the one obtained with SplitsTree (Figure 9).

4. TWO SPECIFIC ISSUES

4.1. *Internalization of nodes*

In phylogenetic methods as well as in the first stage of the weighted support method (see 3.1.3), the objects are situated at the end of an edge. This makes sense in most biological situations, but in case of manuscript traditions, it is possible that both the model and its copy or copies survive. In other words, intermediates between the ancestral archetype and the most recent copies may exist in any manuscript tradition. In a tree representation they should be brought up to an internal branching point.

Spencer *et al.* (2004: 507) discussed the criteria to internalize an object. An object is likely to be an intermediate, if the length of its branch is very small. On figure 6, as well as on figures 9 and 10 (distances are less clear on figure 5), U, C and S can certainly be internalized and A, B can possibly be internalized. See also the discussion by M. Van Mulken, *supra* (3.1.3).

Another major issue in this process of internalization is to determine how deeply the element should be internalized. This question is not clear either for philologists. In both applications of mathematic and stemmatic methods, participants often stopped at the first level of internalization.

4.2. *Exemplar shift*

Assuming that a manuscript can be copied from more than one model, different methods were proposed to identify the position of a possible shift of exemplar, if any.

The most empirical method consists of splitting the data matrix into two or more parts equal in length, so as to run software on every part of the data and to compare the results. If the position of an object dramatically changes, a shift of exemplar may be suspected. The method is efficient but requires a trial and error approach to identify the location of the shift with precision.

Maynard Smith (1992) proposed a technique (the chi-squared method) to identify the most likely site of DNA recombination between chromosomes. The rationale of the method is to compare

the observed number of differences between sequences with the expected number of differences if no recombination occurred.

The application of this method by Windram *et al.* leads to a very efficient and precise location of the exemplar shift (Figure 11).[11]

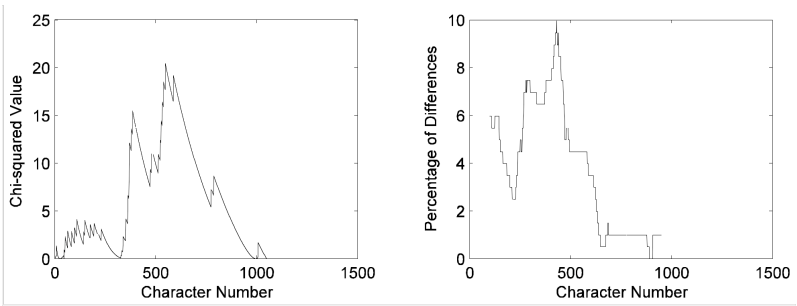

Figure 11 - Chi-squared value
and distribution of difference plots for F/C

It should be noted that the exemplar shift was clear enough to be identified by most philologists on the basis of the collation of the manuscripts (see 3.1.2.).[12] Finally, in the weighted support method, it is also possible to verify whether every copy derives from a single exemplar: it is checked whether the text tradition is best reflected in more than one pedigree; this search for a shift in kinship is called 'shock wave analysis' (see Wattel and Van Mulken, 1996a).

5. COMPARING THE METHODS

We may now compare the results produced by each method with the real stemma of the artificial tradition. For each method, we may ask:

[11] For details, see Windram *et al.* (2005) and the contribution by Windram *et al.* in this volume.
[12] In the case of the method used in 3.2.1., however, the small amount of data kept did not allow to identify this shift of exemplar.

a. Were the groups/clusters correctly identified? On the true stemma, three 'groups' are visible: the group of C, J (being a mono-element group), and the group of U. It is acceptable that J is not recognised as a 'group'.
b. Is the distance between manuscripts indicated? The distance is an important piece of information, showing the extent of the differences between the texts of two witnesses.
c. Does the method recognize the possibility for a manuscript to have more than two copies (polytomy)? It should be noted that a polytomic representation can bear different meanings: for example, the polytomy in consensus trees (parsimony) means that several possibilities of branching are equally possible, it does not imply that one object has several descendants (see 3.2.1.2).

The purpose is to identify the main differences between the methods when they are used in an automated mode and to compare them with the not-automated philological methods. It is clear, though, that in most cases, methods will be combined and/or applied to subsets of data in order to test hypotheses about the tradition. Moreover, the analyses were achieved by interdisciplinary teams composed by philologists and biologists or mathematicians and these interactions lead to combinations of different points of view.

TABLE 5 - Comparison between the different methods

Team	1/2	3	4/5	4/5	6	7/8
Method	Stem	WS	NJ	Pars	DC	Netw
Paragraph	3.1.1-2	3.1.3	3.2.1.1	3.2.1.2	3.2.1.3	3.2.2.1-2
AUTOMATED	no	partly	yes	yes	yes	yes
CLUSTERING	ok	ok	ok	ok	ok	ok
DISTANCES	no	yes	yes	possibly	yes	yes
POLYTOMY	yes	not at the 1st stage	no	yes	no	yes

Classical Stemmatics (teams 1 and 2) gives relatively good results compared with the true stemma, but it is at the same time an analysis and an interpretation of the results of the analysis, while

the results given by the other methods are before any interpretation. In addition, classical stemmatics relies on *a priori* assumptions about 'primary' and 'secondary' variant readings. Furthermore, the processing time of these methods, as they are not automated, is much more dependent upon the quantity of data than the others.

None of the methods used was able to lead exactly to the 'true' history of the text. Yet, despite the difficulties peculiar to this tradition, all were able to define correctly the groups among the tradition, with a hesitation on the position to give to A/J.

The fact that the different methods used came to very convergent results can be explained by the combination of two elements:

a) the nature of the methods: the different methods share a common rationale (parsimony, minimum evolution, clustering based on resemblance) and most of the differences derive from secondary aspects such as the mathematical tools used or the mode of representation of the results.
b) the nature of the data: the provided data were not especially informative, some aspects of the tradition were clearly documented, while some others could only be suspected, by guessing and interpreting the results.

Two important issues have been left aside so far: the sequence of the tradition and its root. The question of the sequence is linked with the problem we have raised about the internalization of the nodes (see 4.1). There are other aspects in this question. In the true stemma, there was a suppressed witness (Ω), which was intermediate between two extant manuscripts (U and B): Ω could not be identified by any method. Conversely, stemmatological methods often suppose missing witnesses which never really existed. This is probably the weakest aspect of stemmatic methods, and maybe a side-effect of their 'obsession' with the root. This fixed idea of the philologists is not shared by people working with phylogenetic methods: none of them proposed a temporal structure or a clear root of the tradition. This is a key characteristic of phylogenetic methods, which typically work to hypothesize

unoriented trees of relationship, where the ancestor could (theoretically) be located at any point within the tree. Rooting is then postponed to a further step and most of the time based on external evidences.

6. General conclusions

The objective of this paper was to offer the possibility of a comparison of a wide range of methods, not to fully explain any of them. Leads for several further developments were given but would still need to be explored. The focus of the 'automated' methods was on the identification of clusters and the sequence of texts. Other aspects were addressed by some ad hoc 'manual' solutions.

The nature of the data had a significant influence on the results. The number of witnesses was relatively small (13) as well as the number of informative variants.[13] Such situations occur in the real world of philology (a short text in a few witnesses), and this has the advantage of providing a more manageable set of data. What was very different from a real philological situation is that there was no available information about the manuscripts as historical objects, about their place and time of copy and the factor 'time' did not play any role in the evolution of the tradition, since all the witnesses were contemporaneous.

Due to the limited size of the tradition and the purpose of this experiment, technical and computing aspects were not major issues. It is noteworthy that the interest of computerized methods may be their capacity to process very large amounts of data and their ability to deal quickly with new or modified data. This might become particularly important with large textual traditions and different scholars working in partnership to provide new transcripts and collations.

[13] For Peter Robinson, a key figure is that the tradition provides only 64 parsimony-informative characters which is the low edge of reliability. As a rule of thumb, 200 informative characters are needed for reasonable secure results. In the case of the *Canterbury Tales* in the Miller's Prologue and Tale, for example, there are 1735 places of variation which are parsimony informative.

In any event, one of the most important points raised by the common work on this artificial data set is that any stemma or genealogical tree of a textual tradition is always a theoretical reconstruction. One may draw two conclusions. Firstly, the advantages of computerized methods are clear: they are objective, reproducible, able to deal with a large amount of data, and they can produce results which are reliable. Secondly, computerized methods do not seem capable, of themselves, of producing the most complete account of witness relationships. We have mentioned the need for scholarly analysis to determine direction of variation and hence tree orientation: this seems beyond the scope of computer methods. Further, computer methods have difficulty with polytomy (more than two witnesses copied from a single source): one may judge that the position of certain witnesses within a computer-generated tree is consistent with polytomy, but it is difficult to see how this judgement could be reliably left to the computer. Taken together, this suggests that a partnership of computer methods and traditional analysis might give the best results.

Acknowledgements

The editors wish to thank the following people for their helpful advice and corrections: B. Bordalejo, M. Spencer, C. Peersman, M. Van Mulken, H. Windram and M. Albu.

REFERENCES

ALMASY L., TERWILLIGER J. D., NIELSEN D., DYER T. D., ZAYKIN D., BLANGERO J., *GAW 12*: *Simulated genome scan*, *sequence*, *and family data for a common disease*, «Genetic Epidemiology», XXI (2001), 332-338.

BANDELT, H.-J., DRESS, A., *A Canonical Decomposition Theory for Metrics on a Finite Set*, «Adv. Math», XCII (1992), 47-50.

BENEDETTO D., CAGLIOTI E., LORETO V., *Language trees and zipping*, «Physical Review Letters», LXXXVIII (2002), 048702-048705.

BOVENHUIS H., VAN ARENDONK J. A. M., DAVIS G., ELSEN J.-M., HALEY C. S., HILL W. G., BARET P. V., HETZEL D. J. S., NICHOLAS F. W., *Detection and mapping of quantitative trait loci in farm animals*, «Livestock Production Science», LII (1997), 135-144.

BRYANT D., MOULTON V., *Neighbornet*: *an agglomerative method for the construction of planar phylogenetic networks*, «Lecture Notes in Computer Science», MMCCCCLII (2002), 375-391.

DEN HOLLANDER A. A., *De Nederlandse bijbelvertalingen 1522-1545*, De Graaf, NieuwKoop, 1997.

FELSENSTEIN, J., *PHYLIP - Phylogeny Inference Package* (Version 3.2), «Cladistics» V (1989), 164-166.

HUSON D. H., *SplitsTree*: *analyzing and visualizing evolutionary data*, «Bioinformatics», XIV (1998), 68-73.

LANTIN A.-C., BARET Ph. V., MACÉ C., *Phylogenetic analysis of Gregory of Nazianzus*' *Homily 27*, in G. PURNELLE, C. FAIRON, A. DISTER (eds.), *Le poids des mots. Actes des 7èmes Journées Internationales d'Analyse statistique des Données Textuelles* (*JADT04*), Presses Universitaires de Louvain, Louvain-la-Neuve, 700-707.

MAAS P., *Textual criticism*, transl. from German by B. Flower, Clarendon, Oxford, 1958.

MAYNARD SMITH J., *Analyzing the mosaic structure of genes* «J. Mol. Evol.», XXXV (1992), 126-129.

ROBINSON P. M. W., *Collate*: *Interactive Collation of Large Textual Traditions*, Oxford University Centre for Humanities Computing, Oxford, 1994.

ROBINSON P. M. W., *Collation Rationale*, in *The Miller*'*s Tale on CD-ROM*, Scholarly Digital Editions, Leicester, 2004.

SAITOU N., NEI M., *The neighbor-joining method*: *a new method for reconstructing phylogenetic trees*, «Molecular Biology Evolution», IV (1987), 406-425.

SPENCER M., DAVIDSON E. A., BARBROOK A. C., HOWE C. J., *Phylogenetics of artificial manuscripts*, «Journal of Theoretical Biology», CCXXVII/4 (2004), 503-511.

SPENCER M. and HOWE C. J., *Collating texts using progressive multiple alignment*, «Computers and the Humanities», XXXVIII (2004), 253-270.

SWOFFORD D. L., *PAUP*. Phylogenetic analysis using parsimony* (**and other methods*), Sunderland (Mass.), 1998.

TOMBEUR P., BOULANGER J.-C., SCHUMACHER J., *Génération automatique d'un stemma codicum*, in *La pratique des ordinateurs dans la critique des textes Colloque international du CNRS*, Paris, 29-31 mars 1978 (Colloques Internationaux du C.N.R.S., 579), Publications du CNRS, Paris, 1979, 163-183.

WATTEL E., *Constructing initial binary trees in stemmatolgoy*, in P. VAN REENEN, A. DEN HOLLANDER, M. VAN MULKEN (eds.), *Studies in Stemmatology* II, Benjamins, Amsterdam/Philadelphia, 2004, 145-165.

WATTEL E., VAN MULKEN M. P., *Shockwaves in text traditions*, in P. T. VAN REENEN, M. P. VAN MULKEN (eds.), *Studies in Stemmatology* I, Benjamins, Amsterdam/Philadelphia, 1996a, 105-122.

WATTEL E., VAN MULKEN M. P., *Weighted formal support of a pedigree*, in P. T. VAN REENEN, M. P. VAN MULKEN (eds.), *Studies in Stemmatology* I, Benjamins, Amsterdam/Philadelphia, 1996b, 135-168.

WEST M. L., *Textual criticism and editorial technique*: *applicable to Greek and Latin texts*, Teubner, Stuttgart, 1973.

WINDRAM H. F., SPENCER M., HOWE C. J., *The identification of exemplar change in the* Wife of Bath's Prologue *using the maximum chi-squared method*, «Literary and Linguistic Computing», XX (2005), 189-204.