Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets

Teemu Roos*
Department of Computer Science and the National Library of Finland, University of Helsinki, Helsinki, Finland

Tuomas Heikkilä
Department of History, University of Helsinki, Helsinki, Finland

Correspondence: Teemu Roos, Department of Computer Science, P.O. Box 68, FI-00014 University of Helsinki, Helsinki, Finland
E-mail: teemu.roos@cs.helsinki.fi

Abstract

Given a collection of imperfect copies of a textual document, the aim of stemmatology is to reconstruct the history of the text, indicating for each variant the source text from it was copied. We describe an experiment involving three artificial benchmark data sets to which a number of computer-assisted stemmatology methods were applied. Contrary to earlier similar experiments, we propose and use a numerical criterion to evaluate all the solutions. Moreover, our primary data set is significantly larger than used before. The results suggest the superiority of two computer-assisted methods amongst those tested: the maximum parsimony method implemented in the PAUP* software package and a related compression-based method we have proposed in earlier work.

'No book is published without some discrepancy in each one of the copies.
Scribes take a secret oath to omit, to interpolate, to change.'
(Jorge Luis Borges: *The Lottery in Babylon*,
in *Labyrinths: Selected Stories & Other Writings*, 1962)

1 Introduction

Before the development of the art of printing, pioneered by Johannes Gutenberg in the fifteenth century, written works were copied by hand. This resulted in numerous unintentional errors and intentional modifications. They accumulated in copies of copies, copies of copies of copies, etc. Consequently, a text of any importance ended up existing in a group of different versions, a so called *tradition*, some of which are all but identical to the original, some perhaps hardly recognizable. Connecting each version to its *exemplar*, i.e. the version from which it was copied, gives a tree-like structure called the *stemma*, with the original version as the root. The aim of *stemmatology* is to recover this structure given a set of surviving variants.

There is an obvious analogy to the transmission of textual information through the stemma in evolutionary biology. Namely, the transmission of genetic information and the development of species,

*Present address: The Helsinki Institute for Information Technology HIIT.

doi:10.1093/llc/fqp002 Advance Access published on 14 March 2009

often visualized as a phylogenetic tree or, more poetically, the 'Tree of Life', has the same characteristics of unintentional errors and iterative multiplication as 'manuscript evolution'. The methods developed for phylogenetic analysis have been fruitfully adapted and applied to stemmatology, see e.g. (Robinson and O'Hara, 1992; Spencer *et al.*, 2002).

In July 1991, Peter Robinson posted an announcement about a Textual Criticism Challenge on various internet bulletin boards. The challenge was based on 44 versions of the *Svipdagsmal* narrative, consisting of poems *Grougaldr* and *Fjolsvinnsmal* in Old Norse, written in about 1650–1830. The two poems together are about 1,500 words long. In the challenge, the objective was to divide the manuscripts into groups of related documents, to identify the readings that characterize each group, and to find out the relationships between the groups. In addition to these, one was to identify cases where different readings have originated by copying from multiple sources, i.e. *contamination*.

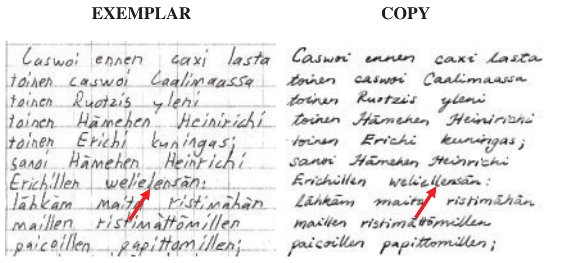
These tasks would have to be solved by analysing the texts alone, about which the available information was a table of agreements and disagreements between the variants in each reading. While the tradition used in Robinson's challenge was real, and hence, the exactly correct solution was not known, there was a reasonably good understanding of the relationships between the manuscripts founded on external evidence, which was not available to the participants of the challenge. Three submissions were entered to Robinson's challenge. One of the contestants, Robert O'Hara used the software package PAUP, which was originally developed for phylogenetic analysis, the study of the relatedness among various groups of organisms. O'Hara obtained his solution in mere 5 min, achieving spectacular success in reconstructing the main groups of manuscripts and the relationships between the groups. The PAUP software is not designed to handle contamination, and hence that part of the challenge was left unanswered. Nevertheless, O'Hara's (and PAUP's) success was a decisive demonstration of the applicability of computer-assisted methods in stemmatology.

More than 15 years have now passed since Robinson's ground-breaking challenge. Meanwhile, a number of different methods of computer-assisted methods have been proposed and applied to both real-life and artificial textual traditions; for experiments with artificial traditions, see e.g. (Spencer *et al.*, 2004; Baret *et al.*, 2006). While the application of such methods still faces some scepticism[1]—a number of scholars of philology, for instance, still consider the nineteenth century methods of textual criticism the best available tool in shaping a textual tradition—they have become a standard tool in a stemmatologist's toolbox.

In order to gauge the current situation in the field, we organised (with Petri Myllymäki) the Computer-Assisted Stemmatology Challenge.[2] The main idea was to construct a collection of traditions that could be used to benchmark different stemmatological methods, with special emphasis on artificial traditions for which the correct solution is known exactly. Unlike Robinson's challenge where a real data set was used, experiments with artificial data enable objective and precise comparison of alternative solutions against a ground-truth solution.

2 Data Sets

The primary data set, called *Heinrichi*, was created by volunteer scribes, who copied a given text by hand, following an imaginary stemma (Fig. 3 below). Care was taken to simulate the procedure of medieval copying of texts in other respects, too. As the aim was to be as realistic as possible, a real text and its manual copying was preferred to simulating the copying process by computer. The original text was a late medieval Finnish folktale *Piispa Henrikin Surmavirsi* ('*Death Psalm of Bishop Henry*'), written in the seventeenth century and published in (Neovius, 1912). The text was chosen for the purpose because it is written in old Finnish, only partially recognizable by the scribes. This was intended to resemble the situation faced by medieval scribes copying Latin or vernacular texts. Although still a living language and a language of relatively fixed grammatical rules and spelling, Latin was no more spoken or written as the mother tongue by anyone during the Middle Ages. Medieval or early modern Finnish of the *Heinrichi*



EXEMPLAR

Caswoi ennen caxi lasta
toinen caswoi Caalimaassa
toinen Ruotzis yleni
toinen Hämehen Heinirichi
toinen Erichi kuningas;
sanoi Hämehen Heinrichi
Erichillen welielensän:
lähkäm maita ristimähän
maillen ristimättömillen
paicoillen papittomillen;

COPY

Caswoi ennen caxi lasta
toinen caswoi Caalimaassa
toinen Ruotzis yleni
toinen Hämehen Heinirichi
toinen Erichi kuningas;
sanoi Hämehen Heinrichi
Erichillen weliellensän:
lähkäm maita ristimähän
maillen ristimättömillen
paicoillen papittomillen;

Fig. 1 An example of a typical error introduced in the process of copying. The word 'welielensän' (meaning 'to his brother') in the exemplar (left) is transformed into 'weliellensän' in the copy (right) by duplicating the letter *l*

data, in turn, was not a language of strictly fixed spelling. However, as the copyists of the simulation approached their copying task from their modern linguistic background of a fixed language, it is reasonable to assume that there are significant similarities between the medieval copying process and the simulation.

A majority of the copyists—15 out of 17—were Finnish native speakers studying at a university. Of the two others, one had Swedish and one Hungarian as their mother tongue, but both had good command of Finnish, as well. Some of the volunteer scribes copied the text twice, some of them thrice. Even if the copyists may thus have benefitted from their previous knowledge of the textual contents when copying their second or third version, the situation corresponded to the medieval copying of a text, the contents of which were more or less familiar to the scribes. This was often the case with a variety of liturgical and biblical texts, for instance. The practical copying of the *Heinrichi* data set was done according to a plan known only to the organizers of the Challenge. Care was taken to include a number of contaminated textual witnesses in the material in order to make the artificial tradition resemble an actual historical case. An example of a typical case of an error introduced by misreading or miswriting is shown in Fig. 1.

The *Heinrichi* text was approximately 1,200 words long. The total number of variant texts created was sixty-seven, of which thirty were held back from the challenge data set in order to simulate the realistic scenario where a significant portion of the manuscripts are missing. Furthermore, we deleted significant parts of some of the manuscripts, in order to simulate the cases where the manuscripts are partially destroyed. The greater number of manuscripts and significant number of missing manuscripts and parts of the texts are the major differences between this data set and those used previously in experimenting with artificial textual traditions. Therefore, the created *Heinrichi* data set is not only the clearly largest and most complex artificial tradition created so far, but also represents a more realistic case than the earlier examples described below.

In addition to *Heinrichi*, two more artificial traditions were included in the challenge data. First, there was a collection of 13 copies of the text *Notre besoin de consolation est impossible à rassasier* (Dagerman, 1952), provided by Caroline Macé. The data have been used in a similar experiment as ours, where eight different methods were applied to the collection. The data and the experiment are described in (Baret *et al.*, 2006). Second, we were provided a collection of 21 copies of the medieval German poem *Parzival* (von Eschenbach, 1980) by Matthew Spencer and Heather F. Windram. In order to make the task more realistic, we held back five manuscripts from the *Parzival* data set, so that the analysis had to be performed using the remaining 16 manuscripts. The original collection, and results of its analysis using several stemmatological methods, are described in (Spencer *et al.*, 2004) (Table 1).

In addition to the three artificially created textual traditions, one real-life data set was included in the Challenge. The *Legend of St. Henry* (in Latin *Legenda s. Henrici*) is a Latin text composed in Finland towards the end of the thirteenth century. The text is known in more than fifty different medieval versions dating from the early fourteenth century to the early sixteenth century (Heikkilä, 2005). Thus, it represents a real case where the textual witnesses were written during a time-span of four centuries, and hence, where the aspect of time and the changes in the traditions of writing are present. These are aspects that even the most carefully executed

Table 1 Summary of the data-sets

Data	Length in words (max)	Parsimony informative words	Number of MSS	Number of missing MSS	Number of contaminated MSS[a]
Heinrichi	1,208	805/617[b]	67	30 (45%)	4 (11%)
Parzival	855	88	21	5 (24%)	0 (0%)
Notre besoin	1,035	71	14	1 (7%)	1 (8%)
Legend	1,185	325	52[c]	Unknown	Unknown

The three first data-sets are artificial. A word is "parsimony informative" if there are at least two variants that are both present in at least two manuscripts.

[a] Relative to the number of observed (non-missing) manuscripts.

[b] Number of informative words after removal of three outlier manuscripts (*Da, I, J*).

[c] Number of known surviving manuscripts.

simulation of the copying procedure cannot reproduce. Of course in this case the true solution is not known, but there are several clues about it which can be inferred from external evidence available in the properties of the actual manuscripts, such as the style of writing and the materials. Furthermore, the applying of traditional textual criticism to the extant textual versions allows the shaping of significant parts of the stemma of the *Legend of St. Henry*.

3 Measuring Distances between Stemmata

In order to enable objective comparison of different proposed solutions when a correct stemma is available, we introduce a general distance measure between two stemmata. Earlier distance measures have been restricted to strictly tree-shaped structures and, in particular, to bifurcating trees where each node is connected to either three or one other nodes by an edge,[3] see (Waterman and Smith, 1978; Critchlow *et al.*, 1986). However, the correct stemma is usually not strictly a tree due to contamination, and even less likely, a bifurcating one.

Our a*verage sign distance* depends on the number of edges between pairs of nodes, ignoring possible edge length information (although an edge-length dependent version is easily obtained by replacing the distance $d(A, B)$ below by the sum of edge-lengths along the shortest path between nodes A and B.). For each pair of nodes A, B, we define the true distance, $d(A, B)$, as the number of edges on the shortest path between A and B in the correct stemma. Clearly, the distance from A to B is the same as the distance from B to A: $d(A, B) = d(B, A)$. Similarly, we denote the same quantity computed from a proposed stemma, by $d'(A, B)$. If the proposed stemma is correct, then we have $d'(A, B) = d(A, B)$ for all pairs of nodes, and otherwise the two values differ for some A and B. Given a three node A, B, and C, we can measure the distances $d(A, B)$ and $d(A, C)$. We consider which one of these two distances is greater than the other, or whether they are equal. Let $sign(d(A, B) - d(A, C))$ denote the sign of the difference between the two distances, so that it takes value -1 if the distance from A to B is less than the distance from A to C, and the value $+1$ if the opposite holds. If the distances are equal, the sign function takes value 0.

Now define for any triplet A, B, C, the index

$$u(A, B, C) = 1 - \frac{1}{2}|\mathrm{sign}(d(A, B) - d(A, C)) - \mathrm{sign}(d'(\mathrm{A}, \mathrm{B}) - d'(A, C))|,$$

where the $|.|$ denotes the absolute value. The index measures the correctness of the proposed stemma in terms of the order of the distances from A to B and from A to C; the index equals one if the proposed stemma agrees with the correct one about which one of the nodes B and C is closer to node A (or if they are equally close). If the two distances are equal in one but only one stemma, then the index equals ½, and if the order of the distances is wrong, the index equals 0. Table 2 gives the value of the index for all combinations of the relative order of the relevant distances.

Table 2 The value of the index $u(A, B, C)$ for all combinations of the relative order of distances of nodes B and C from node A in the correct stemma and proposed stemma

Correct stemma	Proposed stemma	$u(A, B, C)$
$d(A, B) < d(A, C)$	$d'(A, B) < d'(A, C)$	1
$d(A, B) < d(A, C)$	$d'(A, B) = d'(A, C)$	½
$d(A, B) < d(A, C)$	$d'(A, B) > d'(A, C)$	0
$d(A, B) = d(A, C)$	$d'(A, B) < d'(A, C)$	½
$d(A, B) = d(A, C)$	$d'(A, B) = d'(A, C)$	1
$d(A, B) = d(A, C)$	$d'(A, B) > d'(A, C)$	½
$d(A, B) > d(A, C)$	$d'(A, B) < d'(A, C)$	0
$d(A, B) > d(A, C)$	$d'(A, B) = d'(A, C)$	½
$d(A, B) > d(A, C)$	$d'(A, B) > d'(A, C)$	1

The *average sign distance* is now given by the average of the index $u(A, B, C)$ over all *distinct*, *observed* triplets A, B, C, i.e. triplets where each node is observed (the text is available), and none of the three nodes are equal to each other. The distance is applicable to any pair of graphs, both of which include all the observed nodes. Either graph may include any number of additional nodes, which need not be the same for the correct and the proposed stemma.

The computation of the distance requires that the pair-wise distances between all nodes are computed and recorded in a table. After this, the average can be computed in about n^3 steps, where n is the number of observed nodes.

4 Methods

There were two submissions to our challenge. In addition, we applied a simple hierarchical clustering heuristic, a compression-based method we have developed earlier (Roos *et al.*, 2006), the PAUP software (version 4) (Swofford, 2003), and the SplitsTree4 software (version 4.10). In order to compare our results with earlier work, we also ran the average sign distance calculation on several solutions to the *Notre besoin* tradition published in (Baret *et al.*, 2006).

The challenge solution by Rudi Cilibrasi was obtained by the software CompLearn, developed and maintained by Cilibrasi and others.[4] The core component of CompLearn is a universal distance metric, based on Kolmogorov complexity, see Li and Vitanyi (1997). The universal metric minorizes a large class of distance metrics in the sense that if two objects are close to each other according to any metric in the class, then the objects are also close to each other according to the universal metric, at least asymptotically as the complexity of the objects grows. In practice, Kolmogorov complexity is uncomputable and has to be approximated by actual compression algorithms, like LZ78 (Ziv and Lempel, 1978), which gives an approximation of the universal distance metric. Having computed the pairwise distances of the manuscripts, CompLearn employs a clustering algorithm based on comparing the relative distances between each quartet of the manuscripts.

A second solution, by George Giannakopoulos and Ilias Zavitsanos from the National Center for Scientific Research Demokritos and University of Aegean, was obtained by a method where the texts are represented as character n-gram graphs of various n-gram ranks. The manuscripts are clustered according to a heuristic where a parent is inserted for every pair of vertices (manuscripts) that are most similar to one another, but have lower similarity that the average similarity between all pairs of manuscripts plus the standard deviation of the latter.

In order to establish a baseline level of performance, we applied a basic hierarchical clustering method. All the manuscripts were first aligned in a rectangular matrix where each row corresponds to a certain location in the text, so that if two or more manuscripts have the same word at the point in question, then these words are on the same row in the matrix. Aligning the texts was a relatively obvious task, and was performed by hand although there are several tools that could be used to achieve this automatically, see, e.g. (Notredame, 2007) and references therein. Having constructed the matrix, pairwise agreement ratios A_{ij} between each pair of manuscripts (i, j) were computed by counting the number of words for which the two manuscripts match, and dividing this number by the number of rows in the matrix. This yields a ratio between zero (all words differ between the two manuscripts) and one, achieved when all the words match. The agreement ratios were used to obtain a hierarchical clustering by standard agglomerative clustering with the so called complete-linkage criterion (Johnson, 1967).

Table 3 The accuracy of different methods for the three artificial traditions measured as the average sign distance (see Section 3): the percentage value is obtained as 100% × average sign distance

	Data		
Method	***Heinrichi*** **(%)**	***Parzival*** **(%)**	***Notre besoin*** **(%)**
RHM	**76.0**	79.9	76.9
PAUP*			
Parsimony	74.4	77.8	74.5
Parsimony BS[b]	73.6	85.4	77.3
Neighbour Joining	64.4	81.5	76.2
Neighbour Joining BS[b]	62.9	**87.1**	77.4
Least squares	64.2	81.5	70.2
Least squares BS[b]	62.6	79.8	73.0
n-Gram clustering	64.4	79.3	66.4
SplitsTree4			
NeighborNet	59.1	77.8	70.2
SplitDecomp.	53.1	74.5	73.1
ParsimonySplits	56.8	83.7	71.6
CompLearn	52.7	81.5	70.6
Hierarchical clustering	51.4	72.6	60.2
'Classical' method A[a]			74.4
'Classical' method B[a]			**85.1**
Weighted support method			66.3
Neighbour joining A			76.0
Neighbour joining B			75.0
Parsimony			74.4
Data compression			62.0

The result of the best method(s) for each tradition is shown in bold face. The results for the last seven methods are known only for the *Notre besoin* data, and they are based on stemmata published in (Baret *et al.*, 2006).

[a] Non-computer-generated solutions.

[b] BS, bootstrap consensus tree.

In an earlier paper (Roos *et al.*, 2006), we have presented a method for stemmatic analysis, dubbed RHM in the following, the core of which is a compression-based criterion for comparing alternative stemmata. The method uses a combination of stochastic optimization and dynamic programming to simultaneously search for a tree structure and the texts in the missing interior nodes that optimize the criterion. An outline of the method is given in the Appendix.

In default operation, the PAUP software uses the so called maximum parsimony criterion to rank alternative tree structures. A maximally parsimonious tree minimizes the total number of differences between directly connected nodes—species, individuals, or manuscripts that are directly related—possibly weighted by their importance. For traditions with more than a handful of manuscripts, the number of possible trees is too large for exhaustive search, and hence, a heuristic is used to find as good a tree as possible. Other criteria available in the PAUP software include neighbour joining (Saitou and Nei, 1987) and least squares. We ran the software with the parsimony, neighbour joining, and least squares criteria using default settings. For all these, we provide results with and without bootstrapping (consensus tree, level 50%).

In contrast to the other methods, the SplitsTree4 software (Huson and Bryant, 2006) provides methods that construct non-tree-shaped stemmata. We applied the NeighborNet, SplitDecomposition, and ParsimonySplits methods with default settings.

5 Results

Table 3 summarizes the performance of the thirteen methods applied to the artificial challenge data sets,

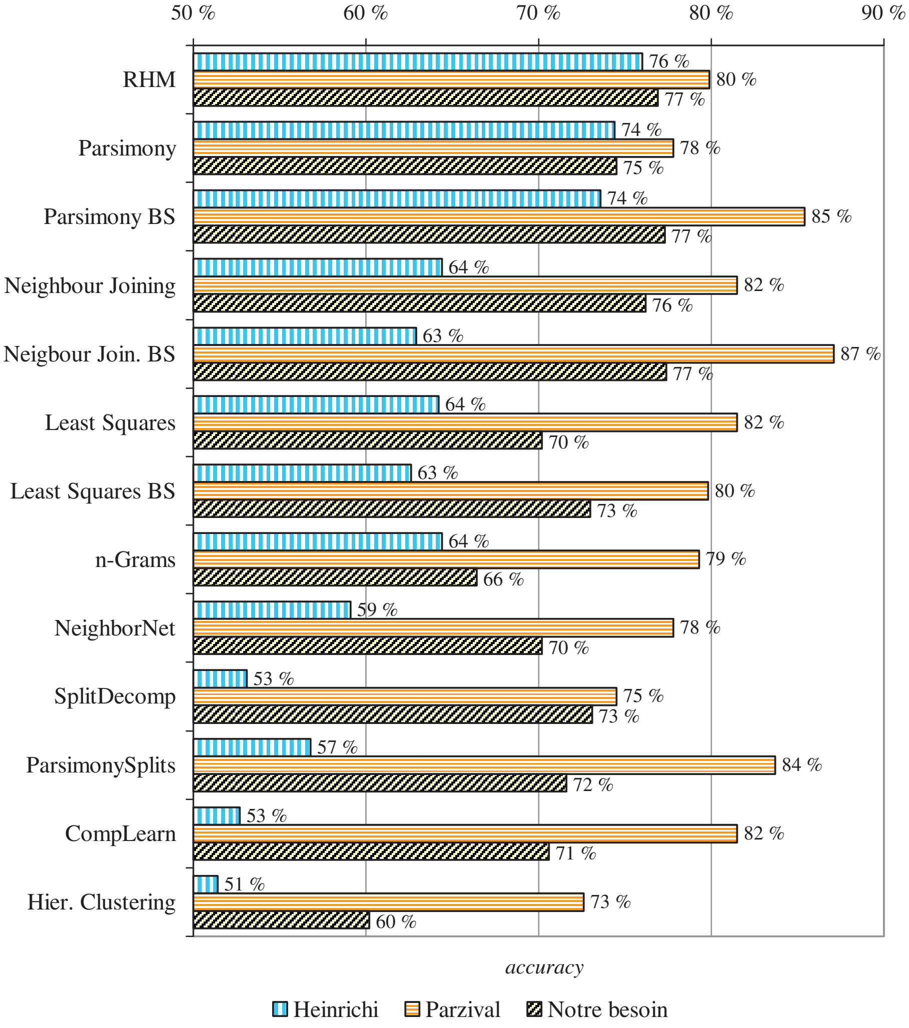


Fig. 2 The accuracies of Table 3 for methods for which the result is known for all three data sets (*Heinrichi, Parzival, Notre besoin*). Note that the *x*-axis starts at 50% in order to enhance readability; in practice no accuracies less than 50% are ever obtained

and seven other methods applied to the artificial *Notre besoin* data.[5] The performance of the thirteen first methods, for which the result is known in all three data sets, is also shown in Fig. 2. The overall conclusion is slightly mixed, different methods achieving the best score on different data sets, but in general a compression-based method, labelled RHM, and PAUP with the parsimony criterion achieve scores that are consistently near the best ones in all three cases. Here we focus on some interesting aspects of the solutions in order to understand when the methods work and when not, and why.

It should be stressed, however, that the comparison between the usefulness of the different methods is made solely based on their success in finding the correct stemmata. Other factors worth taking into consideration are the amount of data a method

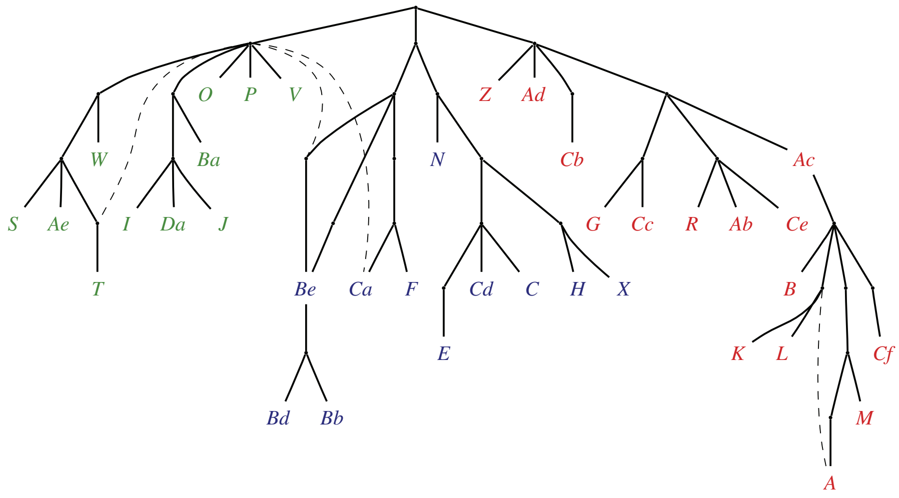


Fig. 3 The correct stemma of the artificial tradition *Heinrichi*. In case of multiple exemplars per copy, i.e. contamination, a dashed edge indicates the secondary exemplar. Colours emphasize the three main groups

needs to reconstruct a useful tree, on one hand, and the degree of preprocessing of the data, on the other. These are both aspects closely related to the usability of the different stemmatological methods. Due to the nature of the challenge, where the participants were only requested to analyse the given data sets as a whole, the study of the effect of the amount of data is left for future investigation.

None of the methods was able to reconstruct exactly the correct pedigree of the copies of a text in any of the cases—this would show as an accuracy value of 100% in Table 3. Still, all the methods marvelled when it came to reconstructing the account of witness relationships in *Parzival*, whereas the scrutiny of data set *Heinrichi* returned the most unreliable results. Some of the results obtained on *Heinrichi* were astonishingly unsatisfactory and could, from a textual scholar's point of view, hardly contribute to the study of the imaginary copying history and dissemination of the text. What is more, some of the methods achieved very different scores on the three data sets.

Both the overall tendency of the applied methods to get the best results of the very simple *Parzival* data set and the poorest of the most realistic data set *Heinrichi*, and the very different quality of the results on different data sets have clearly to do with the complexity of the three artificial traditions. The most obvious way to elucidate the degree of complexity of the pedigree of texts behind a data set is to look at the stumbling stone of the more traditional methods of textual criticism, namely the contamination between different textual versions. Whereas *Parzival* does not contain any contamination at all, 8% of the textual witnesses of *Notre besoin* and 11% of *Heinrichi* are results of copying from more than one exemplar. It is hardly surprising to discover that the degree of contamination and the number of missing manuscripts correlate with the quality of the results: the higher the number of contaminated or missing manuscripts, the poorer the results. Currently, contamination is still ignored in many computer-assisted methods, and it remains a real challenge for computer-assisted stemmatology—based on our experiments, this applies even to the methods in the SplitsTree4 software which produce non-tree-shaped stemmata. Promising attempts to tackle contamination in a specific setting (exemplar shift) have been made by Windram *et al.* (2005).

Even if the mentioned general tendency of getting poorer results in more difficult data sets is shown in all the methods of our challenge, RHM and PAUP show more consistency in their results of different sets of data. The outcome of their

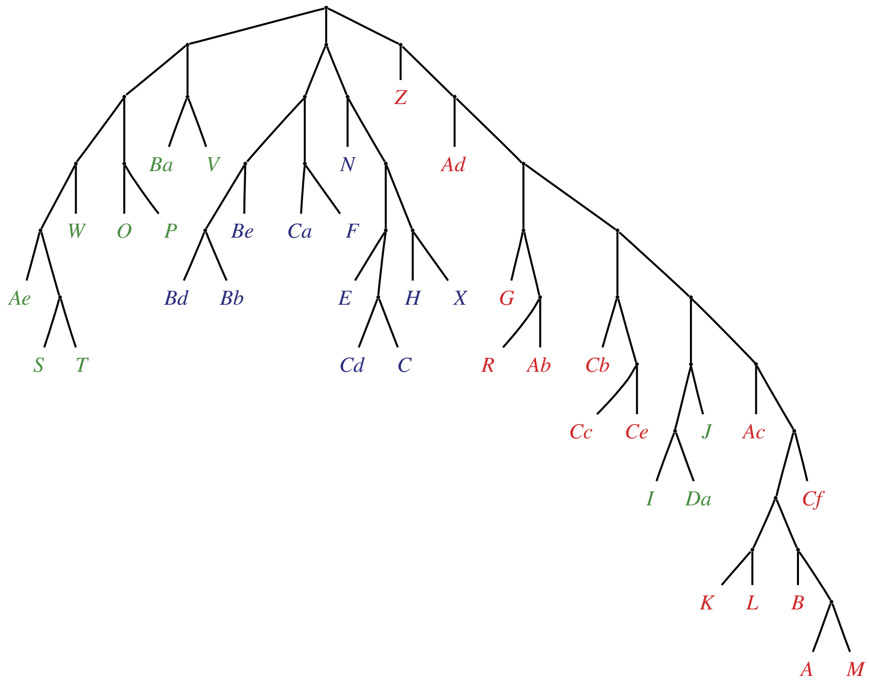


Fig. 4 The stemma obtained by the RHM compression-based method of Roos, Heikkilä, and Myllymäki (Roos *et al.*, 2006) for the artificial tradition *Heinrichi*. The tree was manually rooted in order to facilitate comparison to Fig. 3.

reconstruction of the stemma remains reliable in all three data sets.

Although none of the methods were able to recognize a textual witness with several exemplars, i.e. *contamination*, many of them managed to locate them surprisingly well within the stemma. More importantly, RHM and PAUP (using parsimony criterion) were able to identify nearly all groups of manuscripts correctly. This can be seen in Figs 3 and 4, which show the correct and reconstructed stemma for our primary data set, *Heinrichi*, respectively. The method recognized ten of the twelve groups of textual witnesses totally or almost correctly: Ae-S-T, O-P, Bb-Bd, Ca-F, C-Cd, H-X, Ab-R, Da-I-J, K-L, and A-M. Of the other two groups the close relationship between Ad and Z was recognized, but they were not understood as copies of the same version but of two versions descending from each other. In fact, there was only one group that was not identified at all: Cc-G, the versions of which RHM located rather far away from each other. As the method combines in its result Cc with Ce that was copied by the same scribe, the inability to identify Cc's and G's real context might have to do with the characteristics of the copies written by a certain copyist. Still, no such features could be identified in this scribe's work.

Some witnesses of the otherwise correctly identified groups went astray in the results of RHM. Version V, which was copied from the same exemplar as O and P, finds itself some steps further away, in the company of Ba—which is also rather similar to it. The failure to combine V with O and P is probably due to the omission of large passages of the text in V.

The identification of a group of versions normally goes hand in hand with the more or less correct localization of the group within the stemma. The group Da-I-J is an exception. All three versions were copied from an exemplar that contained lengthy passages written in Latin among the Finnish text. Whilst the version in which the Latin

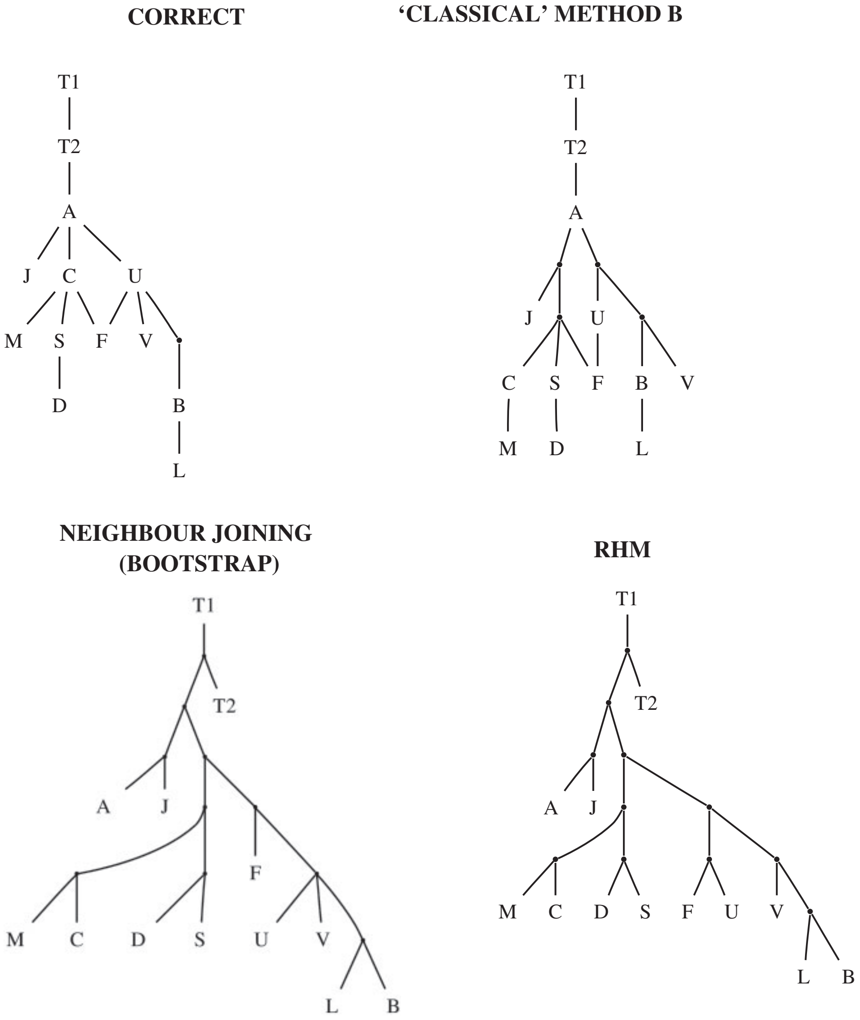


Fig. 5 The correct stemma (left) and three solutions for the *Notre besoin* artificial tradition. The manually constructed solution, labelled 'Classical' Method B (top-right), achieves the best average sign distance (85.1%) and is the only solution where contamination is identified; Neighbour Joining with bootstrap (bottom-left) is the best among computer-generated solutions (77.4%). The RHM tree (76.9%) is very similar. Trees were manually rooted in order to facilitate comparison with the correct solution.

passages where first introduced was not included in the material of the stemmatological study, all three of its descendants were combined by the Latin passages. Thus, it was easy to identify the group, but there was not much material to help with the locating of it within the stemma.

In the data set *Notre besoin*, one of the solutions, produced by J. Noret and labelled as 'Classical' Method B, achieves a remarkably high score, 85.1%. Figure 5 shows the correct stemma and the solutions obtained by 'Classical' Method B, and the second best method Neighbour Joining with bootstrap with score 77.4%, as well as RHM with the score 76.9%. In fact, the classical methods proposed by the philologists were the most successful in identifying contamination on the basis of the collation of the actual manuscripts, and evidently, the difference in performance between the classical method and the other methods is mainly due to this feature—even if the result of the 'Classical' Method B on the contaminated section of the stemma is not totally correct. Baret *et al.* do not explain in detail how the solution was obtained, but the method is *not* computer-generated and includes very much manual work. While in small data sets where contamination is an issue, a carefully hand-crafted solution is often the best one, it is out of the question for larger data sets.

6 Discussion

The outcome of our challenge points out RHM and PAUP (with parsimony criterion) as the computer-assisted methods returning consequently the most correct hypotheses according to the proposed numerical criterion. Therefore, the comparison between the results of RHM, PAUP and those of the methods used in previous tests on the same data set is well justified.

Two of the three artificial textual traditions used as material of the challenge have been used previously for similar comparisons of methods: *Parzival* by Spencer *et al.* (2004), and *Notre besoin* by Baret *et al.* (2006). The outcome of both experiments was positive, but it 'did not indicate strong superiority of one approach over the other' (Spencer *et al.*, 2004). The comparison between the results of Spencer *et al.* and Baret *et al.* on one hand, and our results, on the other, is somewhat hampered by the relative easiness of the *Parzival* and *Notre besoin* materials. Most real manuscript traditions are much more complicated. The limited amount of witnesses and contamination in the two artificial traditions prevents unambiguously distillation of the pros and cons of different methods applied in the experiments.

Nevertheless, there are certain observations that repeat in all three experiments. Running the average sign distance calculation on the previous solutions to the *Notre besoin* tradition, published in the article by Baret *et al.* (2006), gives very similar accuracy values for both neighbour joining and parsimony methods, see Table 3. The answers of RHM and PAUP (with parsimony criterion) are among the very best even when compared with a greater number of different methods. As the *Notre besoin* material is very limited, a 'Classical', i.e. manual, method of comparing the variants of the small number of witnesses succeeded in getting the most correct stemma. Still, the usefulness of a manual method correlates reversely with the complexity of the material: the higher the number of variants and witnesses of a given textual tradition, the less reliable the stemma obtained by using only manual methods gets.

The only earlier *numerical* results we are aware of are presented by Spencer *et al.* (2004). They measure so called partition distances and triplet symmetric differences for tree-shaped stemmata they obtained using neighbour joining and parsimony criterion for the *Parzival* material. No clear conclusion was obtained. Depending on whether bootstrapping was used, either neighbour joining or parsimony was found to be better. The single best result was achieved by neighbour joining with bootstrap. Our experiments, where a different scoring criterion (average sign distance) is used, confirm this in the sense that neighbour joining with bootstrap gives the best results in the two smaller data sets (*Parzival* and *Notre besoin*). However, RHM and the parsimony method of PAUP outperform neighbour joining in the more extensive *Heinrichi* tradition by a large margin.

It is interesting that PAUP—the method so successful in Peter Robinson's Textual Criticism Challenge already in 1991—manages to get better results than many of the other approaches even today. Generally, methods based on parsimony and neighbour joining algorithms seem to rank high across different data sets. The three data compression-based methods, RHM, CompLearn, and 'Data Compression' (see Table 3) differ strongly in their results, RHM being the most consistent performer among all computer-assisted methods, whereas the other two exhibit less impressive performance.

Naturally, the ranking of these three different principles and the methods based on them would require more tests on a number of artificial methods. Still, the outcome of our experiment of running the average sign distance calculation on a great variety of different methods indicates that the methods based on parsimony, neighbour joining and compression are all well worth developing further. As the vast majority of scholars interested in stemmatology and history of texts are not familiar with computer science and the different principles behind the methods, another aspect worth taking into consideration is the easiness of use of the different methods proved reliable.

As stated above, there are several methods that manage to identify textual versions or copies close to each other in a reliable way. This alone is of great help for textual scholars wishing to test or verify their hypotheses about the classification of copies of a text based on more traditional methods of textual criticism. However, the decisive step to locating of the groups within a stemma and trustworthy reconstruction of vast textual traditions on grounds of the remaining textual witnesses by computerized methods has not yet been taken. For instance, polytomies, the cases in which several copies were made of one single exemplar, have proven out to be difficult to detect for most of the proposed methods of computer-assisted stemmatology. On the other hand, the restriction to (only) bifurcating trees or stemmata has been a traditional stumbling stone of the textual criticism from the very beginning of the discipline, already before computers; see the groundbreaking article (Bedier, 1928).

Even the most successful methods applied in the Computer-Assisted Stemmatology Challenge, RHM and PAUP, share many of the same drawbacks, as do most other computerized methods. On one hand, in spite of some promising attempts like (Barbrook *et al.*, 1998; Spencer *et al.*, 2004; Windram *et al.*, 2005; Huson and Bryant, 2006) the methods still have great difficulties in recognizing contamination, the copying of an exemplar from multiple sources. On the other hand, they tend to situate the different textual witnesses always at the end of an edge. By doing this, i.e. by avoiding the locating of a text version as an internal node within the stemma, the methods create edges and nodes that may never really have existed in the real-life stemma. This is, however, a problem constantly present even in the classical non-computerized methods of textual criticism: it is very difficult, if not virtually impossible, to estimate the changes a copyist made when copying a text from another. Consequently, there is no real way of telling how many missing links there were between two close versions of the same text. The addition of disappeared exemplars to the stemma as a result of an educated guess to explain the differences between closely related textual witnesses is used as an emergency exit as commonly in classical textual criticism.

The tendency of most computerized methods to place witnesses in the leaf nodes has the disadvantage of making the resulted stemma even more complicated than necessary, as well as reproducing relatively close text versions further apart from each other. In addition, this feature of the methods based on parsimony, neighbour joining and compression may make the results obtained by using them seem less adequate than they actually are, when measuring the distances between stemmata to evaluate the quality of a proposed stemma. In fact, Spencer *et al.* (2004) argue that the internalization of certain witnesses to internal nodes seems to be an intractable problem, since by doing it one important aspect of the computer-assisted stemmatology may be lost, namely that of objectivity.

In this study, we have not paid attention to the length of the edges in the stemma. The length of an edge could indicate the amount of differences between two objects, and might thus be useful in deducting whether or not there could be lost copies hidden in the stemma, or in internalizing some of the versions (Spencer *et al.*, 2004). Furthermore, the stemmatological questions of textual scholarship are normally closely connected to editing a text, where special attention is paid to those passages that contain much variation. From a textual scholar's point of view, however, this—in spite of being an important issue—is often not quite as critical a problem as identification of the groups and their localization within the stemma, since passages containing much variation can often be identified relatively easily during the process of collation.

One of the fundamental differences between the modern computer-assisted stemmatological methods and the more traditional ways of textual

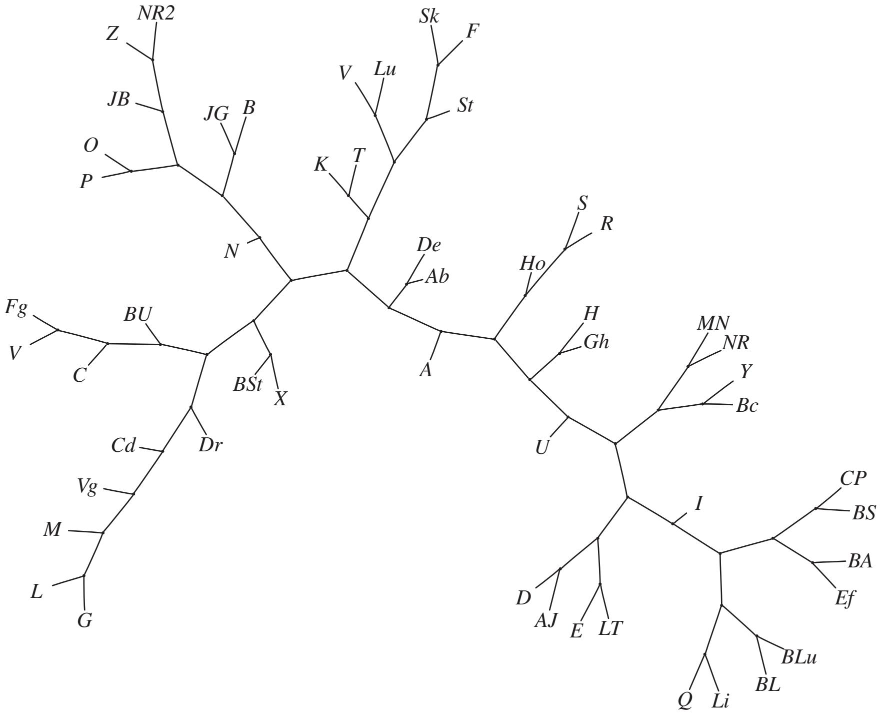


Fig. 6 The stemma of *Legend of St. Henry* obtained by RHM. The result is compatible with current understanding of the history of the legend based on external evidence not included in the textual content of the manuscripts

criticism lies in the form of the outcome. As the former provides a scholar with an undirected network of relationships, the latter gives a directed tree-shape figure with time dimension. The directing of a network is not always a problem: in most cases it is relatively easy to find the root of a network manually, knowing the textual contents of the witnesses. The adding of time dimension to the stemma may bring great advantages, but it may also endanger objectivity. In spite of the fact that the age of the carrier of a certain textual witness has only little to do with the relative age of the text version within a tradition, the manually executed textual criticism tends to overemphasize text versions whose physical contexts are the oldest among the material. Therefore, the more objective means of computer-assisted stemmatology that treat the witnesses only as texts without their physical context probably has great advantages in the first phase of the scholarly study of a complex tradition.

7 Epilogue: The Legend of St. Henry

What about the uses of the outcome of the challenge for real-life study of the textual traditions, the history and dissemination of texts? One of the tasks of the participants of the challenge was to reconstruct a stemma based on the real data set *Legend of St. Henry*. The results were divergent. Whereas neither of the hypotheses provided by the actual participants of the challenge was considered plausible or useful from the textual scholar's point of view, the outcome of RHM (Fig. 6) was very encouraging, indeed. Even if the correct answer is not known in a real set of data, the previous scholarship on the *Legend of St. Henry* has resulted in good knowledge of the groups of textual versions and a number of known relationships between the texts of different manuscripts. Therefore, it is possible to evaluate the plausibility of the proposed stemmata in spite of the fact that the whole true answer is not known. Previous studies have shown ten pairs of very closely related textual versions, on one hand, and a branch of the stemma consisting of a group of texts of nine manuscripts, on the other.

All ten pairs of versions known to be very closely related were chosen for the purpose of comparison: Q-Li; BL-BLu; MN-Y; E-LT; D-AJ; R-S; B-JG; JB-NR2; O-P; L-M. The submissions were able to get five or six of the ten checkpoints right, whereas RHM managed to find all of them, thus indicating the good reliability of the result. What is more, the hypothesis of RHM was able to find and locate a branch of nine different but closely related text versions known to exist in the stemma: Vg-M-L-G-Cd-Dr-C-Fg-V. Among the methods we have applied, at least PAUP (with parsimony criterion) achieves similar good results, identifying the same group of nine manuscripts, and almost all of the ten pairs (result not shown).

The success of the method in tracing the textual tradition of the artificial material of *Heinrichi*, and its evident success with the real *Legend of St. Henry* material give reason to optimism. In fact, as both data sets are very similar in nature—with a relatively high number of both extant and lost witnesses, much variations and modifications within the text, lots of very fragmentary manuscripts, and possibly many places of contamination—the good results of the method with the artificial material probably indicate that it gets rather correct answers of the real-life *Legend* material, as well. Due to the verifiably very good results in the artificial material and the plausibility of the answers got of the ascertainable parts of the stemma of the real data set, the RHM method has really been able to provide the scholarship on the *Legend of St. Henry* with a sound, justified hypothesis that can be examined in the future research of the text.

What is more, the knowledge of the mistakes the method makes in an artificial data set helps to identify and possibly correct the flaws in a stemma representing a real-life tradition. As an example, the outcome of the examination of the differences between the hypothetical stemmata and the correct solution of the artificial data suggests the internalization of a number of witnesses of the stemma obtained of the real *Legend of St. Henry* material. As it will thus be possible to aim the spotlight to certain aspects of the hypothetical stemmata of the real textual traditions, the time-consuming checking of the relations between members of a group of manuscripts, variant by variant, can be done to ensure the correct relationships between different parts of the stemma.

All the above-said opens up vast new perspectives for the future study of the text. As the *Legend of St. Henry* was the most influential text of the Finnish Middle Ages in many respects, one can hardly overestimate the significance of the outcome of applying the stemmatological methods to the material for the future scholarship. More importantly, this is but one example of the uses of the newest instrument in a textual scholar's tool box. Still, the use of the computer-assisted stemmatological methods does not leave the other, more traditional tools useless. The palaeographical study of the script of the texts, the codicology of the carriers of texts as physical objects, the knowledge of the history of the textual tradition, its writers and copyists, as well as the painstaking analysis of the textual contents itself have not lost their *raison d'être* but are rather given new additional roles.

Acknowledgements

We thank the volunteer scribes and others who helped create the data set *Heinrichi*, Caroline Macé for providing the data set *Notre besoin*, Matthew Spencer and Heather F. Windram for providing the data set *Parzival*, and the participants of the Computer-Assisted Stemmatology Challenge. We also thank the anonymous referees for thoughtful comments that have significantly improved the paper. This work has been supported in part by Emil Aaltonen Foundation (project 'Suomen keskiajan kirjallinen kulttuuri'), the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and the Academy of Finland (projects Minos, Kukot).

References

Barbrook, A. C., Howe, C. J., Blake, N., and Robinson P. (1998). The phylogeny of *the Canterbury tales*. *Nature*, CCCXCIV: 839.

Baret, P. V., Macé, C., and Robinson P. (2006). Testing Methods on an Artificially Created Textual Tradition. In Macé, C., Baret, P., Bozzi, A., and Cignoni, L. (eds), *The Evolution of Texts: Confronting Stemmatological and Genetical Methods*. Linguistica computazionale XXIV–XXV. Pisa – Roma: Istituti Editoriali e Poligrafici Internazionali. pp. 255–81.

Bédier, J. (1928). La Tradition Manuscrite du 'Lai de L'Ombre'. Réflexions sur l'Art d'Éditer les Anciens Textes. *Romania*, **54**: 161–196, 321–356.

Cilibrasi, R. and Vitanyi, P. M. B. (2005). Clustering by compression. *IEEE Transactions on Information Theory*, **51**(4): 1523–45.

Critchlow, D. E., Pearl, D. K., and Qian, C. (1986). The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology*, **45**(3): 323–34.

Dagerman, S. (1952). *Notre besoin de consolation est impossible à rassasier*. Paris: Actes Sud. (translated to French from Swedish by P. Bouquet).

von Eschenbach, W. (1980). *Parzival*. London: Penguin Books.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, Massachusetts: Sinauer Associates.

Grumbach, S. and Tahi, F. (1994). A new challenge for compression algorithms: genetic sequences. *Journal of Information Processing and Management*, **30**(6): 875–66.

Heikkilä, T. (2005). *Pyhän Henrikin legenda*. Helsinki: Finnish Literature Society.

Huson, D. H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, **23**(2): 254–67.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **2**: 241–54.

Li, M. and Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*, 2nd edn. New York: Springer Verlag.

Neovius, A. (1912). Akter och undersökningar rörande Finlands historia intill år 1401. *Historiallinen Arkisto*, XXIII, I, 3.

Notredame, C. (2007). Recent evolutions of multiple sequence alignment algorithms. *PLoS Computational Biology*, **3**(8): e123.

Pearl, J. (2007). Statistics and causal inference: A review. *TEST*, **12**(2): 281–345.

Robinson, P. and O'Hara, R. J. (1992). Report on the textual criticism challenge 1991. *Bryn Mawr Classical Review*, **3**(4): 331–7.

Roos, T., Heikkilä, T., and Myllymäki, P. (2006). A Compression-Based Method for Stemmatic Analysis. In Brewka, G., Coradeschi, S., Perini, A., and Traverso, P. (eds), *Proceedings of the 17th European Conference on Artificial Intelligence*. Amsterdam: IOS Press, pp. 805–6.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology Evolution*, **4**(4): 406–25.

Spencer, M., Davidson, E. A., Barbrook, A. C., and Howe, C. J. (2004). Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology*, **227**(4): 503–11.

Spencer, M., Wachtel, K., and Howe, C. J. (2002). The Greek Vorlage of the Syra Harclensis: a comparative study on method in exploring textual genealogy. *TC: A Journal of Biblical Textual Criticism*, 7. http://purl.org/TC (last accessed 28 February 2009).

Swofford, D. L. (2003). *PAUP*: Phylogenetic Analysis using Parsimony (*and other methods)*. Version 4, Sunderland, MA: Sinauer Associates.

Waterman, M. S. and Smith, T. F. (1978). On the similarity of dendrograms. *Journal of Theoretical Biology*, **73**(4): 789–800.

Wehner, S. (2007). Analyzing worms and network traffic using compression. *Journal of Computer Security*, **15**(3): 303–20.

Windram, H. F., Howe, C. J., and Spencer, M. (2005). The identification of exemplar change in the *Wife of Bath's Prologue* using the maximum chi-squared method. *Literary and Linguistic Computing*, **20**(2): 189–204.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, **24**(5): 530–36.

Notes

[1] The classical example of objecting automated and objective methods in stemmatology is provided by the wonderfully entitled article 'The Application of Thought to Textual Criticism' by A. E. Housman (*Proceedings of the Classical Association*, August 1921, Vol XVIII), where Housman fiercely argues that

'textual criticism is not a branch of mathematics, nor indeed an exact science at all.'

[2] See www.cs.helsinki.fi/teemu.roos/casc. All the material used in this paper is available for download in several formats (including plain text and the Nexus format compatible with PAUP*).

[3] Note that when a bifurcating tree, where each interior node has exactly three neighbours, is directed by choosing one of the nodes as root, then the tree branches in two at each branching point (except at the root, where there are three subtrees). For mostly practical reasons, bifurcating trees dominate the type of solutions that computer-assisted stemmatological methods produce.

[4] CompLearn is available as OpenSource software at www.complearn.org.

[5] The SplitDecomposition and NeighborNet methods were also applied to the *Notre Besoin* data already by Baret *et al.* (2006). Our results are very similar to what they report.

[6] The justification of bifurcating trees in phylogenetics is based on two assumptions: first, that genetic information of two or more species is not merged to create a new species, and secondly, that when a population constituting a species splits to distinct subpopulations, which will later become new species, the number of such subpopulations never more than two at a time (later the subpopulations can of course be further divided to more subpopulations).

[7] For instance, if the available variants are (*AAC*, *ABC*, *BBA*, *BBB*), where *A*,*B*,*C* are used in place of the segments, then the possible interior nodes are *AAA*, *AAB*, *AAC*, *ABA*, *ABB*, *ABC*, *BAA*, *BAB*, *BAC*, *BBA*, *BBB*, *BBC*.

Appendix: A Description of the RHM Method

The idea of the RHM method (Roos *et al.*, 2006) is the following. All variants are described, or *encoded*, by picking one of them as a tentative starting point, proceeding along the edges of the stemma tree to the tips of the branches, or the *leafs*, and describing each variant along they way given its already described predecessor. Having described the predecessor of a variant, the new variant can be described concisely if it resembles the predecessor. Hence, a stemma where similar variants are placed in neighbouring nodes gives a shorter code-length than a stemma where similar variants are randomly scattered across different branches.

In order to formally define the length of the encoding of a string given another string we need to choose a specific *code*. As is well-known, the universal code corresponding to Kolmogorov complexity is noncomputable, and moreover, it is defined only up to a constant which may be significant for short strings. In the spirit of a number of earlier authors—see e.g. (Grumbach and Tahi, 1994; Cilibrasi and Vitanyi, 2005; Wehner, 2007) and references therein—we approximate Kolmogorov complexity by using a compression program (gzip, based on the LZ78 algorithm). We also modify the gzip complexity by letting the complexity of a string given itself, $C(x|x)$, be zero for all strings x. It is also possible to ignore certain features known to be uninformative by replacing alternative forms by a standardized form; an example being the replacement of the symbol '&' by the word '*et*'.

For simplicity, and following the common practice in phylogenetics where it is perhaps better justified, we restrict the stemma to a bifurcating tree, i.e. a tree in which all interior nodes have exactly three neighbours.[6] Since in any realistic case, some of the manuscripts are missing, it is not reasonable to build a stemma consisting only of the surviving manuscripts. Instead, the remaining variants are all placed in the leaf nodes of the stemma, and the interior nodes are reserved for the missing variants. Note that even though some of the interior nodes may actually be available among the set of remaining variants, we can always imagine that those variants are duplicated so that the original text is lost and the copy is placed in a leaf node. Missing leaf nodes, i.e., missing variants with no surviving descendants can simply be ignored since they don't affect the analysis in any way. If the code-length of a pair, $C(x, y)$, is symmetric in the sense $C(x, y) = C(y, x)$, which is approximately true in our application, then the total code-length for any bifurcating tree G is given by

$$C(G) = \sum\nolimits_{(v,w)\in E(G)} C(v,w) - 2\sum\nolimits_{v\in VI(G)} C(v),$$

where $E(G)$ denotes the set of edges in G, and $VI(G)$ denotes the set of interior nodes in G. Hence the choice of the root node is irrelevant. In other words, the method gives no indication of the temporal order in the stemma. The question whether the order can be recovered from the texts alone, even in principle, is an intriguing open problem that touches on causal analysis, see (Pearl, 2007).

From an algorithmic point of view, the task of finding both a tree structure and the contents of the missing nodes is a daunting combinatorial optimization problem. Fortunately, given a tree structure, the optimal interior node contents minimizing the total code-length can be found in polynomial time in the number of nodes, under certain restrictions. More specifically, we compute the cost $C(v|pa(v, G))$ as a sum of the contributions of segments of ten–twenty consecutive words, and assume that the possible choices for the contents of each segment in the interior nodes are those appearing in the segment in question in at least one of the available variants.[7] This requires that the variants are aligned so that each segment corresponds to the same part of the text in all variants. To simplify notation, consider a fixed (directed) graph, and a fixed segment. Let the different versions of the segment in the available variants be denoted by $x_1, \ldots, x_m$. Under the restriction that $x_1, \ldots, x_m$ are the only possible choices, given the tree structure, the minimum achievable code-length can be evaluated using dynamic programming with the following recursion at the interior nodes, see (Felsenstein, 2004):

$$\begin{aligned}\text{cost}_i(j) = \min_k[C(x_k|x_j) + \text{cost}_a(k)]+ \\ min_l[C(x_l|x_j)+\text{cost}_b(l)],\end{aligned}$$

where a and b are the children of node i. The recursion is initialized at the leaf nodes by letting

$$\text{cost}_i(j) = \{0, \text{ if } x_j \text{ matches the content of node } i; \\ \infty \text{ otherwise.}$$

The total cost of the tree is obtained by summing over the segments the minimal costs

$$\min_j \text{ cost}_{\text{root}}(j)+C(x_j).$$

Assuming that computing the code-length $C(x_k|x_j)$ can be done in constant time for all k and j, the time-complexity of the algorithm is of order $O(knm^2)$, where n is the number of nodes, k is the number of segments, and m is the maximum number of different versions of a segment. In the worst case, all the versions of all segments differ, in which case we have $m = n$, and the time-complexity is of order $\Omega(kn^3)$.

With respect to the tree structure, the optimum cannot be found in closed form. The number of different bifurcating trees is super-exponential. Hence exhaustive search is infeasible, and no feasible alternative guaranteed to find the optimal tree is known. We use simulated annealing, accepting random modifications to the tree with probability

$$p = \min\{1, \exp(\text{total cost}_{\text{old}} - \text{total cost}_{\text{new}})/\text{T}\},$$

where T is a temperature parameter that is slowly decreased to zero. When evaluating the total cost, the algorithm also takes advantage of the fact that small modifications require only partial updating of the dynamic programming tables. With a large enough initial choice of T, the starting point in the tree search has no significance. We ran several runs up to 2.5 million iterations, each of which usually resulted in a very similar final tree structure and total cost.