

STUDIES IN STEMMATOLOGY

STUDIES IN STEMMATOLOGY

Edited by

PIETER VAN REENEN
MARGOT VAN MULKEN

With the assistance of

JANET DYK

JOHN BENJAMINS PUBLISHING COMPANY
AMSTERDAM/PHILADELPHIA



The paper used in this publication meets the minimum requirements
American National Standard for Information Sciences — Permanence
Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Studies in stemmatology / edited by Pieter van Reenen, Margot van Mulken : with the
assistance of Janet Dyk.

p. cm.

Includes index.

1. Manuscripts. 2. Manuscripts, Medieval. 3. Transmission of texts. I. Reenen, Pieter Th.
van. II. Mulken, Margot van. III. Dyk, Janet.

Z105.S78 1996

O91--dc20

ISBN 90 272 2153-7 (Eur.) / 1-55619-507-9 (US) (alk. paper)

96-665

CIP

© Copyright 1996 - John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any
other means, without written permission from the publisher.

John Benjamins Publishing Co. • P.O.Box 75577 • 1070 AN Amsterdam • The Netherlands
John Benjamins North America • P.O.Box 27519 • Philadelphia PA 19118-0519 • USA

Table of Contents

Prologue	vii
<i>Pieter van Reenen and Margot van Mulken</i>	

PART I: Methodological Approaches

Cladistics or the Resurrection of the Method of Lachmann	3
<i>Ben J.P. Salemans</i>	
Computer-Assisted Stemmatic Analysis and ‘Best-Text’ Historical Editing	71
<i>Peter M.W. Robinson</i>	
Shock Waves in Text Traditions	105
<i>Evert Wattel and Margot van Mulken</i>	
Clustering Stemmatological Trees	123
<i>Evert Wattel</i>	
Weighted Formal Support of a Pedigree	135
<i>Evert Wattel and Margot van Mulken</i>	
The Analysis of Early Islamic Traditions and Chains of Transmission .	169
<i>A.D. van Reenen</i>	

PART II: Subject-Related Observations

Medieval Dossiers and Modern Stemmas	209
<i>J.B. Voorbij</i>	
Textual Incompatibility and Many-Pronged Stemmata	233
<i>William R. Veder</i>	

PART III: Applications to Manuscript Traditions

Which Variants are Useful in Discovering the Deep Structure of the
Manuscript Tradition of a Text? 249
Karl-Heinz Uthemann

From Variant to Pedigree in the *Charroi de Nîmes* 263
Pieter van Reenen and Lene Schøsler

Index 305

Prologue

State of the Art

Progress in science may come from unexpected areas. An example of this is radiology, an auxiliary discipline of medicine: it has been accepted that recent developments in this discipline deserve to be taken seriously, and that several main disciplines in medicine, surgery for instance, can benefit from the progress made in this domain. In a similar manner, stemmatology may be regarded as an auxiliary discipline to text history; only recently have philologists come to the awareness that new developments in stemmatology may influence their findings in text history.

We consider stemmatology, the study of the relations between texts, as one of the two sciences basic to the study of older languages. The other basic discipline is the study of the linguistic variation found within and between texts, a study which concerns not only phonology and syntax, but also genre and the location in time and space of the language of these texts. Together these two disciplines are fundamental to text history. Disregarding them has made our understanding of older languages, impressive as it may be in many respects, seriously biased. Thanks to computer science this lacuna in our study of older languages is no longer necessary.

Stemmatology has always been regarded as an inevitable ‘malus’ in text history. Every self-respecting philologist consecrates the first chapter of his text edition to the text genealogy of the treated manuscript tradition, mostly observing that the tradition is extremely limited or, on the contrary, so complicated that any conclusions as to the kinship of the manuscripts are well-nigh impossible. Philologists usually restrict themselves to a brief résumé of all existing manuscripts, to which a short chronological, dialectological and narratological evaluation is added. In general, philologists do not believe that the outcome of stemmatological research will or can influence their choices of text edition.

This minimalist viewpoint is due to the evolution in stemmatology up to the pre-war period. Since Bédier’s pessimistic approach to stemmatology and his proposition to restrict oneself to the ‘best manuscript’, philologists

in the medieval field have often adopted this framework, discarding all methodological approaches to stemmatology. In the field of biblical or classical text criticism, scholars usually adhere to the neo-Lachmannian approach promoted by Maas. The latter renewed interest in this positivist method by systemising the detection and classification of variants, but he continued to stress the responsibilities and decisions of the philologist. In fact, in the neo-Lachmannian method the editorial and directional choices are always linked to the scholar's preconception of an archetype, and cannot be dictated by a fixed, verifiable, objective method.

Since the 1970s, however, new initiatives have been taken to renew interest in systematics in stemmatology. Especially in the Netherlands, philologists have tried to restructure research in the field of stemmatology. Thanks to Anthonij Dees, a theoretical revival of the traditional practice of text editor took place. Dees was the first to prove that the number of possible pedigrees is not infinite. Though the number may be enormous, increasing exponentially whenever a new manuscript is added to the collation, Dees has shown that Bédier was wrong in pronouncing the character of the stemmatological enterprise to be inconclusive. Further, Dees has pointed out that the vast number may be narrowed down by strictly discerning the distinct working stages, that is, the three levels (see below) in stemmatology. Dees' theoretical works have inspired many scholars. Since the 1990s, several colloquia have been organised in order to allow scholars to report on their findings in text genealogy and to familiarise them with the application of the newest technology to manuscript traditions.

These colloquia, held at the Vrije Universiteit in Amsterdam, have had a largely interdisciplinary character: scholars from classical, biblical, medieval and modern language studies have met with mathematical and computer scientists and a truly synergetic process has emerged. The results of these meetings are brought together in these proceedings.

In this volume, the papers are grouped according to three fields of interest:

- Part I. Methodological Approaches (six contributions)
- Part II. Subject-Related Observations (two contributions)
- Part III. Effective Applications to Authentic Manuscript Traditions (two contributions).

This volume may be considered as a working atelier, in which several artists agree to exhibit the state of their art. Functioning as a platform this volume gives each contributor the floor: he or she can explain what procedures are being followed, to what point the research has thus far progressed and why

the described choices have been made.

To novices in stemmatology we recommend starting with the articles on the application of stemmatology methods in Part III of this volume.

Computational Analysis

In cases of extensive manuscript traditions or highly contaminated traditions, scholars are nowadays ready to admit that their cognitive capacities are inadequate to encompass the complexity of the kinship relations in the text tradition. It comes as no surprise that those who wish to solve the problems of large and entangled manuscript traditions invoke the help of the computer science. Only recently, the synergy resulting from the collaboration of philologists and information scientists has produced relevant modifications in the methodological approach to stemmatology. All contributors to this volume agree that in order to detect the kinship relations among manuscripts three stages should be discerned: the unrooted deep structure, the underlying intermediate structure and the rooted oriented structure.

Embracing the coming of the computer science to the field of stemmatology implies that the consequences of this science have been thoroughly thought through. The advent of the computer is not only seen as a 'handy tool', a mere calculator or text editor; rather the implementation of the computer has fundamental theoretical implications. If the computer enables us to work with 'difficult' text traditions, the entire stemmatologic process has to be redesigned, and philologists have to learn to relativise their own decisions. The heightened awareness of the limitations of the researcher's own capacities and those of the computer may be considered as a true revolution: scholars are ready to work and live with uncertainty. The outcome of the stemmatologic enterprise, the stemma, is no longer seen as an authoritative prescriptive scheme which an editor should blindly apply to his manuscript tradition. None of the contributors to this volume wish to see the stemmatologic project as sacrosanct. The resulting pedigree, or, rather, the table of relationships, is but a guiding tool, though a necessary one to justify the decisions made by the editor.

All contributors further agree on the fact that in working with a manuscript tradition, every member of the tradition, even its worst representative, should be integrated. All available information should be made accessible and must be taken into account.

Lachmann Revisited?

In Part I of this volume, both Salemans and Robinson refer to Lachmann when discussing the state of the art in stemmatology. They testify to their indebtedness to this German scholar who was the first to point out the systemics in text genealogy. Lachmann outlined some basic principles:

1. some textual variants in manuscript traditions may betray kinship between texts;
2. only directional variants (common errors) can be used to build stemmas since manuscripts necessarily derive from an anterior archetype.

Both Salemans and Robinson stress the importance of the first Lachmannian principle for current stemmatology. Using cladistics, a recent (biological) ordering method, to draw text historical trees, they state that this taxonomic method can be seen as a modern version of the Lachmannian Method. Salemans and Robinson claim, fully in accordance with Lachmann, that only a few variants may be used to build stemmas.

It is the second Lachmannian principle, the directional character of variance, that Salemans and Robinson submit to reexamination. They work with 'unrooted' trees and show that Lachmannian subjective judgements to decide whether variants are directional or not are not necessary for the construction of trees, since only a few judgements on the originality of variants are required for the rooting of the tree. It is this aspect which exactly meets the design proposed by Wattel and Van Mulken in their second contribution to this volume: the determination of the unrooted deep structure of stemmas. The insights of Salemans, Wattel and Van Mulken, and many other text genealogists, agree with what may be called the main, commonly accepted, achievement within modern stemmatology since Lachmann: the acknowledgement of the dissection of the stemmatologic procedure: the detection of variants, the determination of undirected relationships and, finally, the orientation of the stemma with its editorial repercussions.

Salemans shows that scholars should be aware that the determination of variants remains a delicate process. If the processing of the data can be entirely confined to computer software, the focus of the scholar is still on the fixation of variants: like the traditional Lachmannians, Salemans is convinced that only a few variants are truly relationship revealing.

In the present contribution he demonstrates his automated text-genealogical, neo-Lachmannian or cladistic approach, using seven text

versions of Chrétien's *Yvain*. First, Salemans proposes five general, more or less abstract hypotheses which apply in general to relationship-revealing variants. Second, in the formalisation stage, these hypotheses are translated into eleven single-word oriented characteristics by which text-genealogically important variants can be recognised. Third, in the implementation stage, the computer is taught to recognise the eleven characteristics of variants automatically. The computer is thus enabled to perform the text genealogical theory. Fourth, in the evaluation stage, the computer results, and consequently the text genealogical theory, is evaluated; however, this evaluation presupposes perfect knowledge of the authentic language in the extant manuscripts, *in casu* the Old French dialects. Since Salemans does not possess such knowledge, he will comment on the evaluation stage when applying his approach to a field he is more familiar with, the fourteen versions of the Middle Dutch *Lanceloet van Denemerken*, the subject of his forthcoming dissertation.

Like Salemans, Robinson shows how stemmatologists can benefit from the progress made in cladistics; he, too, illustrates the usefulness of the PAUP-program. Although this program cannot cope with all typical stemmatological problems, such as contamination, it is possible to attest shifts in relationships easily. The combination of computer-aided text manipulation and the paleographic knowledge of the philologist gives a reasonable account of the existence of different underlying exemplars in the main representatives of Chaucer's *Wife of Bath*.

Robinson emphasises the usefulness of computer software in facilitating the difficult choices a text editor is confronted with when preparing an edition. The 'perfect' or 'ultimate' reconstruction of the original is nowadays considered an utopic enterprise: text editors are only too aware of the ephemeral status of their products. It is generally acknowledged that the entire project of text edition is primarily based on the highly interpretative character of manuscript transcription. Editors admit to having to make choices which should always be open to discussion. It is only logical, as Robinson states, that the editor should provide the next generation of text editors with as much available information as possible, thus allowing them to make their own choices and decisions.

It is in this context that Robinson advocates what may be called the 'cybernetic synoptic' edition, which, with the appearance of CD-I software, will become accessible to any scholar.

Graphs and Trees

Wattel, in collaboration with van Mulken, represents yet another direction in the newly recovered field of stemmatology. He shows that the problem of the 'virtual tree', as proposed by Dees, can indeed be solved by decomposing the stemmatologic procedure into several stages. Not only should the idea of unrooted deep and underlying intermediate structures be acknowledged and find a realisation in independent strategies, but also the complexity of large traditions can be reduced by systematically selecting representatives of bundles of manuscripts and by dissecting the phenomenon of contamination.

Wattel has tackled and partly resolved the two main problems with which the modern scholars are confronted, i.e., that of large quantities of manuscripts and that of the contamination. First, he has separated simultaneous contamination, resulting from the concurrent consultation of several sources, from successive contamination, the use of several sources consecutively. Wattel shows that the presence of successive contamination can be proven by the changes in relationships among manuscripts, detectable by sudden differences in distances between pairs of manuscripts. This sudden change may be caused by the scribes' separation of the quires of the underlying exemplar in order to facilitate the transcription.

Wattel's dissection strategy is characteristic of his approach: every logical stage in the stemmatological process can be studied as an independent nucleus. Every element may function on its own, every variable may be modified where deemed necessary and the process may be stopped at any desired moment, when the motivations for doing so are consistent with the overall strategy.

Wattel's stemmatologic enterprise mainly benefits from the recent achievements made in graph theory. In fact, a pedigree is "but a connected acyclic graph". Using this knowledge, the nodes and branches of this graph can be constructed by computing the distances, or, better, differences, between pairs of manuscripts. As in a cladistic procedure, a provisional tree is constructed, but thanks to the manipulation of quadruples (and in an ulterior stage, triples and doublets), and a careful evaluation of the weight of the information, the provisional tree may be calibrated. For intermediate nodes, branches may be contracted, thus optimally computing the information captured in the variants.

By simply offering a tool to extract the 'best' representative of a bundle of manuscripts, Wattel developed a procedure for processing large manuscript traditions. This representative is not chosen on the ground of

'best readings' but on the basis of the relative distances to other text families.

Though Wattel's procedures are highly intricate and especially devised for large manuscript traditions, his work also shows how a philologist may learn to distinguish theoretical choices from design choices. Traditional scholars tend to confuse their own unconscious design choices with methodological ideology and, therefore, are often unable to put their results based on extant manuscript traditions into perspective.

That stemmatology does not only apply to the texts but also to the transmitters of the texts is shown by A.D. van Reenen whose work has to do with the extremely hybrid tradition texts of the early Islamic period. Since the texts themselves offer little direct assistance in disentangling the miscellaneous relationships, the long chains of transmitters, originally added to the texts to give it a more authoritative status, may provide useful information on the internal relations between the texts. Though traditionally considered by Western scholars to be but untrustworthy data, these chains may prove informative when entered into an intelligently conceived relational database, which allows the comparison of large quantities of chains of transmitters. It is possible both to select the text variants and to retrieve a stemma of their chain with the added possibility of identifying the transmitters who are the 'nodes' of such a stemma. A long list of selection criteria can be used and tested, taking into consideration, e.g., the subject and literary form of text variants, the places and dates of birth and death of the 'nodes', and the text collection in which the variants were found. Candidates for (partial) authorship of texts, for example, concerning the prohibition of drinking wine, can be selected. As long as an open and flexible corpus is constructed with a comprehensive typology of texts and a transparent labelling, as free as possible from interpretative evaluations, the hybrid knot of chains of transmitters may be used to obtain accessory information on the affiliation of the texts.

Like Wattel, A.D. van Reenen uses the achievements of graph theory to gain insight into the relations among the transmitters of the words of the traditions and their position on the time scale. The possible links between pairs of transmitters often resembles the phenomenon of contamination in traditional manuscript traditions. By manipulating all possible relations between pairs of transmitters, by computing them, by extrapolation into a large chain, which only can be done with the help of the computer science, the philologist is able to motivate his editorial choices.

In summary, the disciplines of cladistics and graph theory are largely responsible for the new directions which stemmatology has taken. Apart

from these achievements, new insights in other adjacent fields of stemmatology may come to assist the text editor in making his choices.

Auxiliary Disciplines to a Preliminary Discipline?

Paleography, codicology and iconology are known to be as relevant to the determination of manuscript relations as the stemmatologic process itself. With the introduction of the concept of 'dossiers', Voorbij pleads for the usefulness of examining the composition of ancient volumes, in codicological terms as well as in terms of subject matter and content.

Voorbij analyses what may be called external variants by distinguishing not only the concept of dossier but also concepts such as volume, composite manuscript, manuscript and text. He shows that these concepts help to solve the crucial problems of the transmission of the manuscripts of the *Epistola Alexandri ad Aristotelem*, even before the analysis of the internal variant readings in the different texts.

The analysis of Voorbij results in a stemma of the structures and contents of manuscripts belonging to the family I of the *Epistola*. The success of this approach depends on the availability of unambiguous descriptions of the structures and contents of manuscript volumes. Introductions to editions should contain information on the transmission context of the texts.

While Voorbij emphasises the importance of extra-textual evidence for text genealogy, Veder underlines the necessity of intra-textual scrutiny on the basis of two text traditions: the *Scete Paterikon* and the treatise *On the Letters*. His explanation of the extraordinarily open character of the Old Church Slavic manuscript tradition takes into account the fact that this tradition developed in a period when written texts were conceived in a writing system other than Cyrillic, viz. Glagolitic, and that in this same period a totally different conception of text transmission was maintained: the users of the older alphabet allowed free adaptations of the text. Like A.D. van Reenen, Veder argues that respect for the text is a diachronically changing notion: a relatively short older period in which the text tradition exhibits considerable freedom is followed by a period in which the letter of the text is accepted as sacrosanct.

Another consequence of the change to the new spelling system was that the need for the copying and translation of the older manuscripts had suddenly increased. The resulting necessary collaboration of scribes had as effect that during the process exemplars were torn into quires and complete-

ly reorganised. Veder shows that during this stage, perhaps in one and the same scriptorium, all extant text families were born, generating in each of the two text traditions seven idiolectal versions, due to the differing capacities of the respective scribes. By transcribing the text into the new Cyrillic writing system, the scribes respected the authority of the spirit of the text, whereas in a later period the authority shifted to the letter of the text. Again, the phenomenon of contamination in the older phase can be explained by invoking the insights from an auxiliary discipline.

Methods put into Practice

That variants can have quite different ranks is discussed by Uthemann, who compares the neo-Lachmannian approach with the modern view in which three levels are distinguished. In discussing the correspondences and differences between the methods, Uthemann shows that though many drawbacks can be formulated against the Lachmannian point of view, the main advantage of this approach still lies in its focus on the critical notion of 'variant'. Like Robinson, Uthemann draws attention to the fact that the fundamental basis of any stemmatological procedure will always be the interpretative, intersubjective criteria which philologists apply when collecting the building blocks of the stemma: the variants. If a philologist does not critically evaluate the status of his variants, the entire building is subject to doubt. Not all variants can be accepted as relevant, according to Uthemann, especially intentional variants should be sifted out before further judgements on the stemma can be made.

In their contribution, P. van Reenen and Schøsler show that only in practice can methods in stemmatology prove their value. P. van Reenen and Schøsler apply new methods to an existing manuscript tradition, the *Charroi de Nîmes*, to arrive at a reasonable typology of variants. The *Charroi* tradition has proven to be a pilot manuscript tradition that may serve as a model to other applications of stemmatologic methods aimed at an evaluation of their usefulness. Like any other text tradition, the *Charroi* has, of course, its own characteristic deviations, anomalies that are almost inherent to text traditions and with which any method should be able to cope.

In this volume it has become apparent that a difference in medium (oral vs. written, Glagolitic vs. Cyrillic) is often correlated to a difference in the conception of the treatment of the text. Like Veder, P. van Reenen and Schøsler show that the extant contaminated character of the tradition

originated in an anterior stage when texts had an entirely different status. Direct access to this phase is no longer possible.

Conclusion

All contributors have clearly shown that they have confidence in the development of stemmatology. They acknowledge the status of the computer as processor, which a computer will always have, just as in radiology the X-rays remain the instrument; nonetheless, both the surgeon and the philologist are responsible for drawing their own conclusions. The more perfected the tools are that we have at our disposal, the more reasonable and defensible will be the decisions that we can make.

Contamination and large text traditions are the key words in this volume: thanks to recent developments, much progress can be made. Most of the contributors focus on the determination of the unrooted deep structure, and it is at this level that the computational analysis has booked its greatest successes. In the future, scholars will need to concentrate on the rooting procedures: the linking of the directional variants to a motivated orientation of the graph. Only then may we be able to assist in a resulting marriage of the neo-Lachmannian and the 'computerised' scholars.

The state of the art is always but the image of one particular moment: it is clear that stemmatology is in progress, still developing and very much 'alive and kicking'. Certainly in the Netherlands, the interdisciplinary approach has yielded fruit: the annual meeting of scholars interested in stemmatology at the Vrije Universiteit can be seen as a continuous workshop where new ideas can be tested and discussed.

The index to this volume is composed of the items chosen by the individual authors. No attempt has been made to standardise this list of entries for the entire volume. Authors' names occurring in the list of references following each article are not separately mentioned in the index.

Pieter van Reenen

Margot van Mulken

PART I

Methodological Approaches

Cladistics or the Resurrection of the Method of Lachmann

On Building the Stemma of *Yvain*

Ben J.P. Salemans

Press and Information Office

Universiteit van Nijmegen, Nijmegen, NL

0. Introduction and Summary

In 1991 I was presented with the challenge¹ of drawing the stemma of seven French text versions of Chrétien de Troyes' *Yvain* using my text-genealogical software. To that end the first thousand verses of these text versions were offered to me on a floppy. For a moment I hesitated: my knowledge of the medieval French dialects is poor and, initially, my software had been developed for Middle Dutch texts. Neither I nor my software can, for instance, judge the grammatical adequacy of the old French verses well enough, while my text-genealogical principles presume that good variants are part of grammatically correct textual environments. On the other hand, my principles seem to be more or less language independent. They assume, for instance, as we will see, that genealogically significant textual dif-

¹ Dedicated to my children, Bart, Milou and Joost, this text, the follow up of Salemans (1990), reflects the content of two lectures held at the 'Stemmatology' conferences at the Free University of Amsterdam, organised by Dr. M. van Mulken and Prof.Dr. P. van Reenen: the first two parts were presented at the conference of September 1991, and the third part in October 1992. I would like to thank both the organisers and the participants of the symposia who joined the discussions after the lectures. I am also grateful to Drs. R. de Bonth, Dr. P.A. Coppen, Prof.Dr. G. Dibbets, Dr. W. Ellis, Dr. W. Kuiper and Dr. Th. Mertens for reviewing a previous version of this text. Dr. D. Swofford, creator of the PAUP-software, and father Dr. A. Lee, pioneer in cladistic text genealogy, were kind enough to answer questions on PAUP and cladistics. Naturally, only I bear the responsibility for the text, which will be presented in my dissertation in adapted form.

ferences often involve verbs or substantives (by ‘substantives’ I do not mean ‘nouns’). I felt capable of recognising most of these categories in the *Yvain* texts. Curious to find out whether my principles could be applied to a non-Dutch corpus, I accepted the challenge.

In Salemans (1990) four text-genealogical principles were presented, whose backgrounds had been explained in two previous papers in Dutch (Salemans 1987, partly translated in Appendix D below; Salemans 1989). In the first part of this paper I discuss these principles in English, showing their importance and implications by using a passage from Dees *et al.* (1988-89). The creation of variation formulas, by which the shape of stemmas can be determined, is not as simple, undisputable or objective as some text genealogists think. *Only very few textual differences can serve as genealogical, relationship-revealing elements.* Therefore, a selected and theoretically justified choice among them is inevitable. Furthermore, two new *ad hoc* rules are formulated: one on word order and the other on text-genealogically significant word types. Finally, a new, fifth genealogical rule stresses the need of historical-philological knowledge for ascertaining whether variants are truly kinship revealing.

In the second part of this paper I describe briefly the software created to draw the genealogical tree of fourteen *Lanseloet van Denemerken* text versions, and discuss the application to the seven *Yvain* text versions of a part of this software, necessarily adapted for French texts. Eventually, twenty-four definite type-2 variation formulas are chosen. My first attempt to build the *Yvain* chain, using the algorithm in Salemans (1990:450), turned out to be less than successful. Most likely, this was due to contamination in the *Yvain* text versions. The algorithm advises the use the biological cladistic Wagner network method to build the *Yvain* tree.² Biological cladistics resembles a modern variant of the philological nineteenth century method of Lachmann. The latter, so often incorrectly disqualified text-genealogical method must benefit from cladistics.

In the third part I discuss and apply the cladistic approach for pursuing the text-genealogical relationship of the *Yvain* text versions. In close cooperation with Dr. W. Ellis, I tested the cladistic software package PAUP (Phylogenetic Analysis Using Parsimony), versions 2 and 3, on the twenty-

² See the English Summary of the Dutch paper Salemans (1987:223): “In this paper the basis of a new alternative text critical method has been laid: cladistic textual criticism. The method is deduced from the biological, taxonomic cladistic method of ordering elements (‘taxons’), which was first developed by Hennig.” The method of Wagner is described in Salemans (1987) (see English translation in Appendix D below).

four variation formulas. Recently, we also successfully tested MacClade version 3 (Maddison and Maddison 1992). In this paper I concentrate on PAUP-3 (Swofford 1991), an attractive tool for practicing cladistic textual criticism, with which it was possible to develop the genealogical *Yvain* tree as presented by Micha.

This paper demonstrates that a rather new approach in computerised text genealogy is possible and necessary, in contrast to the unscientific tendency of many 'computerised' text genealogists to develop text-genealogical software without a clear theoretical basis. My approach — perhaps new in philological fields, but more or less common practice in many other parts of science — is as follows. First, text-genealogical fundamentals or *theoretical principles* are formulated in plain English. Logically, these fundamentals should be understandable to a non-computerised audience. Second, in a *formalisation* process, the 'theory' — I consider my text-genealogical rules rather to be a kind of 'recipe' — is transformed into eleven concrete characteristics of text-genealogical variants. Third, these characteristics are transformed into computer software: a process of *implementation*. Then, the computer runs the software, taking text versions, in this case the *Yvain* texts, as input. In other words, the computer performs and tests the theoretical philological ideas. Finally, the analysis of the results should lead to Popperian scientific progress, i.e., the evaluation and, therefore, the refinement or rejection of the theoretical principles and characteristics. It will be clear that such an evaluation requires almost indisputable knowledge of many aspects of the texts.

Besides introducing and applying biological cladistic ideas and techniques successfully in text-genealogical fields, this paper also recommends that true philologists, 'computerised' or not, and the 'method of Lachman' should regain fundamental influence in text genealogy.

1. Theoretical Text-Genealogical Insights

Although a 'computerised' audience tends to be more interested in the results of developed text-genealogical software than in the concepts behind the software, it seems imperative to me to begin this paper on text genealogy with theoretical considerations. Mainly interested in computer-built chains and stemmas, modern text genealogists often do not want to be bothered with 'theoretical chitchat'. Such an attitude is understandable, but is too focussed on the end result, on the output. Do not fear: in the end, an *Yvain* tree will be presented in this paper; however, we must not be too impressed

by ingenious software that produces text-genealogical trees, for a three-year-old child can draw such trees as well. The latter statement is not intended as a trivial remark. The problem is not drawing a text-genealogical tree, with or without the computer, but developing a *trustworthy* text-genealogical tree. Text-genealogical software will draw text trees as prescribed by a text-genealogical theory or idea. Text-genealogical software, in fact, *is* a text-genealogical theory in computer shape. The computer acts as the performer and tester of the theory. When the software produces incorrect results, one or more elements in the theory must be inaccurate. The theory can then be rejected or refined. Logically, to discuss and compare stemmatological software and trees, one of the main goals of the Amsterdam symposia (see note 1 above), *is* to discuss and compare the theories behind them.

1.1. *The four basic text-genealogical rules*

1.1.1. *The determination of genealogical variants: the first rule*

The first text-genealogical basic rule:

A genealogical (or relationship-revealing) variant is a textual difference that fits well and inconspicuously in a text version.

(Appendix to this rule: Because a genealogical variant fits well in a text version, it is plausible that it keeps its form during the text transmission and will not be submitted to changes, except for small, trivial, differences in the spelling.)

A taxonomic ordering like a stemma or a chain³ is the description of the way the ordering tools present themselves in the taxons, *in casu* the text versions. The order or structure is a reflection of the ordering tools. When the tools are inadequate, the order produced by them is likely to be inadequate as well. Therefore, it is of utmost importance that text-genealogical methods critically consider the basis on which the ordering tools are chosen.

Text genealogists will often use textual differences, ‘variants’, as the tools by which the kinship of text versions can be discovered. The inexhaustible computer can help to detect all variants quickly. Yet, not all variants are *genealogical*, in the sense that they possess relationship-revealing powers. An editor of a text-critical edition will be interested in all variants, but text genealogists will mainly be interested in variants that reveal something about the kinship of the text versions.

³ A stemma is the family tree of the text versions. A chain is the deep structure of a stemma in which the originality of variants, i.e., textual differences, is unimportant.

According to the first rule, when dealing with variants we must ask ourselves whether it is likely or possible that one variant can be easily altered to produce other occurring variants. If so, these variants should not be used as text-genealogical tools. Only when it is plausible that a variant does not change into other variants, but has a fair chance of remaining the same, can we use it as an indicator of a relationship between text versions. In other words, only elements that fit well in the text have reasonable chances to 'survive' and obtain a stable place in a text version. Only such variants can serve for text-genealogical purposes. For instance, a copyist can or will rectify obviously incorrect words which are present in the text he copies, the so-called 'exemplar' (see Willis 1972:228) or 'layer'. Logically, grammatically incorrect text elements and clearly misspelled words cannot be used to build text-genealogical trees. This is demonstrated in fig. 1. Evidently incorrect elements attract attention and ask for correction, change or elimination. Small differences in spelling between variants, as expressed in the Appendix of the first rule, are unimportant as well.⁴ Genealogical variants must be nearly irreversible textual elements that fit well into the text.

The vague word 'plausible' in the Appendix of the first rule implies that every textual difference has to be evaluated before it can reach the status of genealogical variant. Some scientists will shudder at the idea of data needing to be 'evaluated'. They believe that judgements or interpretations are condemnable by definition. We will discuss this disapprobation in §1.1.2 more specifically; however, a few comments can be made here. First, a scientist's subjective judgements are perfectly allowable, when the scientist shows precisely when, and, preferably, on what grounds, he makes them, whereby subjectivity becomes *intersubjectivity*. Second, objective, undeniable text-genealogical facts to build text-genealogical trees simply do

⁴ The element "except for small differences in spelling" in the first rule is based on the fact that in the Middle Ages explicit, generally accepted orthographic conventions for the 'vulgar' (i.e., non-Latin) languages did not exist. Therefore, the same word form (say: 'Karel') can be presented by different copyists in different spelling shapes (say: 'Kaerel', 'Kairel'). These spelling differences are genealogically unimportant, because every copyist will recognise what is meant by the word in different spelling shapes and will usually adapt the spelling of that word into a shape which he prefers (say: 'Kairel'). In verse 45 of the *Yvain* texts (see fig. 3) three competing variants occur: 'pesa' versus 'greva' versus 'greua'. The small spelling difference between 'greva' and 'greua' is genealogically insignificant. Therefore, we may say that in fact two competing variants occur: 'pesa' versus 'greva|greua'. Nevertheless, we must be careful in using a vague category like 'small differences in spelling'. Tools are needed for determining whether variants consist of genealogically unimportant differences in spelling. For the Dutch medieval dialects these tools are the MNW (1885-1952) and MNHwb (1932, including the 1983 Supplement).

not exist. As explained above, not all textual differences between text versions offer text-genealogical information. A careful analysis of the variants is necessary. Third, judgements are not necessarily arbitrary: they can often be formulated in concrete rules or characteristics, as I will try to do. Moreover, to think that a theory has become completely objective once it has been formulated in computer terms is a misunderstanding. Subjective and *ad hoc* elements can be introduced and programmed into computer systems and software as well.

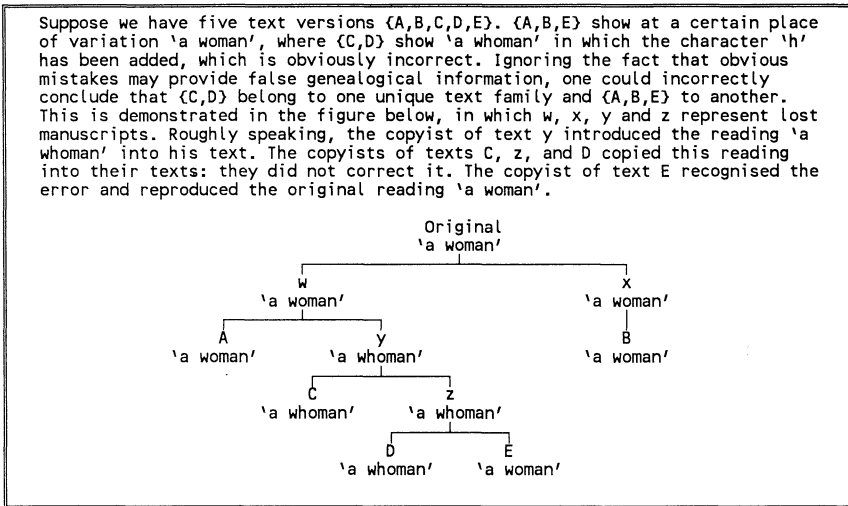


Figure 1. Consequences of not taking evident mistakes in texts into consideration

A good example of a variant type which is clearly not revealing kinship is the *parallelism*, also known as *homoplasy*, *coincident variation* or *accidental (variation)*. Parallelism is the phenomenon that the same variants show up in text versions spontaneously or by accident, while these text versions do not trace back to a common ancestral text.⁵ Logically, parallelisms do not offer

⁵ See Salemans (1987:202). I borrowed the notions 'parallelism' and 'homoplasy' from the biological sciences. I judge these terms to be more adequate than the rather vague term 'accidentals'. Additionally, in text-genealogical literature the term 'accidental' is usually used for 'small, inferior differences in spelling' or 'small differences in which the sense is not affected' (Dearing 1974:34). As we will see, these 'small differences' are just a part of the possible types of parallelism. Havet (1911: par. 543: 'Parallélisme') uses the notion 'parallelism' as well. This is a rather peculiar and illustrative example of the phenomenon 'parallelism'. Philologists familiar with the works of Havet could mistakenly assume a dependency of my ideas on those of Havet. Though Havet and I use the word 'parallelism'

reliable information on the relationship of text versions. The same goes for contamination (see fig. 2b), also known as 'conflation', 'text bastardy' and 'change in relationships' (see van Mulken 1993:111).⁶

Textual parallelism, as described in Salemans (1989),⁷ is shown in fig. 2a, which offers us artificial examples of parallelisms. Suppose an author has described, in a lost original text O, the white skin of his beloved. Six copyists transcribe it according to the pattern of the stemma in fig. 2a.

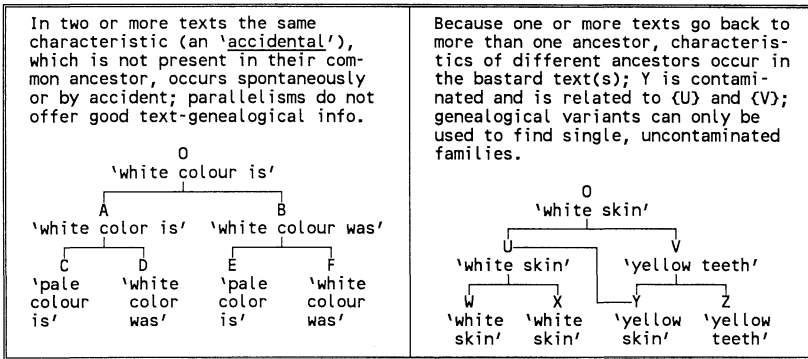


Figure 2a. Parallelism or homoplasy Figure 2b. Contamination

Several types of textual parallelism exist (see fig. 2a):

- a. *Synonymous parallelism* ('white' $\longleftrightarrow$ 'pale'). Text versions C and E show the adjective 'pale', while the other texts have 'white'. We must ask ourselves repeatedly whether it is possible that copyists spontaneously introduce in text versions more or less synonymous variants. A 'white face' can easily be altered into a 'pale face'. Like all parallelisms, synonymous parallelisms do not offer information with which to build genealogical trees. Were this to be neglected, it could be concluded incorrectly

in the same way, I am not influenced by him, for I used the term before becoming acquainted with parts of Havet (1911)⁵. The fact that we both use the word 'parallelism' is purely coincidental or parallelistic, and provides no information on a relationship between our works.

⁶ A contaminated text contains characteristics or variants that have been derived from several ancestral texts. Most current text-genealogical methods assume, as I do, that variants point to a single line of relationship. In fig. 2b the contaminated text version Y shows the variant 'skin', which points to a relationship to the U-family. The variant 'yellow' denotes other, contradicting family ties, i.e., to the V-family. At the moment no convincing solution for contamination is known.

⁷ See Appendix A.

that texts {C,E} must go back to one common text, which also has 'pale'. Synonymous parallelism not only concerns adjectives, but also nouns, conjunctions, pronouns, etc. Words within these categories can often be interchanged easily without much damage to the meaning of a sentence.⁸

- b. *Regional parallelism* ('color' $\longleftrightarrow$ 'colour'). 'Color' and 'colour' are not used in this example as trivial spelling differences,⁹ but as the American and the British English forms expressing the same word or idea. Every region has its own specific words and linguistic peculiarities. Copyists from the same region can introduce, independently of each other, the same words in their text copies. Sometimes they borrow words from other dialects.¹⁰ It is likely that an American copyist will write 'color' and an English copyist 'colour', no matter what reading the exemplar contains. The American texts {A,D,E} show 'color', not because they go back to the same ancestral text version, but because their copyists come from the same region.

A related type of parallelism is the *idiolectic parallelism*. Copyists have their own personal preferences in using a language. For instance, a copyist can prefer certain word orders, while other word orders are equally possible.¹¹

⁸ Compare the sentences: 'one can say ...', 'you can say ...' and 'we can say ...'. These sentences differ in the subject pronoun, but can have identical meaning. In the Dutch language many words are derived from other words by the use of affixes, like '-tje', '-schap', and prefixes, like 'be-', '-ge', '-her', '-ont' (ANS 1984, § 2.6 and § 8.5). Often the words with the added suffixes have a meaning which is closely related to that of the words from which they originate. This implies that the words can be interchanged rather easily. Logically, when Dutch text versions show suffix variants, we should be aware of the danger of synonymous parallelism (see also Appendix A, points j and k).

⁹ If 'color' and 'colour' were to be considered as trivial spelling differences, the problem of parallelism would not exist. The first rule states that spelling differences are ungenealogical, irrelevant.

¹⁰ Van der Wal (1992:121-122) remarks that texts often show characteristics of several dialects. Often copyists introduced, intentionally or unintentionally, characteristics of their dialects into texts. Furthermore, van der Wal quotes the famous Flemish medieval author Jacob van Maerlant to demonstrate that it was common practice that authors, in search of rhyming words, borrowed words from other dialects. Van Maerlant remarks in his hagiography, *St. Franciscus*, that for the cause of rhyme an author can take refuge in *misselike tonghe* 'different languages or dialects'. See also the remarks of van Mulken (1993:124-152) on the analysis of rhyming words.

¹¹ Consider the place of the auxiliary verb in modern Dutch subordinate clauses, which is dependent on the taste of the native user. It can be put at the end or near the end of the clause. The Dutch translation of the sentence 'he says that he has been punished' can be

- c. *Inflectional parallelism* ('is' $\longleftrightarrow$ 'was'). We know that the tenses of verbs, with or without an aspect auxiliary¹² can change rather easily, often even without changing the meaning of a statement (cf., for example, the *praesens historicum*). Therefore, for the moment, I assume that tenses of verbs are parallelistic. Analogously, substantives and other nouns occur in various numbers and cases. It seems that also these forms and cases do not offer trustworthy information on the kinship of text versions.¹³ To avoid the danger of parallelism, we choose to work with the basic, uninflected¹⁴ forms of substantives and verbs as genealogical informants. Once we have prepared the text-genealogical tree with nearly undisputable variants, we can check whether the hypothesis that tenses and inflections are parallelistic makes sense in practice or should be adjusted. Possibly it will turn out that the tenses of verbs are not parallelistic.¹⁵
- d. *Diachronic or historical parallelism* (not displayed in fig. 2a). This type of parallelism is caused by the development of languages throughout the ages. The lexicons of languages are not static but dynamic. Some words will become obsolete and will be replaced by new words. Independently from each other, copyists can replace the same old-fashioned words by the same new words.

The absence of small, highly frequent words in text versions is likely to be parallelistic, too.¹⁶

dependent on the taste of the native user. It can be put at the end or near the end of the clause. The Dutch translation of the sentence 'he says that he has been punished' can be 'hij zegt dat hij is gestraft' as well as 'hij zegt dat hij gestraft is'. Therefore, the place of the auxiliary in a subordinate clause in Dutch text versions is text-genealogically speaking not informative.

¹² In Dutch the auxiliaries of the perfect, the present and past perfect tenses, are 'hebben' and 'zijn'. The auxiliary of the future tense is 'zullen'. Van der Wal (1992:151) states that throughout the Middle Ages the frequency of the use of the perfect tense grew considerably. Logically, we could say that the use of the perfect tense in text versions is a 'diachronical' parallelism rather than an 'inflectional' parallelism.

¹³ Compare: 'Houses without a roof are unusual' and 'Houses without roofs are unusual'.

¹⁴ This is in agreement with Dees (1975:4): "abstraction est faite de variations graphiques et de différences de déclinaison casuelle".

¹⁵ In §1.2.1, just before the formulation of the final version of the fifth hypothetical rule, I will discuss the strategy that can be followed when there is any doubt on the text-genealogical relevance of a textual variant.

¹⁶ Dain (1949/1975:48): "Certaines fautes sont plus communes que d'autres. Il en est au moins deux que tous les copistes, sans exception, commettent couramment. C'est d'abord l'omission des petits mots. Les termes courts, et surtout ces petits mots qui constituent les

Parallelism emphasises the importance of the first text-genealogical rule in that parallelistic forms are by definition unstable. Thus, only non-parallelistic variants can be text-genealogically informative. Incorrectly, many text genealogists assume that all textual differences between versions, sometimes with the exception of small differences in spelling, can be used to draw text-genealogical trees. From the four types of parallelism presented we can see that huge non-spelling variants may offer us false information on the kinship of text versions. It could be argued now that parallelisms will not occur frequently in text versions, and, therefore, could be filtered out by statistical or mathematical analysis, like cluster analysis. In the third part of this paper we will see that for text genealogy this is a dubious approach. Furthermore, I would like to express here explicitly that the notion ‘parallelism’ is not a strictly Lachmannian notion.¹⁷ This notion was introduced in Salemans (1987:202), on the basis of a study of commonly accepted biological and taxonomic literature (see also note 5 above). Finally, it should be noted that it may indeed turn out that a parallelism offers genealogical information that is not in conflict with the true historical, genealogical relationship of text versions. The point is that parallelisms can offer false information and, therefore, they are not a reliable source for the construction of genealogies.

1.1.2. *The presentation of the used variants: the second rule*

The second text-genealogical basic rule:

(Genealogical) variants used for the construction of a chain or stemma must be presented in an apparatus of variants or in a synoptic edition, by which the stemma and the variants can be checked and falsified.

This rule says that we have a right to see from which observable, ‘positive’,¹⁸ elements in the text versions a text-genealogical tree has been

« utilités » de la phrase — liaisons, conjonctions, particules, prépositions — ont une tendance étonnante à disparaître sous la plume des copistes, au moins quand il s’agit des langues anciennes. Le verbe « être » et, chose à peine croyable, les négations n’échappent pas à cette règle. (...) Un second type constant d’erreurs de copie, depuis longtemps étudié, est *le saut du même au même*. C’est de beaucoup la faute la plus commune.”

¹⁷ Compare van Mulken (1993:29): “Parallelism, as a tool for deciding the kinship-revealing qualities of variants, includes the notion of ‘piège à copistes’. This is a typical Lachmannian notion: it implies a hierarchical order of the readings. At an earlier stage of the text transmission, a manuscript presented a reading which almost inevitably forced scribes to make a mistake, the variant conditioning the further transmission like a scribal trap.”

¹⁸ Dom Quentin (see Quentin 1926) and his ‘successor’ Zarri (see Zarri 1977) developed a

developed. All readers interested in text genealogy, who, like traditional 'Lachmannians', draw stemmas without the use of the computer, must be able to check how the trees are produced. I realise that genealogical software may produce an enormous apparatus of variants that cannot be completely printed in publications. In that case at least a part of the apparatus should be printed in order to give the readers an impression of how the stemma or chain has been built. Another possibility is to put the complete apparatus on a server computer of, for instance, a university library, connected to world-wide computer networks like the Internet. With file transfer ('ftp'), users can copy the apparatus to their own computers and read it.

The second rule is not a strictly text-genealogical rule. I consider it to be a normal scientific principle, here formulated because it is often violated by 'computerised' text genealogists, like, for instance, Dees.¹⁹ Trees are too often presented without disclosing the fundamental apparatus of genealogical variants; variation formulas are presented as though these formulas were obvious objective facts. In the formulas, the text versions are clustered into groups that show the same variants. Because these text genealogists do not offer the variants with which the formulas are composed, it is impossible to check their adequacy.

This disregard of variants is caused by the fact that in 'modern' text genealogy the influence of non-philological, statistical or mathematical methods has grown drastically, almost beyond philological control. Up to the sixties, text genealogy generally belonged to the philological branch of

method to draw chains by using so-called 'intermediate text versions'. A text version Y is an intermediate of the text versions X and Z, when X and Z together show no differences with Y. i.e., Y contains no variants which are not present in X as well as in Z. This is also called the 'zéro caractéristique' of the intermediate. Suppose that X shows the variants {a,b,c,d,e} and Z {x,y,a,b,c}. When Y is the intermediate between X and Z, it must contain all variants which occur in both X and Z, namely {a,b,c}. Logically, when Y contains, for instance, the variants {x,a,b,c,d}, it can be the intermediate between X and Z. In tracing the intermediate text versions, the chain, the deep structure of the stemma, becomes clear. I consider the basis of the construction of the chain by intermediates to be too passive or non-positive: the detection of a non-positive 'zéro caractéristique', namely, the absence of indications that a text version is not an intermediate text version, is used as a positive sign or tool for the determination of genealogical relations. When a text genealogist does not find such differences, it may be that his or her text comparison is inadequate. To quote Dekker (1987:64; translation BS), an intermediatist with his own interpretation of intermediacy: "When determining intermediacy, we will always have to keep in mind that the absence of indications (c.g. enough indications) that a manuscript is not intermediate does not guarantee that the manuscript actually is intermediate." Text trees preferably are composed from concrete, determinable, 'positive' elements in text versions.

¹⁹ See §1.1.5 below and References.

science. Philologists were focussed on the question of which variants in text versions could provide text-genealogical information. The fundamental idea of the influential, old text-genealogical method of Lachmann (see Maas 1957) was that text versions with the same non-original variants go back to the same common ancestor. With the knowledge of some of these so-called 'common errors' a stemma could be built. The problem was and is that it is hard to find out whether a variant is a common error. Then, some decades ago, the importance and influence of the mathematically oriented natural sciences grew enormously. Within the humanities more and more scientists fell back on the 'empirical' paradigm.²⁰ In this paradigm the importance of gathering and ordering little facts is emphasised upon. Often theories, hypotheses and judgements were regarded to have an inferior, dangerous and unscientific character. Probably due to the strong general admiration of the humanities for the methodology and requirements of the natural sciences (e.g., prediction and testing), many members of the humanities chose for the empirical methods of the natural sciences to legitimise the scientific character of their own branch of science. Textual criticism with its hypotheses and subjective judgements, like Lachmannian errors, were generally judged by these 'modern' text genealogists to be unscientific. Textual criticism, including the method of Lachmann, had to be reformed into a new empirical part of science.

The first results were promising. It was discovered that stemmas could be built in two steps, an idea now accepted by most modern text genealogists.²¹ The first step is to build a chain with the aid of genealogical variants, without judging their (un)originality. The second step is the transformation of the chain into a stemma, using the knowledge of the originality of only a few variants. This implied that the philological judgement on variants as text-genealogical informants seemed to have become irre-

²⁰ The philosopher Kuhn formulated the philosophy of the paradigm. For a long period of time, researchers within a scientific discipline would build, intentionally or unintentionally, their works on one central slumbering but undisputed framework of thinking, a paradigm. See van Buuren (1988:34-40), and Salemans and de Bonth (1990-91:210).

²¹ A third step, the so-called 'input step', concerns the development of variation formulas from the variants. I perform the input step before the chain-building step. First, I try to find significant, text-genealogical, variants and build variation formulas from them. With these formulas the chain can be created with which the stemma can be developed. A stemmatologist like Dees performs the input step later. Roughly speaking, he first draws all possible chains for a certain amount of text versions. He then performs the input-step and scratches off all textual 'constellations' that are not in accordance with the variation formulas created in the input step.

levant. Modern text genealogy started to focus on developing methods to create textual trees with a given amount of 'unjudged' variants. Generally, all observable textual differences, except maybe for small spelling differences, were considered to be sound text-genealogical variants. Omnipotent mathematical, computerised principles could build text trees from nearly any set of variants.

It is time that philologists regained fundamental influence in text genealogy. Duinhoven (1988-89), for instance, has shown that 'modern' text genealogy has produced impossible stemmas, due to the incorrect use of variants. I do not claim here that all modern empirical or statistical text-genealogical methods are useless, nor do I say that empirical sciences are to blame for the current identity crisis of text genealogy; however, text-genealogical trees have to be built with strictly text-genealogical building elements. Historical or relational trees require historical or relational building blocks. This is the fundamental idea behind all of the text-genealogical concepts and rules expressed in this paper. Only after philologists have developed clear ideas or methods to detect informative variants, they can request their 'empirical' colleagues to produce trees from the selected variants, if necessary.

1.1.3. *The width of the place of variation: the third rule*

The third text-genealogical basic rule:

A place of variation is that part of a text in which the extant text versions show one or more different ('competing') variants, which are formed by at least one word. The place of variation must be kept as small as possible to avoid the danger that it contains several genealogical variants introduced during independent stages of textual transmission.

This rule says that it is preferable to use single words²² in order to be on the safe side in choosing good variants and variant places. A syntactic unit will often contain more than one word. These words can be different genealogical variants, introduced at independent stages of textual transmission.

²² I will not define here the complex concept 'word' in depth. I consider it to be the smallest, independent unit of meaning in a language. For modern Dutch, the spelling of words is regulated by law. Apart from a few problems, modern native speakers will recognise the same words in a sentence. Until about 1600 spelling prescriptions in 'orthographies' did not exist, nor did a universal Dutch language. We see that one copyist writes one word (e.g. 'opstont', 'euerzwyn'), where the other copyist writes two words (e.g. 'op stont', 'euer zwyn'). It is, thus, difficult to define the notion (written) 'word' in old Dutch texts. I consider the differences in word boundaries to be potential parallelisms.

Therefore, I choose, as a principle, single words as places of variation rather than syntactic units.

My fundamental choice for single words seems to be in conflict with two other text-genealogical concepts involving word order and the omission or addition of complete verses.

First, I think that different word orders can offer genealogical information. Naturally, these word orders have to be grammatical, as a result of the first genealogical rule. The problem is that a word order involves more than one word, thus breaking the single word boundary of the third rule. I solved this problem with a flexible definition of 'word order' at word level as 'a left to right enumeration of single words'. Now it is possible to formulate:

The (tentative) first *ad hoc* rule or hypothesis on word order:

When two or more related (parts of) sentences or verses in different text versions show the same, preferably three or more, words but in a different, yet syntactically adequate, order, the places in the text versions with the different order of words may be considered as genealogically informative places of variation; the word classes to which these words belong are unimportant.

By definition, places of variation with a change in word order must have a length of at least two words. While discussing the idiolectic parallelism, we saw that the place of auxiliaries in Dutch in subordinated clauses is rather arbitrary (see note 11 above), offering no genealogical information. Therefore, it is preferable that a place of variation with a change in word order is greater than two words.

According to Dr. M. van Mulken, with whom I discussed the tentative rule on word order several times, a change in word order can be a regional or a diachronic, historical parallelism.²³ Presently, I agree with her criticism when the change in word order concerns adverbs. A superficial study of some old Dutch texts²⁴ convinced me that adverbial adjuncts, built from single adverbs that denote place, time, etc., may take a rather free position in a sentence. Logically, when a difference in a word order merely consists of such adverbs, this can be a text-genealogically uninformative parallelism. Hopefully, in the near future, using, among others, the stemma of the *Lanseloet van Denemerken* text versions, we will be able to determine whether adverbs (or other word categories?) may offer genealogical infor-

²³ See van Mulken (1993:28): "... Salemans assigns changes in word order too quickly to the genealogically significant category ('heavy weight variants'), since changes in word order can well be the result of dialectical translation."

²⁴ See Salemans and Wackers (1986); Salemans and Schaars (1987-...; 1990; 1993).

mation that is now presumed to be false. First, I will try to build a *Lanseloet van Denemerken* tree from (nearly) indisputably correct variants. Then I will check whether changes in word order are in accordance or in conflict with the determined relationships of the text versions. Possibly, we will find out that changes in word order are parallelistic indeed. For the time being, I think it is good to add at the end of the tentative rule a restriction for adverbs:

The (final) first *ad hoc* rule or hypothesis on word order:

When two or more related (parts of) sentences or verses in different text versions show the same, preferably three or more, words but in a different, yet syntactically adequate, order, the places in the text versions with the different order of words may be considered as genealogically informative places of variation; the word classes to which these words belong are unimportant, provided that they are not adverbs.

There is a second, more complex, problem connected to my fundamental choice for single words as text-genealogical units: I am convinced that omissions or additions²⁵ ('interpolations') can sometimes possess relationship-revealing power. Can we say that text versions, showing the same 'hole' or extra word(s), are genealogically related? Is an empty place in texts a variant?

Currently, I do not see clearly how a 'hole' or 'empty place' can be defined in terms of single words. I am convinced that variants must be 'positive', observable.²⁶ This implies that one group of text versions contains one word, while the other group of text versions have another word. Variants must have a content and, therefore, cannot be empty. One could state now that a hole in a text version has to be considered as a 'zero word', a word without a content; however, this is not an accurate solution. I fear

²⁵ For the development of a chain, the difference between omission and addition is not important.

²⁶ See also note 18 above. Suppose that we have seven text versions A to G and that their first sentences are 'And this is the first sentence' in texts {A,B,C} and 'This is the first sentence' in texts {D,E,F,G}. At first sight this seems to result in a type-2 variation formula ABC|DEFG: the first three texts with 'And' versus the other texts without 'And'. The 'zero spot' without 'And' in the four texts is, however, not observable or 'positive'. It is, for instance, possible that an ancestor text of texts D and E had 'So, this is the first sentence', while an ancestor text of F and G had 'Obviously, this is the first sentence'. The independent omission of the words 'So' and 'Obviously' resulted in the same line for texts {D,E,F,G}, while these texts do not necessarily go back to the same common ancestor. Therefore, the correctness of the type-2 variation ABC|DEFG is doubtful.

that empty words in text versions can be parallelisms, introduced at different stages in the transmission of a text. To put it differently, one empty word is not necessarily equal to another empty word. The fact that text versions show the same omitted words does not necessarily imply that they have a direct text-genealogical relationship. To detect whether the addition or omission of words in the *Lanseloet van Denemerken* text versions can be parallelistic, I will follow the same procedure mentioned above: I will build the text genealogical tree from sound variants and then I will check whether omissions or additions of words are in accordance with the determined relationships of the texts.

We see that the use of omitted words as text-genealogical informants is tricky. Can the same be said of omitted (epic) verses? Obviously, the omission and interpolation of verses break the single word boundary of the third rule; nevertheless, I think that omitted or interpolated verses can offer us important text-genealogical information. Assuming that the verses in the original text all rhymed and that single, non-rhyming verses did not occur,²⁷ when we are confronted with a non-rhyming verse, it is likely that we have spotted an 'error' introduced during the transmission of the text.

Suppose that the rhyming scheme looks like 'a-a-b-c-c-d-d'. The third verse, with 'b', does not rhyme, while the verses in its direct neighbourhood do. One possible explanation is that a second verse with 'b', possibly accompanied by verses 'x-x- etc.', has been deleted during the transmission of the text. Another explanation is that the 'b'-verse (and accompanying verses) has been added, interpolated. Another possible situation is that the rhyming scheme is like 'a-a-b-c-d-d-e-e': the third and fourth verse do not rhyme. In this case it is not necessary that the deletion or interpolation of a verse did occur. It is, for instance, possible that a rhyming word of one of the verses concerned has been deleted or took another place in the verses. Other situations could be suggested as well. Presently, I only want to propose that if a break in the rhyming scheme can easily be repaired, thus implying a philological judgement, the break has no text-genealogical value. When it cannot be repaired easily, it does have.

²⁷ It is possible that a complete text version W is copied by a printer or copyist who produced a complete text version X as a copy of W. Then, to rebind W, a bookbinder cuts the pages of W too small, which causes the loss of some first and last lines of pages in W. Then W is used by another copyist to produce a text version Y. Y is copied into Z. Y and Z show the same loss of certain text lines. This loss does not occur in W. Nevertheless, X, Y and Z go back to the same ancestral text W. Thus, we should treat with caution the loss of text lines as an indicator for kinship.

The same can be argued with respect to the inversion of two rhyming verses. We conclude that the omission or interpolation or inversion of complete verses can sometimes be genealogically relevant.

My genealogical concepts are, partly, verse-oriented. This is due to the fact that the text versions of *Lanseloet van Denemerken*, the subject of my dissertation, are all in verses. The advantage of working with rhyming verses instead of prose lines is that the rhyming words cannot be replaced by other words as easily as non-rhyming words in prose texts. When a copyist changes one rhyming word into a new word, he will also have to find a new rhyming word in the other verse. Rhyming words seem to take stable positions and are fundamental key words in verses. The fact that verses generally seem to have a more stable shape than prose line is convenient for text-genealogical research.

1.1.4. *The use of type-2 variations: the fourth rule*

The fourth text-genealogical basic rule (or the 'type-2 limitation'):

Only if all text versions show at a place of variation exactly two genealogically significant variants, and if each variant does occur in at least two text versions, can these variants be directly used for the determination of the structure of the stemma (the 'chain'). This fundamental variation is called a 'type-2 variation'.

A type-2 variation occurs when a place of variation shows exactly two competing genealogical variants that occur in exactly two true groups²⁸ of text versions. (As is mentioned in Salemans [1990:463], a variation is the phenomenon that text versions show variants at places of variation; a variation formula is the notation or symbolic expression of a variation.) The type-2 approach requires the presence of at least four text versions.²⁹ The two true groups in a type-2 variation are called *end groups*, because they divide the chain into two groups situated at both ends or sides of the chain.³⁰

²⁸ A true group of text versions consists of two or more text versions showing the same variant. In Salemans (1990:446-447) several variation types are discussed, using five text versions A, B, C, D and E. In simple variations, text versions show two variants: A|BCDE (= type-1, with one true group), AB|CDE (= type-2, with precisely two opposing true groups); in complex variations text versions show more than two variants: AB|C|D|E (= type-3, with one true group), AB|CD|E (= type-4, with two or more true groups).

²⁹ A 'quadruple' is a type-2 variation formula with the minimum amount of four text versions.

³⁰ To express it differently, an end group in a chain is formed by the complete group of text versions all of which are connected to the rest of the text versions by one single line in that

Salemans (1990) explains why true groups in type-2 variations offer direct information about the chain, which is the deep structure of a stemma. It is claimed there that complex variations, with three or more competing variants, can only be used for orienting the chain into a stemma.³¹

The fourth rule is in accordance with Greg (1927).³² Dearing (1974)³³ claims that new type-2 variation formulas can sometimes be deduced from complex variation formulas. Salemans (1990) demonstrates that Dearing's interesting approach can unfortunately lead to incorrect results; therefore, these manipulations have to be considered with scepticism. Mind the for-

chain. The end group can be determined by cutting the chain into two pieces at a point on that single line.

³¹ Salemans (1987:211-220, to be found in translation in Appendix D, §3.3 to §3.3.3, below) and Salemans (1990:438-442) explain that chains of taxons can be built when the transformation order of the character states is known. The determination of this transformation order can be troublesome, too, unless binary character states or groups of variants are strictly used. In that case the transformation order is uncomplicated and clear. In textual variations binary groups of variations occur in so-called 'simple variations' or 'type-1 and type-2 variations'. Since type-1 variations are genealogically unimportant, it is the best to use type-2 variations to build chains. This is the motive behind the fourth genealogical rule. The process of producing a stemma from a chain is delicate. Dearing (1974:44-56) mentions that the following variations are 'directional', offering information on how a chain has to be oriented: confusion of similar letters, misinterpretation of contractions, mistaken combination or separation of words, transposition ('anagrammatism' or 'metathesis'), retention of once meaningful signs in contexts where they have no place, imperfect corrections, failure to repeat, omissions resulting from similarities in words or syllables, simple omission, omissions of standard lengths, insertions from the margin, miscellaneous bibliographical evidence and other possibilities.

³² See Greg (1927:20-23): "It follows, therefore, that only those variants which give rise to at least two groups of more than one manuscript each can be described as (genetically) *significant variants*. And only those which give rise to groups all of which are of more than one manuscript can be described as *completely significant*. By *significant groups* we shall understand true groups (i.e., of two or more manuscripts) arising from significant variants. ... [A] type-1 variant can never be significant Provided they are numerous enough, type-2 variants afford us all the evidence of which we can in the calculus make use. ... [It] will be apparent that it is only such variation as we see in type 2 that is fundamentally significant."

³³ Presently, I am studying the approach of Dearing (1974) of drawing a chain using only the smallest groups of type-2 variations. This approach is also used in the algorithm in Salemans (1990). When Dearing finds a type-2 variation formula as ABC|DEFG, he uses only ABC for the development of the chain. He does not explain why the larger group DEFG may not be used. Possibly, this is a normal mathematical procedure, but I mistrust it. To put my doubts into cladistic terms (see Part 3 of this paper), maybe the larger group can be an informative, monophyletic, (end) group, too.

mulation: I do not say that every attempt to build chains from non-type-2 variations is incorrect by definition.

The strict use of only type-2 variations in building the chain is a severe limitation, for many observed variations will not have a type-2 character. Therefore, I do not feel at ease with my own type-2 limitation.

For the automatised determination of the text tree of the fourteen *Lanseloet van Denemerken* text versions, the subject of my dissertation, I am confronted with the fact that some of the versions show only parts of the text. One text version contains only about thirty verses, another version only the verses spoken by Lanseloet and Sandrijn, etc. Actually, no variation place is available which occurs in all *Lanseloet van Denemerken* texts. Thus, no type-2 variations covering all fourteen text versions can be determined. My opponents, not hindered by limitations or text-genealogical principles, will possibly smile. My own severe genealogical rules, especially the fourth one, seem to trap me here. During the past years I have studied possibilities of introducing flexibility into the limitation without violating the idea that only type-2 variations offer sound information on the shape of a text tree. Long after my two Amsterdam lectures, I found a way to 'ease' the type-2 limitation, which will be presented in my dissertation in the near future.³⁴

³⁴ Until 1993 I thought that the best way to treat the *Lanseloet van Denemerken* text versions was as follows. First, the computer searches for type-2 variations in thirteen text versions, as if one text version, in particular the one with only about 30 verses, was not extant. Then, the computer searches for type-2 variations in combinations of twelve text versions, as if two texts were not known to us. And so on. This leads to an apparatus of 'partial' type-2 variations, with which the chain of the fourteen text version has to be built. This way of breaking the type-2 limitation is cumbersome, even for a computer: type-2 variations have to be detected in hundreds of combinations of twelve, eleven, ten, etc., text editions. At the end of 1993, I found that it is sometimes possible to rewrite some complex variations into 'artificial' type-2 variations. Suppose we have six texts A to F, and that {A,B,C} have reading '1', where {D,E} show '0' and {F} has no reading, a so-called missing value '?'. Then the variation formula is 'ABC|DE|F'. The formula looks like '111|00|?', when we substitute the text sigla by their values. The formula is not of type-2, because it does not have exactly two true groups. Notice that we do not know whether F once had the value '0', '1' or another value. Biological cladistic literature teaches that it is possible to build Wagner networks ('chains') using type-2 variations with missing values (see also footnote 74, below, about working with missing values while building Wagner trees). Then texts with missing values are treated as if they had not been transmitted to us. The variation formula is treated as if it were 'ABC|DE', which is a type-2 variation. Suppose now that at a certain place of variation texts {A,B} have 'x', {C,D} have 'y' and {E,F} show 'z'. The corresponding variation formula is 'AB|CD|EF'. We can split up this formula into three type-2 formulas 'AB|CD' (as if {EF} have missing values, 'AB|EF' and 'CD|EF'). In this way the type-4 variation formula is broken into three type-2 variations. Each of the three artificial type-2 variation formula has a weight of one third or '0.333'. In

Possibly, biological cladistics, in particular the Fitch parsimony³⁵ for unordered multi-state variations, can help us to leniate further the type-2 limitation.

1.1.5. *The implications of the four genealogical rules shown by an example*

Dees *et al.* (1988-89) offer a 'stamboomreconstructie'³⁶ of the Dutch medieval text versions of *Karel ende Elegast*. Two passages (pp. 97-102), presented here in a rather free English translation,³⁷ will be used to demonstrate the implications of my four text-genealogical rules. I do not dispute here Dees' value as the founding father of Dutch computerised text genealogy.³⁸ However, both passages show that Dees *et al.* do not work with clear text-genealogical fundamentals. The point is that they produce complex variation formulas from observed textual differences between the *Karel ende Elegast* text versions, without sufficient theoretical consideration. Handmade formulas are fed into the computer, which seems to be able to create a text-genealogical tree from them.

When a philologist departs from 'common errors' in his research into the relationships between the different text versions, he is in principle on a dangerous path, because he is dealing with two problems at the same time. He distinguishes sub-

that way the three formulas together have a weight of '1', like the original type-4 formula.

³⁵ The PAUP software package (see Part 3 below) offers facilities to build chains from non-type-2 variations. Wagner parsimony, with which I work, has an alternative called Fitch parsimony, which can treat unordered multi-state characters, which are non-type-2-characters (see the quotation in §3.3 below). I have not studied Fitch parsimony thoroughly enough to oversee its undoubtedly interesting possibilities. In PAUP, Wagner and Fitch parsimony are implemented, as well as Dollo parsimony and Camin-Sokal parsimony. More information on these parsimony methods is to be found in the worthwhile PAUP manual (= Swofford 1991).

³⁶ This process could perhaps more correctly be called 'construction' instead of 'reconstruction'.

³⁷ For the translation of the first passage I used Hogenhout-Mulder (1988:205).

³⁸ Perhaps one could judge the tone of this paper and of Salemans (1990) to be too critical towards esteemed modern text genealogists, especially Dearing and Dees. Dees *et al.*, of course, could not follow my four rules which were first published in Salemans (1990)³⁸12. Possibly it is an adage to say that I show my respect for them by criticising them. I still consider their text-genealogical books and papers to be the most important and influential text-genealogical studies of the last decades; through their works I have learned what text genealogy is. It would be disappointing to have my criticism interpreted as a total rejection of all of their text-genealogical ideas. One is reminded that one of the purposes of this paper is to show that the sharp and partly correct criticism of elements of the method of Lachmann incorrectly led to a general rejection of it.

families on the basis of agreements and differences in readings and he passes judgements on 'earlier' and 'later' by virtue of his perception of what is correct and corrupt. The first point, distinguishing subfamilies, is based on objective facts, which can be verified by everyone. The second, indicating corruptions or errors, is subjective and here philologists tend to differ in opinion.
(Dees *et al.* 1988-89:97)

When we want to be able to distinguish subfamilies, we must have insight in the mutual relationships of the text versions. This insight can be gained by mounting a list of relationship formulas, in which is denoted per verse which text versions share the same reading and which deviate from it. For example, in verse 1 the text versions A and M (after A) show: *FRaeye historie ende al waer*. In B and C the verse starts with met *EEn*, while *al* is missing. D and E offer a main clause: *Dit is een schoon historie ende waer* (after D). Finally, K has a main clause too: *DEse hystorie is al waer*. This results into the following formula: AM|BC|DE|K.
(Dees *et al.* 1988-89:102)

In the first passage, Dees *et al.* assert that subfamilies can be found with objective facts that can be verified by the reader; in the second passage, they show how a variation formula can be created, using the first verses of the text versions as an example.³⁹ The formula AM|BC|DE|K is presented by Dees *et al.* in such way that there can be no doubt about its correctness. We, however, have many reasons to doubt it.

First, it is not clear what criteria are used by Dees *et al.* to find variants and to build variation formulas like AM|BC|DE|K. The quoted passage only mentions vague notions, like 'the same verses', 'starting with the same words while another word is missing' and 'the syntactic (main clause) shape of a verse'. A clear 'Dees method', in which the complete process of building chains and stemmas according to certain principles is described, simply does not exist. The quoted passage shows the lack of sharply defined principles to find the variants and to build the formula mentioned.⁴⁰ I do not claim here that the formula AM|BC|DE|K is necessarily incorrect; however, the creation of this formula is not as simple as Dees *et al.* lead us to believe. Since they do not describe their method at all, their formulas are not expressions of objective facts, as they claim. Because they do not present a systematised approach for building variation formulas, their formulas must be considered as highly subjective judgements and representations of unverifiable subjectivity *in optima forma*. Summarising, Dees *et*

³⁹ The verses are presented in Appendix B of this paper.

⁴⁰ See Appendix B.

al. do not pay attention to the question which elements in text versions can reveal the genealogical kinship, a requisite covered by my first rule.

Second, as expressed above, Dees *et al.* do not present an apparatus of variants from which the variation formulas and the textual tree are created (see also the beginning of §1.1.2); thus it is impossible to verify or check them. This contradicts the principle that science should be open and verifiable, as is expressed by the second rule.

Third, Dees *et al.* do not account for the fact that they treat the complete verse as one place of variation, an issue covered by my third rule. On the one hand, they state (pp. 98ff.) that subfamilies of text versions can be determined by means of similarities and differences at morpheme level, so-called 'readings' or 'variants'. On the other hand, they choose to build variation formulas on the basis of complete verses. The link between singular (morpheme-oriented?) readings as relationship-revealing units and the complete verses as places of variations is unclear. It is possible that several variants in one verse are the result of independent text-genealogical processes. Dees *et al.* do not motivate their choice of one variation formula per verse, whereby several textual developments in the verse are combined.

Fourth, treatment of the whole verse as one place of variation, leads to more complex variation formulas, like AM|BC|DE|K, than when smaller parts of the verse would have been used. We already saw that the treatment of complex variation types as sources of genealogical information should be considered critically, as stipulated in my fourth rule. Text-genealogical software that builds text-genealogies is not convincing, as long as the theoretical basis upon which this software has been developed has not been made clear. Unfortunately, since the software of Dees *et al.* and the method or theoretical thoughts behind it have not been described, their software remains a black box.

1.2. *Can genealogical variants be recognised with the aid of the computer?*

1.2.1. *The second ad hoc rule on text-genealogical significant word types: words in rhyming position, verbs and substantives*

This section is focussed on the question of whether textual variants belong to certain syntactic or word categories. Theoretically and practically it is more attractive to discuss and determine the genealogical power of a few word categories than to discuss the text-genealogical power of thousands of single words.

Many elements may be involved in the development of text genealogies, but the question is: which elements are genealogically informative? Besides

textual variants, as defined by the first text-genealogical rule, bibliographical and codicological information on, for example, the use of pictures, lombards and paragraph signs in text versions, can possibly provide genealogical information as well. For pragmatic reasons I only use textual variants as stemmatic building tools. The main topic of my Ph.D. research, based on the text versions of the Middle Dutch drama play *Lanseloet van Denemerken*, is to study how the computer can be used for text tree development. At the moment the computer can easily detect textual differences between texts. The comparison of, for instance, pictures with the help of the computer (see Salemans 1994-95) is only in its initial phases.

It must be clear by now that as a text genealogist I am only interested in textual variations in text versions that provide us information on their historical transmission. The first text-genealogical basic rule defines such a kinship-revealing variant as a textual difference that fits well in a text version, through which it is plausible that it maintains its form during the text transmission and that it will not be submitted to changes, except for small differences in the spelling.

My second, rather intuitive,⁴¹ *ad hoc* rule or hypothesis is as follows:

The (tentative) second *ad hoc* rule or hypothesis on text-genealogical significant word types:

Nouns and verbs are the most suitable word categories to build text genealogies from; variants in rhyming position in verses, regardless of the word class they belong to, are text-genealogical as well.

Epistemologically speaking, it is not strictly necessary to explain a hypothesis. It simply applies as long as it has not been falsified. The first four text-genealogical principles or hypotheses seem to be more or less general text-genealogical ideas, about which I have but little doubt. The hypothesis on text-genealogical word types and the second rule on text-genealogical significant word types have a different status, expressed in the words '*ad hoc*'.

Let us take a closer look at some elements of the rule. First of all, the final phrase, "variants in rhyming position in verses, regardless of the word class they belong to, are text-genealogical as well", claims that all textual variants in rhyming position in verses are relationship-revealing. The word class they belong to is unimportant; variants in rhyming position can belong to other word classes than nouns and verbs. The thought behind this is that

⁴¹ Popper (1979:355): "...we should dare to put forward bold hypotheses that open, if possible, new domains of observations, rather than those careful generalisations from 'given' observations which have remained [ever since Bacon] the idols of all naive empiricists."

a rhyming word is one of the most essential parts of a verse, if not the most essential part. We may expect that a copyist tries to (re)produce rhyming text copies. Even a creative copyist, who changes numerous elements of his source text, will be confronted with the fact that the verses he produces or copies have to rhyme. The change of a rhyming word has its implications for the rhyming scheme. Furthermore, the final position of the rhyming words in verses attracts attention, which gives these words an even more stable position in the verses. When rhyming verse pairs in two or more text versions are different from the verse pairs at the same variation places in all of the other text versions, this will be, most of the time,⁴² highly informative.

The choice of 'word categories' in the rule above is related to the word-oriented third rule. This choice, however, implies that we will often need syntactic information to discover the word category to which a word variant belongs. Unfortunately, no computerised syntactic and morphological analysers, 'parsers', for the medieval Dutch dialects exist, nor do I have the opportunity to develop them myself. For the development of the *Lanseloet van Denemerken* tree I built a thesaurus with the help of a concordance of the text versions. Roughly speaking, in this thesaurus many word forms, especially the highly frequent ones, are grouped according to word categories like 'adjectives', 'pronouns', 'conjunctions', 'articles', etc. Sometimes the same word occurs in more than one word category. When building type-2 variation formulas from the words occurring in the text versions of *Lanseloet van Denemerken*, the computer is programmed to reject a type-2 variation formula which concerns, for instance, an article or variants that belong to different word categories. In these cases syntactic analysis is not strictly necessary. (I did not build such a thesaurus for the *Yvain* texts.)

The vague element 'most suitable' has been integrated to prevent the impression that the second *ad hoc* rule forbids looking at other word categories as text-genealogical informants. Suppose that one group of text versions

⁴² Notions like 'most of the time' are vague and unwanted in clear definitions or expressions. Nevertheless, I use the expression here to draw attention to the fact that rhyming words or verses can be additions or 'fill-ups' (the Dutch word is 'stoplappen'). Should a copyist be convinced that rhyming words in the exemplar he copies are corrupt, non-original, fill-ups, he or she may change them into new fill-ups. I thank Dr. W. Kuiper for this observation. Clearly recognisable fill-ups do not fit inconspicuously into a text and, therefore, violate the first text-genealogical rule, and do not have text-genealogical power. I realise that it will often be difficult for us to recognise and evaluate fill-ups. For this and other reasons a fifth text-genealogical rule will be formulated in §1.2.3. Compare also the remarks of van der Wal on rhyming words in note 10 above.

read 'this man is always drunk', while the other group has 'this man is never drunk'. The opposition of the adverbs 'always' versus 'never' may offer sound text-genealogical information. Adjectives, which often play an inferior role in a sentence, can be left out or added easily. Because the risk of parallelism is high, the general category adjectives⁴³ has not been mentioned in the *ad hoc* rule.

The choice for 'nouns and verbs' has been derived from the first text-genealogical rule. The rule expresses that a textual element possesses 'survival power' when it is plausible both that the element maintains its form during the text transmission and that it will not be submitted to (further) changes. To me, the word categories nouns and verbs are the fundamental lexical cores or semantic kernels of natural languages or dialects. They may generally be expected in every verse or sentence. Word categories like articles, adverbs, conjunctions and, possibly, adjectives are more marginal and seem to play a more trivial role in the meaning of a statement. They can be left out, added or replaced by resembling elements without much damage to the meaning of that statement. An omission or misplacement of verbs and nouns will more easily attract the attention of a copyist and solicit correction, often successfully.⁴⁴

Possibly, the choice for nouns and verbs can be made more plausible by pointing out that scientists of all ages have claimed that the noun and the verb are the central elements of languages. When we browse through the works of, arbitrarily, Aristotle,⁴⁵ Sapir⁴⁶ and Chomsky, we will easily find

⁴³ In an earlier version of this paper I stated, for instance, that nouns, verbs and adjectives are the main word categories that offer genealogical information. Afterwards I decided to drop the adjectives because this word category in general does not take as stable a position in a sentence as do a noun or a verb. Furthermore, some adjectives (e.g. the pleonastic 'red lips', 'white teeth') are predictable and can be left out and added again easily, which implies a great danger of uninformative synonymous parallelisms. On the other hand some adjectives are so specific that the danger of a copyist introducing them spontaneously is acceptably low. My approach is as follows. First, we try to find type-2 variations built with nouns and verbs. If the variations are well spread, we will have enough chain-building material. If they are not, we must consider other word categories. Dr. P.A. Coppen brought to my attention that currently many linguists claim that the adjective has characteristics of both the noun and the verb.

⁴⁴ Evidently, the process of text copying by copyists and printers has to be studied (more) profoundly.

⁴⁵ In discussing the fundamentals of the rule with my promotor Prof. Dr. G. Dibbets, he pointed out that it has a rather Aristotelian character (see Dibbets 1985:443). To quote Vorlat (1975:42-44): "[Aristotle's] basic distinction is one between 'kategoremata' and 'sundesmoi' ... 'Kategorēmata' are words with a lexical meaning and thus, according to

quotations with such purport. Unfortunately, I do not have the time for a thorough research of this 'noun-verb' case; I will not even offer here a definition of these classes, for the sake of convenience assuming that everyone knows what nouns and verbs are.⁴⁷

It would be interesting to measure the influence of categories like nouns, verbs, adjectives and articles when building a text genealogy. I assume that the use of incorrect genealogical word categories, like articles, offer contradictory and inconclusive information on the shape of the stemma. The 'consistency' or 'trustworthiness' of a tree will be higher when better genealogical word categories are used. With a so-called Consistency Index (CI), the solidity of a text genealogy is expressed. Fortunately, many genealogical software packages, like PAUP (see Part 3), offer CI measurement facilities. We can find out which word categories offer the most trustworthy genealogies. This approach can be repeated for other texts, so that a basis will grow that enables us to find the best word categories from which to build text genealogies, a matter of interest for all text genealogists, also the non-computerised text genealogists. With this material, generated by the computer, the second *ad hoc* rule or hypothesis can be falsified or adjusted. It is even possible to replace this deductive hypothesis by an inductive rule based on these 'empirical' CI-observations, although this does not seem to be necessary to me.

In other words, my strategy to build trustworthy genealogies is as follows. First, I will build a theory on the elements in text versions that provide us dependable information on the kinship of text versions. When I have serious, possibly highly subjective, doubts on the text-genealogical

Aristotle, able to function independently, or as head words in the sentence. 'Sundesmoi', on the other hand, are lexically empty, when standing in isolation. They perform nothing but a grammatical and dependent function with regard to the other category of words, whose functioning they guarantee in the sentence. Aristotle further classifies the lexically full words into 'onomata' and 'remata', i.e., nouns and verbs, as these terms have commonly been translated. This distinction is built on logico-philosophical grounds, viz. the belief that a logically built sentence must predicate an action of an agent. The 'onoma' is the name of the agent of which the action is predicated, whereas the 'rema' denotes the predicated action."

⁴⁶ See Sapir (1921), at the end of chapter V: "No language wholly fails to distinguish noun and verb, though in particular cases the nature of the distinction may be an elusive one. It is different with the other parts of speech. Not one of them is imperatively required for the life of language."

⁴⁷ For *Lanseloet van Denemerken* I have chosen a pragmatic solution. A word is a noun or a verb when the MNW (1885-1952) or the MNHwb (1932, including the 1983 Supplement) labels it as such.

power of textual elements, like, for instance, the adjectives, I simply do not use them for the development of textual trees. Once the text-historical tree, which offers us a view on the historical relationship of the text versions, has been constructed, we can examine how, for instance, the adjectives are in accordance with this tree. Maybe we will find out that the adjectival variants agree completely or partly with the produced historical tree. Then, we may decide to alter our theoretical thoughts and incorporate the category 'adjectives' into the hypothesis on text-genealogical significant word types.

The meaning of the categories 'nouns' and 'verbs' has to be limited to minimise the danger of parallelism. While discussing the synonymous parallelism in §1.1.1, we saw that pronouns are not trustworthy as text-genealogical informants. In the same section it was shown that the inflection of nouns and verbs can present false genealogical information, too. Of the nouns only the substantives seem to have a stable form, which is not easily changeable, except in its declination. Therefore, I would like to reformulate the tentative second *ad hoc* rule:

The (final) second *ad hoc* rule or hypothesis on text-genealogical significant word types:

Substantives and non-auxiliary verbs, regardless of declination or conjugation, are the most suitable word categories to build text genealogies from; furthermore, variants in rhyming position in verses, regardless of the word class they belong to, are text-genealogical as well.

1.2.2. *Eleven characteristics of genealogical variants*

In the previous sections several theoretical genealogical ideas were presented. We will now deduce eleven concrete, recognisable characteristics of genealogical variants from them: a formalisation of the theory.

1. Text-genealogical variants are textual differences, in, preferably, single words, that occur in the *same place of variation*.
(Source: first and third rule.)
2. They are part of a *type-2 variation*.
(Source: fourth rule.)
Mind that this characteristic is not bound to a single variant but to a group of variants.
3. They must be *grammatically adequate* and be positioned in a grammatically correct⁴⁸ context.

⁴⁸ The judgement whether a sequence of words or word groups is grammatically adequate or

(Source: first rule, element 'fits well and inconspicuously'.)

4. a. Text-genealogical variants belong to the *same word classes*.
 b. They are *substantives or verbs*, except auxiliaries of tense (in Dutch: 'hebben', 'zijn', 'zullen'), possibly in argumented combination with other surrounding words that do not belong to these word categories.
 (Source: final second *ad hoc* rule.)
5. Another possibility is that they are *variants in rhyming position* in verses, regardless of the word class they belong to.
 (Source: final second *ad hoc* rule.)
6. The textual differences between text-genealogical variants may *never* be *accidental*, like:
 - a. Orthographic or diacritical differences.
 - b. Differences in word boundaries are considered to be orthographic, and, therefore, accidental.
 - c. Nonsense words.
 - d. Obvious slips of the pen.
 (Source: first rule.)
7. Textual differences are not genealogical (i.e., may not be used for the development of chains and stemmas) when it is likely or imaginable that these differences are parallelistic. Apart from the accidental (orthographic) parallelism, mentioned here as characteristic 6, we know four other, not strictly monolithic, *types of parallelism*: inflectional parallelism ('is' $\longleftrightarrow$ 'was'); synonymous and idiolectic parallelism ('white' $\longleftrightarrow$ 'pale'); regional parallelism ('color' $\longleftrightarrow$ 'colour'); diachronic or historical parallelism. To minimise the danger of parallelism, I formulate the following rules:
 - a. The differences between genealogical variants may not be mere differences in inflection. Therefore, the variants are converted into original 'kernels' or 'roots' (i.e., uninflected *basic lexical forms*, ignoring their gender, number, inflection and suffixes), to be found as 'lemmas' or 'entries' in dictionaries. Inflected word forms can easily take another shape (e.g. 'was' $\longleftrightarrow$ 'has been'; 'have' $\longleftrightarrow$ 'had'). Were such word forms to be used as genealogical variants, the danger of inflectional parallelism is too great.

acceptable, can be different for poetical texts, like drama plays, and prose texts. Sometimes, in poetry the grammatical rules may be slightly violated and yet be acceptable because of rhythm and rhyme. Nevertheless, we may say, generally, that heavy violations of grammatical rules in the poetry of the Middle Ages would have been unacceptable in the eyes of its readers and copyists.

(Source: final second *ad hoc* rule; §1.1.1: the types of parallelism.)

- b. The difference between variants or their roots must not merely be a (phonetic) difference in a *range of vowels*.⁴⁹ Textual differences that consist of mere differences in vowels will always carry the danger of parallelism and can better be passed over as text-genealogical variants. I am interested in heavy textual differences with nearly undisputed text-genealogical power!
- c. All variants must be words that are *generally known* in the languages or dialects of the text versions. This characteristic is mentioned to prevent regional, idiolectic and diachronic parallelisms (see §1.1.1); when variants are caused by differences in languages or dialects, they do not provide dependable genealogical information.
- d. The variants or their roots must be *rare*, in the sense that it is not likely that a variant is easily interchanged with another variant. This characteristic is to prevent the use of synonymous parallelisms (see §1.1.1). Some words can be turned into other more or less synonymous words easily, especially words that express a 'speech act', like 'say', 'speak', 'tell', 'il fait', 'il dit' or words that are names. Admittedly, the word 'rare' in this characteristic is vague.⁵⁰

An element which has not been mentioned earlier is the 'difference in a range of vowels'. When we observe several Dutch medieval dialects we see, roughly speaking, that many differences between regional words are differences in vowels, though changes in consonants also occur. It is likely that the same goes for the French medieval dialects. I am assuming that changes in vowels do not alter the 'word image' as much as changes in consonants do: consonants are more characteristic for a word shape than vowels (cf. Semitic languages where texts are written without vowels; I thank Dr. J.W. Dyk of the Vrije Universiteit in Amsterdam for this observation). A copyist will often understand a word of another dialect when he or she transforms one or more vowels in it. The alteration of consonants in a word is a greater task, simply because there are more consonant characters than vowel characters. Because of the danger of parallelism, differences in vowels between words will normally not provide any text-genealogical information. One could object that this treatment of vowels is not supported by theories, that differences in vowels between word forms are more complex and subtle, and that my vowel approach may sometimes cause incorrect rejections of good text-genealogical variants. I am interested in some clear, nearly indisputable genealogical variants. Having a few variants missing does not imply that it is impossible to develop a text-genealogical tree. Once we have developed dependable genealogical trees we can check if differences in vowels are indeed genealogically uninformative.

Concordance software may help us here. For instance, let the computer produce a frequency list of words used in the total corpus of text versions. Take, for instance, the 150 most frequent words and determine the basic forms of these words: their roots. When in a place of variation both roots belong to the class of highly frequent roots, we should hesitate to use them as genealogical variants.

8. A difference in a syntactically adequate *word order* has to be considered as a genealogical variation as long as the difference in word order does not merely concern a different placement of an adverb in a verse or sentence.
(Source: the final first *ad hoc* rule on word order; because of remaining doubts about the correctness of this characteristic, we must be careful in using it.)
9. When the variants are part of rhyming texts and they stand in a rhyming position, they have to *obey rhyming conventions*.
(Source: first rule, element 'fits well and inconspicuously'.)
If they violate these conventions, this may be due to the interpolation or omission of one or more verses. Such a deletion or addition can be genealogically important.
(Source: end of §1.1.3).
10. The *inversion* of (rhyming words of) verses is genealogical informative when these verses fit well in the text.
11. The *addition* (or *interpolation*) and omission of words (11a) or complete verses (11b) is genealogically informative when these words or verses fit well or offer no crucial information. Additions and omissions of words or verses can be easily detected by the computer. Their genealogical importance has to be determined by a philologist. The computerisation of this judgement is not possible at the moment, because a study of the wide context of the verses concerned is necessary, far beyond the single word boundary of my current text-genealogical concepts. We have to beware that the absence or addition of small, highly frequent words (like 'so') can be parallelistic (see note 16 above).

Evidently, these characteristics of text-genealogical variants work in combination. First, a text-genealogical variant has to be simultaneously in accordance with characteristics 1, 2, 3, 4 or 5, 6, 7 and 9; when one of these characteristics is violated, the textual difference is not a text-genealogical variant. Second, when the variation concerns the order or addition / omission of words or verses, one of the characteristics 7, 8, 9 or 10 can be used to determine whether such variation is text-genealogical.

1.2.3. *The computerised detection of the eleven characteristics: the fifth rule*

The *Lanseloet van Denemerken* text versions show tens of thousands of textual differences. With the eleven characteristics we deduce which textual differences are genealogical. Judging all of these differences manually is too laborious. At the moment many of these characteristics can be detected by the computer automatically (chars. 1, 2, 5, 6a, 6b, 7b, 8, 9, 10) or semi-

automatically (chars. 4 and 11).⁵¹ With the aid of a syntactic and morphological parser, we could have started to automatise the recognition of the other remaining characteristics (chars. 3, 6c, 6d, 7a, 7c, 7d) as well.

Suppose that all eleven characteristics of genealogical variants could be detected completely by the computer.⁵² Then the discovered genealogical variants could be handed over to software like PAUP-3 (see Part 3 below), which will build one or more possible text trees from this material. The question arises whether we should be satisfied with such a complete computerisation of chain or stemma building. The answer is positive and negative. Positive, because I have confidence in the power of my theoretical 'fundamentals' (the rules and the eleven characteristics) in computer shape: it is fascinating to see the computer perform a theory (or a theoretical recipe) accurately and produce one or more text trees. Negative, because a

⁵¹ The computerisation of the search for chars. **1** ('textual differences at the same place of variation'), **2** ('type-2 variation'), **5** ('variants in rhyming position'), **9** ('obey rhyming conventions') and **10** ('inversion of verses') presents no problem. Because the words of all text versions are rewritten in a shorter shape by the computer, most of the accidentals of the original text versions are removed (char. **6**: 'no accidentals', especially char. 6a and 6b). A textual difference concerning one or more vowels (char. **7b**: 'difference not only vowels') will be detected by the computer as well. The detection of char. **8** ('differences in word order') can also be automatised (at the moment this automatised detection is slightly limited, because my software only compares the word order of the basic text with the word orders of the other text versions). The detection of char. **4** (4a: 'same word classes'; 4b: 'verbs or substantives') has been computerised on the basis of a computer-built Dutch word list in which I tagged non-substantive, non-verbal and non-adjective word forms manually. Some homologous word forms have two or more tags and must be analysed when the candidate type-2 formulas are presented. At the moment, the basic lexical form of a word cannot be detected by the computer (char. **7a**). With a parser, a computer program that can analyse the syntactic structure of a sentence, and a dictionary, the computer would be able to find these basic lexical forms of words and could judge the grammatical adequacy of text fragments (char. **3**), get rid of homologies and possibly find out whether a word is generally known (char. **7c**), but rare (char. **7d**). The detection of an interpolation or omission of words (char. **11a**) or verses (char. **11b**) can be automatised; the determination of whether this interpolation or omission is genealogical is left to the philologist (see also the fifth genealogical rule).

⁵² When I started my PhD research in 1986, the objective was to see whether the process of building the text-genealogical tree of the *Lanseloet van Denemerken* text versions could be computerised completely. At that time I saw possibilities of automating this process totally. Within one year I had developed a software package which performed more or less this magical trick, mainly based on the works of Dearing and Dees. After a period of euphoria I felt that my software lacked theoretical background and started to investigate and develop theoretical concepts behind current computerised text-genealogical approaches. A text-genealogical theory or 'recipe' must precede the automation of the development of text trees.

philologist must give the final verdict on the computer-detected genealogical variants.

Several arguments support the last, negative answer. First, variants need to be checked as to whether these variants are indeed genealogical. Possibly, unexpected false variants occur in the variants found by the computer. If so, we can detect how and where the computer program, and, more important, the theory behind it, went wrong. Maybe we can sharpen or even reject the theory and the software, or parts of them. Second, my text-genealogical concepts and software are fundamentally word-oriented, even if parsers for this material were to exist.⁵³ Some variants have to be studied carefully beyond word level before they can reach the status of genealogical variants. A thorough historical-philological study of the surrounding context, of historical events and the development of dialects and language will often be necessary.⁵⁴ Since this historical-philological knowledge is not computerised, we need philologists to check the computer's efforts at developing an apparatus of variants with which text trees are to be built.

At the start of this paper I mentioned that I accepted the challenge to build an *Yvain* stemma with my software, although my knowledge of the medieval French dialects is poor, not to speak of my knowledge of French historical-philological matters. Now I realise that I perhaps trusted too much on my ability to detect the word categories of substantives or nouns and verbs. It takes more than that to build a trustworthy apparatus of variants. Especially as a warning to modern text genealogists, I formulate here the fifth genealogical rule, which is related to my earlier remark that true philologists need to regain fundamental influence in text genealogy:⁵⁵

The fifth text-genealogical basic rule:

Genealogical variants must agree with philological-historical insights.

Furthermore, as mentioned above, presently the recognition of characteristics 3, 6c, 6d, 7a, 7c and 7d cannot be automatised. For these characteristics the efforts of a philologist are necessary.

⁵³ At the moment, without the use of parsers, the software will reject from ninety-five to ninety-nine percent of the textual differences as genealogical variants, because one or more genealogical characteristics are violated. I am pleased with this result. The computer filters out nearly all uninformative variants from the tens of thousands of *Lanseloet* variants. I guess only a few hundred potential type-2 variations will remain which would be worthwhile to check and study carefully.

⁵⁴ See, for example, characteristics 7c and 7d; see also Appendix A, point j below.

⁵⁵ See also Salemans (1987:210) or Appendix D, at the end of §3.2 below.

2. The 'Lanseloet' Genealogical Software applied to the Yvain Texts

2.1. Description of the 'Lanseloet' genealogical software to find type-2 variations

For the drawing of the chain of the fourteen versions of *Lanseloet van Denemerken*, I have developed several computer programs to find genealogical variants:

1. A program to analyse and describe the order of the rhyming words in the text versions. Each text version results in a long string of rhyming words. Among others, verses are detected which do not rhyme (*char. 9*).
2. A program to compare the fourteen strings of rhyming words. Inversions and omissions or additions of rhyming words are detected (*chars. 10 and 11b*).
3. A program to put the fourteen text versions into one synoptic text on the basis of these fourteen strings.
4. A program to produce a shorthand version of the synoptic text. I do not claim that the shorthand offers a correct phonetic representation. The verses are only temporarily rewritten in shorthand to free them of trivial spelling differences which do not possess genealogical power (*char. 6*).⁵⁶ In the final variation formulas with which the genealogy has to be built, the original, non-simplified, variants will again be used.
5. A program to compare the words of the shorthand version of the synoptic verses, verse by verse.⁵⁷ Textual differences that occur in the same place

⁵⁶ Striving to detect significant, 'heavy', variants with the aid of the computer, I do not want to be confronted with thousands of trivial accidental or parallelistic variants. The question of whether it is necessary to use a computer to find these heavy variants receives a simple 'no' as answer. A philologist can often find numerous clear variants in the texts without the computer; however, once the software has been developed, the search for the heavy variants can be performed by the indefatigable computer more quickly, more thoroughly and, above all, more systematically, following the theoretical ideas upon which the text-genealogical software is based. No heavy variant will escape the critical eye of the software. Instead of tedious, fallible and time-consuming browsing in text versions, the computer can find the interesting, noticeable variants within minutes.

⁵⁷ One text version is used as the basic text. Each verse of the basic text is compared with the corresponding verses in the other texts. It is possible that the basic text has one single verse where one or more of the other text versions show more verses, due to deletion or interpolation. This deletion or interpolation of verses will have been detected by my software at an earlier stage (step 2). The 'extra' verses, not occurring in the basic text, will not be compared with the single verse of the basic text by the software. At the moment, for practical 'computer memory' reasons, the word order in text versions are only compared

of variation (*char.* 1) are easily detected. The program is word-oriented: differences in word order (*char.* 8) are noted and, on the basis of a thesaurus developed with the help of the computer, a check is made as to whether the words belong to, for the moment, forbidden word classes like articles, adverbs, conjunctions, pronouns (*char.* 4), and whether the differences are not merely vowels or aspects of inflection (resp. *chars.* 7*b* and 7*a*). Potential type-2 variant groups⁵⁸ (*char.* 2) are detected and, on request, also type-1 groups.

The end result is a list of potential type-2 variation formulas from which a text tree can be built. These formulas have to be judged systematically by philologists, especially on their grammatical (*char.* 3) and historical-philological adequacy.

2.2. The application of the 'Lanseloet' software to the *Yvain* texts

Because I was sent a synoptic text containing the first thousand verses of seven *Yvain* text versions, it was not necessary to run the first three computer programs. Although I realised that I would miss the computerised detection of inversions and omissions or additions ('interpolations') of rhyming words, I decided to start with step 4: the building of the shorthand version of the synoptic *Yvain* texts. Rewriting my software, developed originally for the fourteen Dutch text versions of *Lanseloet van Denemerken*, was difficult. Possibly, my adapted programs missed some important French variants, though this is not necessarily a dramatic error. To develop genealogies, not all existing variants are necessary: some sound, well-spread variations will often suffice.

In step 5, the words of the 'shorthand' *Yvain* texts were compared word for word to find genealogical variants, using the mentioned characteristics. For the French texts I only automatised the detection of chars. 1 ('textual differences at the same place of variation'), 2 ('type-2 variation') and 8 ('differences in word order'). Furthermore, I had to assume that no grammatical errors occur in the texts. For this, I inactivated *char.* 3 ('grammatical

with the word order in the basic text.

⁵⁸ If two potential type-2 variant groups show other sigla, shorthand notations or characters to denote a text version, and if they cover together exactly all sigla of the text versions this combination is a potential type-2 variation (formula). Suppose that for seven texts {A,B,C,D,E,F,G} the following type-2 variant groups are found: {A,E}, {D,E,F}, {B,C,D,F,G}, {A,B,C}. Then the only potential type-2 variation formula, covering all sigla once, is: AE|BCDFG. See also note 59 below.

context'). The occurrence of the other characteristics in the remaining textual differences, which have not been filtered out by the three characteristics, had to be judged without the help of the computer. Fortunately, char. 2 filtered out many variation formulas and variants.

To give an impression of the results of step 5, we use fig. 3, in which the treatment of three verses (vss. 2, 8 and 45) of the *Yvain* text versions is displayed. I would like to pay attention to the following elements:

- a. In the right part of fig. 3 we see the original verses of the *Yvain* text versions as delivered to me on floppy. They have a verse number ('002', '008' or '045') followed by the 'siglum' ('a', 'f', 'g', 'h', 'p', 's' or 'v'; a siglum is a shorthand notation or character or sign that denotes the text version). At the left of these original verses we see the verses in shorthand representation as generated by my software. Notice that for practical reasons, which will not be discussed here, the sigla 'a' to 'v' have been changed into '1' to '7'. These new sigla precede the verse numbers.
- b. In the shorthand verses mainly the vowels composed of two or more characters have been changed into vowels of one character. The words 'rois' and 'roys' in the original verses, for example, have been changed into one shorthand word 'ras': the difference between 'rois' and 'roys' is considered trivial. The chance that copyists used both words arbitrarily implies a high risk of parallelism. One may disagree with simplifying words into a shorthand shape; however, I am especially interested in significant textual differences as solid genealogical building blocks. Additionally, as stated in char. 7b, a mere difference in vowels is of small importance.
- c. The results of step 5 are printed below the dotted lines with 'List of potential Type-2 groups and variations'. The program offers us 'obs.s' and 'combin.s'. Obs.s are computer observations of potential type-2 groups made up of two or more text versions. They are preceded by the verse number. Combin.s are computer-generated combinations of the observed type-2 groups into type-2 variation formulas. The sigla of two type-2 groups with totally different sigla are gathered and investigated by the computer.⁵⁹ If these sigla together contain all seven sigla of the text

⁵⁹ Consider the five observed potential type-2 groups of verse 045 in fig. 2: obs1={1-6-}, obs2={1-2-5-6-7-}, obs3={1-2-3-5-6-}, obs4={2-3-4-} and obs5={4-7-}. The computer tests if each combined observation contains all sigla. No duplicate sigla are allowed. Ten combinations (obs1+obs2, obs1+obs3, obs1+obs4, obs1+obs5, obs2+obs3, obs2+obs4, obs2+obs5, obs3+obs4, obs3+obs5, obs4+obs5) will be tested. Only the combination obs3+obs5, together '1-2-3-5-6-4-7-', contains all sigla without duplicate sigla. The

versions, the computer can build a type-2 variation formula of them. Notice, that in the type-2 variation formulas the sigla 1 to 7 have been replaced by the original sigla a to v.

- d. Obs. 1 and 2 at verse 002 show the detection of a change in word order.⁶⁰
- e. The potential type-2 formula of verse 2 'as|fghpv' applies to the shorthand words 'bans' and 'bons'. The computer recognises that the difference between these words to be merely a vowel. The sixth characteristic determines that this difference does not offer trustworthy genealogical information. The computer mentions the formula, but rejects it immediately: '(light: 1 vowel)'.

The four equivalent potential variation formulas 'hp|afgsv' of verse 8 show that too many observations have been combined. Because we are mainly interested in substantives and verbs, the combinations in which 'li' and/or 'la' occur are unnecessary: because 'li' and 'la' do not belong to these word categories, they cannot occur in a genealogical type-2 variation. If we had had a parser or a computer dictionary of old French words, the presentation of these incorrect formulas would have been prevented: 'li' and 'la' would have been recognised as irrelevant. (For the development of the text tree of the *Lanseloet van Denemerken* text versions I have built such a thesaurus; many incorrect formulas are rejected automatically.) Only the fourth 'hp|afgsv' combination ('ras' vs. 'kor') is correct.

combination of, for instance, observation 1 {1-6-} ('tus') and observation 2 {1-2-5-6-7-} ('mo') is {1-1-2-5-6-6-7-}. This combination contains duplicates for texts 1 and 6. After the computer has detected both duplicates, it 'knows' that the combination of observation 1 and observation 2 cannot lead to a type-2 variation formula. Only the combination of obs3 and obs5 results in a potential type-2 variation formula: '1-2-3-5-6- versus 4-7-'; replacing the number-sigla in the original alphabetical sigla, 'a-f-g-p-s- versus h-v-'; or, placing the smallest group in front of the variation formula, 'h-v- versus a-f-g-p-s-'.

⁶⁰ The procedure to detect changes in word order is as follows. First, the computer labels all words of text 1 as numbers: '1=artus', '2=li', '3=bans', '4=ras', '5=de', '6=bretengn'. Then the verse of text 2 is treated. The words in this verse get the same number codes as the words of text 1. Words which occur in text 2, but do not occur in text one are numbered as 'x': '2=li', 'x=bons', '4=ras', '1=artus' '5=de' 'x=bretengn'. The word sequence of text 1 looks like 1-2-3-4-5-6. The one of text 2 is 2-x-4-1-5-x. The one of text 3 is 1-2-x-4-5-x. Notice that in the left-to-right word sequence of text 1 the numbers are increasing. When this increasing sequence is disturbed in the other texts, it is likely that we have detected a change in word order. It will be clear that for the moment this approach is not optimal when a word occurs more than once in a verse.

- f. Of the observations and combinations, only the word-order observation (verses no. 2), the 008a combination 'hp/afgsv' ('ras' vs. 'kor') and the 045a combination 'hv/afgps' ('greva' vs. 'pesa') are useful.

VERSES NO. 2	
1.002 artus li bans ras de bretegn	< 002a artus li boins rois de bretaingne
2.002 li bons ras artus de bretegn	< 002f li bons rois artus de bretaigne
3.002 artus li bons ras de bretegn	< 002g artus li bons rois de breteigne
4.002 artus li bons ras de bretegn	< 002h artus li boens rois de bretaingne
5.002 li bons ras artus de bretegn	< 002p li bons roys artus de bretaigne
6.002 artus li bans ras de bretegn	< 002s artus li boins rois de bretaigne
7.002 li bons ras artus de bretegn	< 002v li bons rois artus de bretaigne
-----List-of-potential-Type-2-groups-and-variations-----	
1.002 obs. 1: 2-5-7- (= f-p-v-) WORD ORDER	"ras-artus" (Type-2)
1.002 obs. 2: 2-5-7- (= f-p-v-) WORD ORDER	"li-artus" (Type-2)
1.002 obs. 3: 1-6- contain	"bans"
1.002 obs. 4: 1-4- contain	"bretegn"
1.002 obs. 5: 2-3-4-5-7- contain	"bons"
1.002 obs. 6: 2-3-5-7- contain	"bretegn"
002a combin.: a-s- f-g-h-p-v- =	"bans" "bons" (light: 1 vowel)
VERSES NO. 8	
1.008 la kor fu a kardal en gales	< 008a la cors fu a cardoil en gales
2.008 la kor fu a kardol en gales	< 008f la cors fu a cardoel en gales
3.008 la kor fu a karduel en gales	< 008g la corz fu a cardueil en gales
4.008 li ras fu a karduel en gales	< 008h li rois fu a carduel en gales
5.008 li ras fu a kardol en gales	< 008p li rois fu a cardoel en gales
6.008 la kor fu a karduel en gales	< 008s la cours fu a cardueil en gales
7.008 la kor fu a kardol en gales	< 008v la court fu a cardoel en gales
-----List-of-potential-Type-2-groups-and-variations-----	
1.008 obs. 1: 1-2-3-6-7- contain	"la"
1.008 obs. 2: 1-2-3-6-7- contain	"kor"
1.008 obs. 3: 2-5-7- contain	"kardol"
1.008 obs. 4: 3-4-6- contain	"karduel"
1.008 obs. 5: 4-5- contain	"li"
1.008 obs. 6: 4-5- contain	"ras"
008a combin.: h-p- a-f-g-s-v- =	"li" "la" (light: 1 vowel)
008a combin.: h-p- a-f-g-s-v- =	"ras" "la" (heavy: consonants)
008a combin.: h-p- a-f-g-s-v- =	"li" "kor" (heavy: consonants)
008a combin.: h-p- a-f-g-s-v- =	"ras" "kor" (heavy: consonants)
VERSES NO. 45	
1.045 si o de tus ki mo pesa	< 045a si ot de teus qui mout pesa
2.045 si o de teks ki mo pesa	< 045f si ot de tex qui mout pesa
3.045 si o de teks ki en pesa	< 045g si ot de tex qui en pesa
4.045 si o de teks ki mol greva	< 045h si ot de tex cui molt greva
5.045 si o de tiks ki mo pesa	< 045p si out de tix qui mout pesa
6.045 et tus i o ki mo pesa	< 045s et teus i ot ki mout pesa
7.045 si o de tieks ki mo greva	< 045v si ot de tiex qui mout greua
-----List-of-potential-Type-2-groups-and-variations-----	
1.045 obs. 1: 1-6- contain	"tus"
1.045 obs. 2: 1-2-5-6-7- contain	"mo"
1.045 obs. 3: 1-2-3-5-6- contain	"pesa"
1.045 obs. 4: 2-3-4- contain	"teks"
1.045 obs. 5: 4-7- contain	"grev"
045a combin.: h-v- a-f-g-p-s- =	"grev" "pesa" (heavy: consonants)

Figure 3. Demo of output of step 5 on the basis of three Yvain verses

Fig. 3 shows a small part of the computer output on the first thousand verses in the seven *Yvain* texts, giving an impression of the way my software treats these verses with their tens of thousands of textual differences. The automatised process of rewriting the texts into 'shorthand' and the search for potential type-2 variants resulted in about 5000 remaining variants. Finally, the computerised recognition of the mentioned three observable characteristics produced about 150 semifinal variations.

formula	freq.	verse	kind of difference
1. af ghpsv	2	303	"rat" "c(h)at"
2. idem		900	"pris" "vif/uif"
3. as fghpv	2	059	"preus et uailans/vaillans" "molt/mout auenans/ auenanz/avenanz"
4. idem		968	word order "sire yuains .. fu" "fu .. sire ivens"
5. av fghps	1	192	"trouai" "entraï"
6. fg ahpsv	1	629	"apelerioie" "estoit moie"
7. fv ahpsv	1	046	"murmur(r)e" "parole"
8. hp afgsv	1	008	"rois" "cors/corz/cours/court"
9. hv afgps	2	045	"greva/greua" "pesa"
10. idem		209	"entrez" "venus/uenus/uenuz"
11. pv afghs	2	268+269	p and v show no text in these two verses
12. idem		952	"contreual" "(et) chiet/ciet aval"
13. afg hpsv	1	300	"uelues" "mo(u)ssues"
14. afh gpsv	1	588	word order "tant .. m aue2s/avez" "m auez/aue2s tant"
15. ags fhpv	1	342	"pleuis" "justis/iustis"
16. ahp fgsv	1	635	word order "puet .. set" "set .. puet"
17. asv fghp	1	531	"souf(f)rir/soffrir" "ferir"
18. fgv ahsp	1	989	"blecie2s/bleciez" "plai(i)e2s/plaiez"
19. fgs ahpv	1	056+057	inversion of verses (rhyming words)
20. fhs agpv	1	108	"dame" "sire"
21. fpv aghs	1	002	word order "li .. rois/roys artus" "artus li .. rois"
22. gps afhv	1	561	word order "mes armes toutes" "to(u)tes mes armes"
23. hvp afgs	2	465	(type-2) "paroit" "auoit"
24. idem		766	"ato(u)rna" "monta"

Figure 4. Scheme of *Yvain* variation formulas; the apparatus of variants

The other eight characteristics could not be detected by the software. The variations that violated at least one of these characteristics had to be deleted manually. Partly, this was an easy job, because the variations often contained uninformative variants like articles. Sometimes it was more difficult, due to my poor knowledge of medieval French.⁶¹ When there were severe doubts about its validity, a variation was removed, in the hope that enough⁶² variations would remain from which to build an *Yvain* tree. The

⁶¹ The process of eliminating the doubtful variation formulas is interesting, because it may reveal weak elements in my eleven characteristics of text-genealogical variants. Unfortunately, there is not enough room here to show all computer detected potential type-2 variation formulas. Ironically, this seems to be the price one has to pay when one proclaims that preferably the complete apparatus of variants has to be presented (see the second rule)! At the moment I cannot distribute the apparatus permanently by way of the international computer networks like Internet or EARN/Bitnet. All I can do is to present in fig. 4 the final, definitive formulas and claim that these are the only significant variation formulas and variants. The apparatus of variants of fig. 4 can be challenged by demonstrating that the verses of the *Yvain* text versions show other significant genealogical variants which obey the eleven characteristics but are not mentioned in the apparatus.

⁶² I do not know precisely how many well-spread variations are minimally needed to build chains. The absolute minimum amount for seven texts {a,f,g,h,p,s,v} is at least four type-2 variations: three type-2 variations with unique type-2 groups made up of two members (e.g. af|ghpsv, gh|afpsv and sv|afghp), and one type-2 variation with a type-2 group with three members (e.g. psv|afgh). For the development of the *Yvain* tree twenty-four type-2 variations were available.

variations are presented in fig. 4⁶³ above and in Appendix C below, together with the original verses.

2.3. *The impossibility of building the Yvain chain from type-2 formulas*

With the type-2 variation formulas of fig. 4 we can try to build the chain of the *Yvain* texts using the algorithm of Salemans (1990:450). The algorithm cannot lead to a tree of the *Yvain* text versions because too many text versions show confusing, unclear relations with too many other text versions: the type-2 variations are in conflict.⁶⁴ Therefore, when applying the algorithm, we soon will be confronted with step 3d of the algorithm:

3d. (...) the type-2 group cannot take its place in the chain. This is caused by conflation (= contamination), parallelism or incorrect variants. Stop all actions. Reconsider your variants or use e.g. the Wagner network method (see Salemans 1987).

The alternative method indicated by the algorithm, the 'biological' Wagner network method, will be discussed in the following part of this paper (see also Appendix D below, a translation of parts of Salemans 1987).

3. Application of PAUP to the Yvain Text Versions

⁶³ Notice that in variations 2 and 13 in fig. 4 and in Appendix C, adjectives are used. The second *ad hoc* rule does not forbid the use of adjectives. When the danger of parallelism is low, adjectives can be adequate text-genealogical informants (see also footnote 43). I did test whether PAUP builds other *Yvain* trees when the 'adjective' variation formulas 2 and 13 are excluded. The result was that PAUP produced the same two trees.

⁶⁴ Dr. Mertens (see note 1) remarked here: "1. Only one insignificant parallelism has to occur in the text versions, and the Salemans (1990) algorithm concludes: consult your cladist! 2. I would like to propose the following (and that is the way I do it: manually, without the computer): I search for heavy variants which are tremendously relationship-revealing. 3. Then I build a hypothetical chain or stemma. Having this chain or stemma, I look which variants do not fit in this hypothesis. 4. Then I look if it is possible that the violating variants are parallelisms. If they are, I reject them. I perform the weighing after building a chain or stemma." First, it would be inconvenient should one contradicting variant prohibit us from developing a chain. Of course, I will reconsider whether this variant is genealogical or, for instance, a parallelism which slipped through. When it is genealogical, we must face the fact that it is not possible to build a chain with the algorithm. Second, I am interested in tracing only heavy kinship-revealing variants as well. Third, I fear that the adjustment of the variant material afterwards will be too *ad hoc*. The text genealogist could easily fool himself.

3.1. *Systematics and text genealogy*⁶⁵

Thus far we have discussed several aspects of philological ways to develop text-genealogical trees. We now will study how biologists develop genealogical trees of animals and plants to see whether we can profit from their genealogical ordering methods.

Systematics is the evolutionary discipline in biology that deals with the theory and practice of capturing the orderliness in nature. This orderliness has resulted from patterns of phylogenetic ('stemmatic') ancestry and descent of 'taxons', or elements to be classified. In other words, systematics is the biological theory and practice of building genealogical trees: it is biological stemmatology.

Many parallels exist between biological and textual genealogy. Like text genealogists, biological systematians struggle with one fundamental question: what are the elements with which trustworthy genealogies can be built? Systematics can provide new perspectives and insights for text genealogy. The idea that text versions may be treated as biological taxons is not new.⁶⁶ Throughout the ages, biological taxons changed because ancestral characteristics were sometimes not ceded fully to the younger generations; similarly, texts have changed throughout the ages, due to the intended and unintended alterations by copyists and printers.

Systematics and text genealogy share many fundamentals and notions. Unfortunately, many of these basic notions are named differently. In order to learn from systematics, it is necessary to become acquainted with certain terms. The systematic term *homology (in the broader sense)* implies, in text-genealogical terms, 'the phenomenon that text versions share the same textual difference or variant'. The systematic notion *strict* or *restricted*

⁶⁵ This section has mainly been derived from Eldredge and Cracraft (1980:5-13), Ridley (1986), Sober (1988:7), E. Riley (personal notes made during lectures followed in Amsterdam, 1986; see Salemans (1987:212), also to be found in Appendix D below, note 21) and W. Ellis (idem, and personal correspondence).

⁶⁶ See Platnick and Cameron (1977:380), as quoted in note 1 of Salemans (1987:191-192, see Appendix D of this paper). See also Sober (1988:6): "The problem of *genealogical inference* is peculiar to historical sciences. A central example is the task of *phylogenetic inference*: how can we tell, for example, whether human beings are more closely related to chimps than they are to gorillas? However, it is not just species that have genealogies. Single organisms have family trees. Languages evolve and are related to each other by varying degrees of propinquity of descent. Ancient texts, copied by scribes whose copies are then copied, also are related genealogically. Indeed, social, political, economic and artistic traditions exhibit descent with modification ..."

homology can be understood as 'the phenomenon that text versions share a non-original variant, a Lachmannian common error'. A *true* 'Hennigian' *monophyletic group*⁶⁷ of taxons possesses a common ancestor and contains all descendants of that ancestral ('stem') species. It is an end group in a chain (see also note 30 above). The text-genealogical concepts 'parallelism' and 'contamination' are known in systematics as *analogy* and *hybridisation*. When two taxons have an analogous characteristic, they show a similarity of form or structure which is not present in their nearest common ancestor. The analogous characteristic has evolved independently in both taxons.

Within current systematics, three, not strictly monolithic, schools exist.

The first school is called *evolutionary systematics*. It builds genealogies from some (not all!) carefully chosen, trustworthy homologies in the broader sense, without analysing whether these characteristics are original or 'derived', i.e., non-original. Taxons are clustered according to the agreement of the homologies in the broader sense. Evolutionary systematics tries to filter out analogies. Because text genealogists like Dees and Zarri work with unanalysed homologies, their works exhibits parallels with evolutionary systematics. Unlike evolutionary systematics, however, Dees and Zarri do not seem to be opposed to the use of analogies.

The second school, *phenetics* or *numeral taxonomy*, clusters the taxons, often known as 'OTUs',⁶⁸ according to an index of overall similarity of unanalysed homologies in the broader sense and analogies. The position of most pheneticists is that, once enough features or characters have been examined, 'real' or 'true' resemblance will outweigh 'false' resemblance due to parallelism. It is claimed that the law of great numbers will minimise the influence of some false observations. Using statistical techniques, taxons with the most unanalysed characteristics in common are collected into groups. Phenetics is performed in text genealogy by, for example, Griffith, Galloway and, in the Netherlands, Brefeld. The problem phenetics has to face is that taxons do share many original, ancestral character values (or better: character states or feature values; in text-genealogical terms: variants) that do not provide genealogical information. Genealogically distinct taxons run the risk of being grouped or clustered together incorrectly when they share many ancestral character states and they are not strictly type-2 variations.

⁶⁷ See also Salemans (1987:211), §3.3, translated in Appendix D below.

⁶⁸ Roughly speaking, the notion Operational Taxonomic Units (OTU) is used in phenetics (and cladistics, the third school) to indicate 'known animal, plant or species'. See also Salemans (1987:212) or §3.3.1 of Appendix D below.

The third school is the popular and influential *cladistics*, also known as *cladism* or *phylogenetic systematics*. This school emphasises that genealogies of taxons must be built from strictly genealogically informative elements: true monophyletic groups. Classifying taxons into groups only by their common ancestry, cladists try to filter out the analogies by analysing the characters or features.⁶⁹ Generally, the 20th century German scientist Hennig is considered to be the founder of cladistics. In cladistics, *cladograms*, our ‘oriented stemmas’, are often built from *networks* or *phylograms*, our ‘chains’. The Wagner network method is a cladistic method for building phylograms. In cladistics the phenetic overall similarity is not important. Arthur Lee, with whom I corresponded in 1988 about the cladistic analysis of text versions, has explained the differences between cluster analysis and cladistic analysis clearly (Lee 1987:non-paginated p. 2):

Cladistic analysis is sharply differentiated from cluster analysis by that which it measures. Cluster analysis groups the object being analysed or classified by how closely they resemble each other in the sum of their variations, using statistical ‘distance measures’. Cladistic analysis, on the other hand, analyses the objects in terms of the evolutionary descent of their individual variants, choosing the evolutionary tree which requires the smallest number of changes in the states of all the variants.

Roughly speaking, in cladistics first all possible hypothetical trees of descent are developed. Second, the character or feature states are spread across these hypothetical trees. At this stage it is not important whether character states are original or derived. The shortest tree, with the fewest changes of character states, is considered to be the best genealogical tree. This shortest tree idea is based upon the principle of parsimony (see §3.3). Taxons will often share many character states, but yet they can belong to different families, or ‘monophyletic groups’, because they disagree on a few fundamental character states. Cladistics will recognise these different monophyletic groups correctly. More information on cladistics is to be found in Salemans (1987) and in Appendix D below.

3.2. *A cladistic eye-opener for Lachmannians: rooted and unrooted trees*

Cladists build genealogies from true monophyletic groups, just like Lachmannians build their stemmas from common errors (or, better, non-

⁶⁹ I claim that it is possible to recognise potential textual parallelisms. Therefore, we are only confronted with the one-headed monster of contamination.

original or derived genealogical variants). Both biological and philological methods are criticised on the same point: it is hard or impossible to determine whether a variant or characteristic is derived or not. Cladists discovered that these judgements on originality often are not necessary. They claim that with the knowledge of the transformation order of character states — our philological variants — *unrooted trees* ('*phylograms*', '*networks*') can be built. These trees can be oriented into *rooted trees* ('*cladograms*'). Modern text genealogists have for decades been acquainted with unrooted and rooted trees as chains and stemmas; however, generally, modern text genealogists present their chain-to-stemma methods as being totally different alternatives for the method of Lachmann, where chains and stemmas are not clearly distinguished. The innovation is that the concepts of these unrooted and rooted trees can be incorporated into the method of Lachmann as well.

3.3. *Cladistics and the principle of parsimony*

In the next sections we will focus on the software package PAUP. The final 'P' of PAUP stands for 'parsimony', to which some attention must be paid. The principle of parsimony⁷⁰ or economy (Latin *parcere* = 'to save; to be economical') states that evolution takes the fewest steps possible, given the distribution of character states among taxons. To quote Swofford (1991:1):

Parsimony methods search for *minimum-length trees* (...): trees that minimise the amount of evolutionary change needed to explain the available data under a prespecified set of constraints upon permissible *character* changes. The best known discrete-character parsimony method, often called *Wagner parsimony* (...), treats *binary* or *ordered multistate* characters and permits free irreversibility. Multistate character may also be left *unordered* (i.e., any *character state* is permitted to transform directly into any other state), sometimes called *Fitch parsimony* after Fitch (1971).

When two taxons show a common derived characteristic, the parsimonious explanation is that they go back to the same common ancestor showing that characteristic. Shared genealogical character states are more likely to be due to common ancestry than to parallelism or contamination. The method of Lachmann, cladistics and other genealogical methods seek the shortest, most parsimonious tree that accounts best for the disagreements and agreements

⁷⁰ See, among others, Sober (1988); Ridley (1986:61; 1986:189); Wiley (1981:20). See also Dearing (1974:10): "Greg was aware of the principle of parsimony (as we shall see, genealogical reasoning is impossible without this principle)."

in the corpus of variants or characters. Sober (1988) shows that the principle of parsimony is not as simple as it may appear;⁷¹ however, since there is no better alternative, as Sober will affirm, we choose to work with it.

3.4. *PAUP-3*

Version 3 of PAUP, acronym of *Phylogenetic Analysis Using Parsimony*, is software for cladistic analysis, created by Dr. D. Swofford. At the moment PAUP-3 (= Swofford 1991) is only available for the Apple Macintosh computer.⁷² PAUP contains an enormous amount of cladistic ordering features. Among others, the weighing of variants is possible. Here I will focus on building cladistic chains, like Wagner networks, with PAUP.⁷³

3.5. *Running PAUP-3 with the selected genealogical variants*

Before we can run PAUP, the apparatus of variants of fig. 4 has to be rearranged into the matrix of fig. 5. The occurrence of the first variant is numbered as '0' (not intended to indicate 'original') and the second variant

⁷¹ Sober (1988:x-xii): "...this method holds that a set of observations best supports that phylogenetic hypothesis that requires the fewest parallelisms and convergences. The question naturally arises as to what the use of this method assumes about the evolutionary process; does preferring parsimonious hypothesis presuppose that evolution proceeds parsimoniously? ... Are explanations that postulate fewer entities or processes to be preferred over ones that postulate more? Not always. Are common cause explanations always preferable to ones that invoke separate causes? Again, not always. ... Attempts to justify parsimony have not been successful, but neither have attempts to show that it is fatally flawed."

⁷² PAUP-3 versions for computer platforms other than DOS or mainframes are not scheduled as yet (personal e-mail correspondence with Dr. Swofford). Recently, a more definite version of PAUP 3.0 has been produced by Swofford. To date I have not received it and could not use it.

⁷³ More recently, I consider MacClade (1992), a sister of PAUP, to be more convenient for input and output matters and for tracing characters/features or character/feature states. MacClade and PAUP can be used together harmoniously; for instance, MacClade is capable of treating PAUP-NEXUS-programs.

as '1'.⁷⁴ The matrix of fig. 5 is rebuilt into 'NEXUS' file format, as presented in fig. 6.

	A	F	G	H	P	S	V
1. "rat" "c(h)at"	0	0	1	1	1	1	1
2. "pris" "vif/uif"	0	0	1	1	1	1	1
3. "preus et uaillans/vaillans"	0	1	1	1	0	1	
4. word ord. "sire yuains ... fu"	0	1	1	1	0	1	
5. "trouai" "entrai"	0	1	1	1	1	0	
6. "apeleroie" "estoit moie"	1	0	0	1	1	1	
7. "murmur(r)e" "parole"	1	0	1	1	1	0	
8. "rois" "cors/corz/cours/court"	1	1	1	0	0	1	
9. "greva/greua" "pesa"	1	1	1	0	1	0	
10. "entrez" "venus/uenus/uenuz"	1	1	1	0	1	0	
11. p and v show no text	1	1	1	1	0	1	0
12. "contreual" "et) c(h)iet aval	1	1	1	1	0	1	0
13. "values" "mo(u)ssues"	0	0	0	1	1	1	1
14. w.o. "tant ... m aue2s/avez"	0	0	1	0	1	1	1
15. "pleuis" "justis/iustis"	0	1	0	1	1	0	1
16. w.o. "puet ... set"	0	1	1	0	0	1	1
17. "souf(f)rir/soffrir" "ferir"	0	1	1	1	1	0	0
18. "blecie2s/bleciez" "plai(i)e2s"	1	0	0	1	1	1	0
"plai(i)e2s (etc)"							
19. inversion of verses (rh.words)	1	0	0	1	1	0	1
20. "dame" "sire"	1	0	1	0	1	0	1
21. w.o. "li ... rois/roys artus"	1	0	1	1	0	1	0
22. w.o. "mes armes toutes"	1	1	0	1	0	0	1
23. "paroit" "auoit"	1	1	1	0	0	1	0
24. "ato(u)rna" "monta"	1	1	1	0	0	1	0

```
#NEXUS
begin data;
  dimensions ntax=7 nchar=24;
  format symbols = "01";
matrix
A 000001111111000001111111
F 001110011111001110000111
G 111110111111010110011011
H 111111100011101011101100
P 111111101100111011110000
S 1100111111111110101001011
V 111101010000111100110100 ;
end;
```

Figure 6.

Figure 5.

With the NEXUS file format, PAUP can build figs. 7a, 7b and 7c within a few seconds, as is explained in the second chapter of Swofford 1991. We see that the two phylograms (or networks, chains, which have to be oriented or rooted into a stemma later) in figs. 7a and 7b differ considerably. I will not discuss fig. 7c, which offers the consensus tree. I hope that this tree, in

⁷⁴ It is also possible to fill in the value '?' when a variant does not occur in a text version. The possibility of filling in a missing value offers another opportunity to break the type-2 limitation of my method, discussed at the end of §1.1.4. Suppose that we have seven texts {A,F,G,H, P,S,V} showing the following variants in variation place no. 25: in {A,F,G} variant {x}; in {H,P,S} variant {y}; and in {V} variant {z}. This is not a type-2 variation because more than two variants occur. The variation formula looks like AFG|HPS|V. When we want to make use PAUP option of the 'missing value', we can act as if text V does not have a variant. In that case we can add at the bottom of fig. 5 the next line: '25. "x" | "y" ---- 0 0 0 1 1 1 ?'. It is also possible to leave out for instance texts {HPS}. Then the line would look like: '25. "x" | "z" ---- 0 0 0 ? ? ? 1'. In this way non-type-2 variations can be used to develop chains or networks. The option of the missing value '?' enables us to split up non-type-2 variations into type-2 variations (see note 34 above, especially on the 'weights' of split-up variations). The software for building Wagner networks, simply ignores texts versions temporarily. It only works with the values '0' and '1'. I would like to thank Dr. D. Swofford, the creator of PAUP, for his kind explanation of this matter.

which the agreements of the phylograms figs. 7a and 7b are presented, will speak for itself.

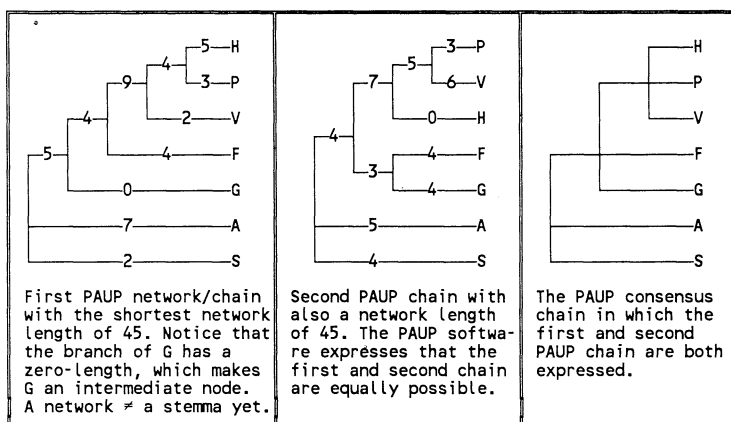


Figure 7a.

Figure 7b.

Figure 7c.

In figs. 7a and 7b the numbers on the branches indicate the lengths or distances between two nodes in a (unrooted) network. In an ideal situation the branch lengths are in accordance with the differences between all character/feature states, '0' or '1' in this case, as pictured in figs. 5 and 6.⁷⁵ This will happen when the character states are not contradictory. The ideal minimum network length, given our twenty-four characters with the states '0' or '1', is '24'.

PAUP has found two minimum-length networks, each with a realised length of '45'. The Consistency Index,⁷⁶ the quotient of the theoretical minimum network length and the realised minimum network length, is '24/45' or '0.533'. The CI of the most trustworthy network is '1', because then the realised tree and the theoretical minimum network have the same

⁷⁵ In fig. 6 we see that text version H has 23 character states with the values '111111100011101011101100'; P has '111111101100111011110000'. This implies that texts H and P have eight different character states or variants. Logically, the minimum distance (*m.d.*) between the two taxons is 8. We see that this *m.d.* has been achieved in both figs. 7a and 7b. When we compare the character states of texts G and P we see that they differ in twelve character states. In other words, the *m.d.* between G and P is 12. Neither in fig. 7a (distance G to P: 0+4+9+4+3=20) nor in fig. 7b (distance G to P: 4+3+7+5+3=22) has this *m.d.* been realised. The *m.d.* between A and G is 12, which has been realised in fig. 7a but not in 7b. The *m.d.* between F and H is 14, which has been achieved in fig. 7b but not in fig. 7a.

⁷⁶ See Salemans (1987:221) or §3.3.4 of Appendix D below.

length. I will not discuss here when a CI is to be considered consistent:⁷⁷ possibly it is better, and at least more convenient, to work with the realised network lengths. The current CI of '0.533' or the realised network length of '45' is, however, not optimal. Clearly, some character states must be contradictory, probably due to contamination.

3.6. Comparing the PAUP trees with the *Yvain* stemma of Micha

It is interesting to compare the PAUP trees for the first thousand verses with the generally accepted *Yvain* stemma, developed by the Chrétien de Troyes authority Micha,⁷⁸ as depicted in fig. 8a.

In fig. 8b the PAUP network or chain of fig. 7a is presented again. In fig. 8c this chain is oriented (in biological systematical terms, 'rooted') into a stemma. Because of my language handicap, I cannot motivate why the point of orientation in the fig. 8a chain lies on or nearby the internode between {H,P}, {V} and {F,G,S}. Another point of orientation would most likely lead to another stemma. All that matters, for the moment, is that it is evident that the stemma of fig. 8c can be derived from the chain of fig. 8b, which is completely in accordance with Micha's stemma, a remarkable and encouraging result.

⁷⁷ Sober (1988:166-172) discusses Felsenstein's attempt to establish a sufficient condition for parsimony to be statistically consequent. Robinson and O'Hara (1992) used PAUP for the development of the text-genealogical tree of 46 manuscripts of the Old Norse narrative sequence *Svipdagsmal*. I share their enthusiasm about PAUP. Unfortunately, they do not discuss why the CI of 0.29 of their *Svipdagsmal* tree is acceptably high. (Our *Yvain* CI of 0.533 cannot be compared with this lower CI of 0.29, because the amount of Norse and French textual material differs.) Furthermore, they do not offer us a glimpse of their, undoubtedly huge, apparatus of variants as provided by the interesting Collate software package. See Maddison and Maddison (1992:114,269-271,278-284,367-368), for more information on the CI (Consistency Index), the EC (Ensemble Consistency), the RI (Retention Index) and the Rescaled Consistency Index (RC).

⁷⁸ See Micha (1966:154). After presenting the stemma, Micha adds some comments, partly concerning the first thousand verses of the *Yvain* texts studied here: "Tel est le schéma général. Mais ici vont surgir les exceptions multiples qui le modifient de façon continue. La plus frappante est celle qui *au début* (1-1000 environ) et *à la fin* du roman (4500-6818) aboutit à la formation HV < P(F)GAS. Les passages pullulent où les manuscrits se répartissent de cette façon" When M. van Mulken offered me, and the other participants of the Amsterdam 1991 symposium (see note 1 above), the verses on a floppy, she did not tell me that their source was Chrétien's *Yvain*. Therefore, I had to produce the text tree without foreknowledge, although it would not have been difficult to trace the source of the verses. In other words, I compared my trees with Micha's tree only after my computer had produced them.

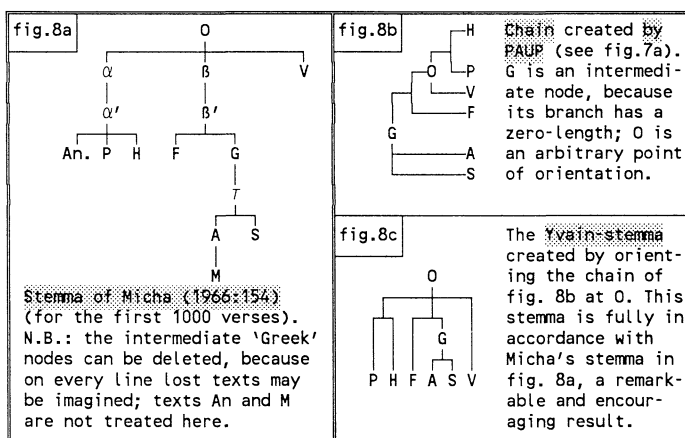


Figure 8a.

Figures 8b and 8c.

3.7. Short evaluation of using PAUP and cladistics for text-genealogical matters

Though PAUP and cladistics are important for text genealogy, the success of PAUP has to be put in perspective. First, we must realise that it was not only PAUP that generated the chain of fig. 7a: initially, my text-genealogical rules and the eleven characteristics produced a very selected set of genealogical *Yvain* variants. Without this filtered set of variants, PAUP probably would not have succeeded in building a chain or stemma as pictured above.

Second, PAUP did not produce one, but two chains, as pictured in figs. 7a / 8b and 7b. Both chains have the same CI of '0.533', thus implying that they are not optimal. Logically, the same goes for the stemma of fig. 8c and possibly also for the stemma of Micha, as Micha would admit (see note 78 above). Furthermore, it is difficult to choose between the two chains, and the rooting of the chain is problematic.

Third, cladists say that a low CI is due to contamination and parallelism. Unfortunately, PAUP does not treat contamination well. It does not visualise contamination in the phylograms and cladograms ('rooted trees') by, for example, convergent lines. Hopefully, cladists, and especially PAUP's creator Swofford, will find ways to handle and express contamination in cladistic trees. This will be an unwelcome task, because PAUP is based on the principle of parsimony, whereby common genealogical characteristics

are explained by common ancestry and preferably not by contamination and parallelism.

4. Final Remarks and Perspectives

Text genealogists can benefit from the insights of their colleagues in the biological fields. Presently in biological systematics, the cladistic school is often preferred. In particular, the text-genealogical method of Lachmann can make use of biological cladistics, because both methods go back to the same basic concepts. The more or less general 'scientific' rejection of this old method should be reconsidered: by means of sophisticated cladistic insights, Lachmann's method has growing potentials and can become again one of the leading text-genealogical methods.

Much has to be done. In the first place, biological ordering methods have to be studied thoroughly. It would be interesting to see if 'evolutionary ordering' scientists like biologists, philologists, historical linguists (relationship of languages), theologians (the original text of the Bible), archaeologists, etc., could profit from each other's ordering methods.

Finally, I want to point out two interesting perspectives for text-genealogical research. First, we should study further possibilities of breaking further the severe type-2 limitation of my 'method', as has been discussed at the end of §1.1.4 and in notes 34 and 74 above. Second, it would be interesting to investigate, with the help of the computer, which (universal?) word categories offer the best genealogical ordering perspectives, as has been mentioned at the end of §1.2.1.

References

- ANS 1984 = *Algemene Nederlandse Spraakkunst*, ed. by G. Geerts, W. Haeseryn, J. de Rooy en M.C. van den Toorn. Groningen/Leuven.
- Brefeld, J. 1994. *A Guidebook for the Jerusalem Pilgrimage in the Late Middle Ages, A Case for Computer-Aided Textual Criticism* (= *Middeleeuwse Studies en Bronnen* XL). Hilversum.
- Buuren, M.B. van. 1988. *Filosofie van de algemene literatuurwetenschap*. Leiden.
- Dain, A. 1949. *Les Manuscrits*. Paris. (Troisième édition revue et augmentée, 1975. Paris).
- Dearing, V. 1974. *Principles and Practice of Textual Analysis*. Berkeley / Los Angeles / London.

- Dees, A. 1975. "Sur une constellation de quatre manuscrits". *Mélanges de linguistique et de littérature offerts à Lein Geschiere par ses amis, collègues et élèves*, ed. by A. Dees, A. Kibédy Varga et R.E.V. Stuip, 1-9. Amsterdam.
- Dees, A. 1976. "Considérations théoriques sur la tradition manuscrite du lai de l'Ombre". *Neophilologus* 60.481-504.
- Dees, A. 1977. "Over stambomen en handschriften". *Forum der Letteren* 18.63-78.
- Dees *et al.* (A. Dees, M. Dekker and M. Hogenhout-Mulder). 1988-89. "Een voorbeeld van stamboomreconstructie: Karel ende Elegast". *Spektator* 18.96-118.
- Dekker, M. 1987. "Intermediariteit". *Corpusgebaseerde Woordanalyse Jaarboek 1986*, 61-69. Amsterdam: Vakgroep Taalkunde, Vrije Universiteit.
- Dibbets, G. 1985. *Twe-spraak van de Nederduitse letterkunst (1584)*. Ingeleid, geïnterpreteerd, van kommentaar voorzien en uitgegeven door G.R.W. Dibbets. (= *Studia theodisca*, 17). Assen.
- Duinhoven, A. (ed.). 1969. *Karel ende Elegast, Diplomatische uitgave van Middelnederlandse teksten en de tekst uit de Karlmeinet-compilatie* (= *Zwolse Drukken en Herdrukken voor de Maatschappij der Nederlandse Letterkunde te Leiden*, 62, twee delen). Zwolle.
- Duinhoven, A. 1988-89. "Stamboomreconstructie: rekenkunde of tekststudie?". *Spektator* 18.119-123.
- Duplacy, J. 1979. "Préalables philologiques à la classification automatique des états d'un texte". *Ordinateurs et critique des textes* 1979.23-33.
- Eldredge, N. and J. Cracraft. 1980. *Phylogenetic Patterns and the Evolutionary Process, Method and Theory in Comparative Biology*. New York.
- Ellis, Willem N. n.d. [1985?]. *TreeTools, A set of programs for the construction of Wagner Networks and Trees for the Macintosh 512 computer, Manual*, draft version. Amsterdam: Institute of Taxonomic Zoology, Dept. Entomology, University of Amsterdam.
- Epp, E.J. 1976. "Towards the clarification of the term textual variant". *Studies in New Testament Language and Text, Essays in Honour of George D. Kilpatrick on the Occasion of his sixty-fifth Birthday*, ed. by J.K. Elliott, 153-173. Leiden.
- Farris, J.S. 1970. "Methods for computing Wagner Trees", *Systematic Zoology* 19.83-92.
- Fitch, W.M. 1971. "Toward defining the course of evolution: minimal change for a specific tree topology". *Systematic Zoology* 20.406-416.
- Froger, J. 1968. *La critique des textes et son automation*. Paris.
- Froger, J. 1979. "La méthode de Dom Quentin, la méthode de distance et le problème de la contamination". *Ordinateurs et critique des textes* 1979.13-22.
- Gaffney, E.S. 1979. "An Introduction to the Logic of Phylogeny Reconstruction". *Phylogenetic Analysis and Paleontology*, ed. by J. Cracraft and N. Eldredge, 79-111. Columbia.
- Galloway, P. 1979. "Manuscript Filiation and Cluster Analysis: the Lai de l'Ombre Case". *Ordinateurs et critique des textes* 1979.87-95.

- Greg, W.W. 1927. *The calculus of variants, An Essay on Textual Criticism*. Oxford.
- Griffith, J.G. 1979. "Non-Stemmatic Classification of Manuscripts by Computer Methods". *Ordinateurs et critique des textes* 1979.73-86.
- Haan, M.J.M. de. 1977. "De filologie en haar hulpwetenschappen". *Geschiedenis van de Nederlandse taalkunde*, ed. by D.M. Bakker en G.R.W. Dibbets, 63-78. Den Bosch.
- Havet, L. 1911. *Manuel de critique verbale appliquée aux textes latins*. Paris.
- Hennig, W. 1982. *Phylogenetische Systematik*, Herausg. Wolfgang Hennig (= *Pareys Studentexte*, 34). Berlin-Hamburg.
- Hennig, W. 1984. *Aufgaben und Probleme stammesgeschichtlicher Forschung*, Herausg. Dr. Wolfgang Hennig (= *Pareys Studentexte*, 35). Berlin-Hamburg.
- Hockey, S. 1983. *A Guide to Computer Applications in the Humanities*. Baltimore / London.
- Hogenhout-Mulder, M. 1988. "The Filiation of Manuscripts of Der naturen bloeme". Van Reenen and van Reenen-Stein (eds.) 1988.205-220.
- Kochendörfer, G. and B. Schirock. 1976. *Maschinelle Textrekonstruktion*. (= *Göppinger Arbeiten zur Germanistik*, 185). Göppingen.
- Kristeller, P.O. 1984. "The Lachmann Method: Merits and Limitations", *Journal of the Society for Textual Scholarship*.11-20.
- Kuiper, W. 1989. *Die riddere metten witten scilde: oorsprong, overlevering en auteurschap van de Middelnederlandse Ferguut*. Amsterdam.
- Lee, Arthur R. 1987. *Numerical Taxonomy Revisited: John Griffith, Cladistic Analysis and St. Augustine's Quaestiones in Heptateuchum*, Paper presented to The Tenth International Conference on Patristic Studies. Oxford, England. Also published in *Studia Patristica* XX. 1989.
- Maas, Paul. 1957. *Textkritik*, 3., verbesserte und vermehrte Auflage. Leipzig.
- MacClade 1992. See Maddison and Maddison 1992.
- Maddison, Wayne P. and David R. Maddison. 1992. *MacClade, Analysis of Phylogeny and Character Evolution*, Version 3. Sunderland, Massachusetts.
- Micha, Alexandre. 1966. "Chapitre V: classement des manuscrits d'Yvain". *La tradition manuscrite des romans de Chrétien de Troyes* (= *Publications romanes et françaises*, 90). Genève (2e tirage; 1. tirage: Paris 1939).
- MNW 1885-1952 = E. Verwijs and J. Verdam. *Middelnederlandsch woordenboek*, Voltood door F.A. Stoett. 's-Gravenhage (11 vols.).
- MNHwb 1932 = J. Verdam, *Middelnederlandsch handwoordenboek*, Onveranderde herdruk en van het woord 'Sterne' af opnieuw bewerkt door C.H. Ebbinghe Wubben. 's-Gravenhage. (Including the Supplement by J.J. van der Voort van der Kleij. Leiden/Antwerpen 1983.)
- Mulken, M. van. 1993. *The Manuscript Tradition of the Perceval of Chrétien de Troyes, A stemmatological and dialectological approach*, dissertation. Vrije Universiteit, Amsterdam.

- Ordinateurs et critique des textes 1979 = *La pratique des ordinateurs dans la critique des textes 1979, Colloques Internationaux du Centre National de la Recherche Scientifique*, No. 579, Paris 29-31 mars 1978. Paris: Éditions du Centre National de la Recherche Scientifique.
- PAUP 1991: see Swofford 1991.
- Platnick, N. and H. Cameron. 1977. "Cladistic Methods in Textual, Linguistic and Phylogenetic Analysis". *Systematic Zoology* 26.380-386.
- Popper, K.R. 1979. "Appendix I: The Bucket and the Searchlight: Two Theories of Knowledge". *Objective Knowledge*, 341-361. Oxford (revision of the first edition of 1972).
- Quentin, Dom. H. 1926. *Essais de critique textuelle (ecdotique) par Dom H. Quentin, moine Bénédictin de l'abbaye de Solesmes*. Paris.
- Reenen, P.Th. van and K. van Reenen-Stein (eds.). 1988. *Distributions spatiales et temporelles, constellations des manuscrits, Etudes de variation linguistique offertes à Anthonij Dees à l'occasion de son 60me anniversaire*. Amsterdam / Philadelphia.
- Ridley, M. 1986. *Evolution and classification, The reformation of cladism*. New York.
- Robinson, P. and R. O'Hara. 1992. "Cladistic analysis of an Old Norse manuscript tradition", *ALLC-ACH 1992 - Conference abstracts*.
- Salemans, B. 1987. "Van Lachmann tot Hennig: cladistische tekstkritiek", *Gramma* 11.191-224.
- Salemans, B. 1988. "Boekbespreking van Buuren (1988)", *Gramma* 12.180-182.
- Salemans, B. 1989. "Varianten als bouwstenen van stemma's: een pleidooi voor eenvoud en openheid bij het opstellen van tekststambomen". *Wat duikers vent is dit!, Opstellen voor W. Hummelen*, ed. by G. Dibbets en P. Wackers, 319-349. Wijhe.
- Salemans, B. 1990. "Text genealogical remarks on Lachmann, Bédier, Greg and Dearing". *Leuvense Bijdragen* 79.427-468.
- Salemans, B. and R. de Bonth. 1990-91. "Opmerkingen bij Duinhovens De Geschiedenis van Beatrijs". *Spektator* 20.197-228.
- Salemans, B. and F. Schaars. 1987-.... *Alfabetische concordanties van Vroeg-nieuwnederlandse geschriften*. Wijhe/Assen. (Until 1994 fourteen parts in this series have been published.)
- Salemans, B. and F. Schaars. 1990. *Concordantie met alfabetische woordenlijst, frequentielijst, retrograde woordenlijst, tekstweergaven en concordantie van hoogfrequente voorden van het dramatisch werk van Joost van den Vondel*, 5 parts. Assen.
- Salemans, B. and F. Schaars. 1993. *Ruusbroecs woorden? Voorlopige concordantie, alfabetische woordenlijst, frequentielijst, retrograde woordenlijst en tekstweergave van Jan van Ruusbroecs werken*, 3 parts. Nijmegen / Tilburg.
- Salemans, B. 1994-95. "Comparing Text Editions with the Aid of the Computer". *Computers and the Humanities* 28.133-139.

- Sapir, E. 1921. "Chapter V. Form in Language: Grammatical Concepts". *Language, An introduction to the language of speech*. New York: A Harvest / HBJ Book.
- Sober, E. 1988. *Reconstructing the Past: Parsimony, Evolution and Inference*. Cambridge / London.
- Swofford, D.L. 1991. *PAUP: Phylogenetic Analysis Using Parsimony, Version '3.0s*. Draft, Illinois.
- Vorlat, E. 1975. *The development of English grammatical theory 1586-1737 with special reference to the theory of parts of speech*. Leuven.
- Wal, M. van der. 1992. *Geschiedenis van het Nederlands*. Utrecht: Aula-boek.
- Willis, J. 1972. *Latin Textual Criticism*. Urbana / Chicago / London.
- Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York.
- Zarri, G.P. 1977. "Some experiments on automated textual criticism". *Association for Literary and Linguistic Computing Bulletin* 5.266-290.

APPENDIX A: A Translation of Part of Salemans (1989) on Parallelisms

Salemans (1989:336-341) describes textual differences, 'parallelisms' or 'accidentals', which can be changed or recognised by copyists. These text-genealogical uninformative variants are, in rather free English translation (*LvD* = *Lanseloet van Denemerken*): '(...)

- a. Differences in use of capitals and small letters ('Karel de Grote' vs. 'karel de grote').
- b. Differences in spelling ('roesen' vs. 'roisen').
- c. Differences in dialect and language ('brood' vs. 'bread').
- d. Differences in use of punctuation marks ('oh! oh!' vs. 'oh, oh').
- e. Differences in boundaries of words ('metten' vs. 'met den').
- f. Differences in clause headers (or incorrect placement of or clear absence of clause headers. BS 1993), when the same clause of a drama play is spoken by different persons (in some *LvD* texts clauses occur with incorrect clause headers; a copyist familiar with the text could detect these false headers and simply correct them).
- g. 'Ungrammaticalities' (ungrammatical sentences can often be easily corrected).
- h. 'Nonsense Readings' (compare Duplacy 1979:28; Epp 1976:168).
- i. Clear copy mistakes ('Karel de Grote' vs. 'Krl de Grote'; compare Epp 1976:168).
- j. Names (...). In the *LvD* text versions is spoken about 'sint Jan' and 'sint iohan'. This is a clear example of a ungenealogical variant (...): the names of 'Sint Jan' and 'Sint Johannes' are still used for indicating the same saint. Different names in text versions, which refer to the same person, only are genealogically relevant (...) when the names concern unknown persons, who do normally not play an important role in the story or 'the world'. (...)
- k. Archaic words. Many copyists will use a more contemporary word when confronted with an archaic word. Therefore, there is considerable chance that copyists working with minimally related exemplars introduce the same more modern word in their copies. The occurrence of the same more contemporary word in the text versions is not due to equal descent, but by diachronic change of language: they are not genealogical variants but parallelisms.'

I would like to add to these a quotation from Salemans (1989:329):

- l. '(...) highly frequent words, which are usually not kinship-revealing (or text-genealogical) and, therefore, must be treated with the highest caution.'

See also Dearing (1974:21-58).

APPENDIX B: Criticism of Dees *et al.* 1988-89 (an Addendum to §1.1.5)

The first verses of the *Karel ende Elegast* text versions are to be found in Duinhoven (1969:part 1, 16-17; part 2:4-5), as is referred to in footnote 8 of Dees *et al.* (1988-89:118):

- A: Fraeye historie ende al waer
- B: Een vray historie ende waer
- C: Een vraye hystorie ende waer
- D: Dit is een schoon historie ende waer
- E: HET is een schoon historie ende waer
- K: Dese hystorie is al waer
- M: Vraye ystorie ende al waer

I repeat here one of the Dees *et al.* (1988-89:102) passages, already mentioned in §1.1.5:

When we want to be able to distinguish subfamilies, we must have insight in the mutual relations of the text versions. This insight can be gained by mounting a list of relationship formulas, in which is denoted per verse which text versions share the same reading and which deviate from it. For example, in verse 1 the text versions A and M (after A) show: **FRaeye historie ende al waer**. In B and C the verse starts with met **EEn**, while **al** is missing. D and E offer a main clause: **Dit is een schoon historie ende waer** (after D). Finally, K has a main clause too: **DEse hystorie is al waer**. This results into the following formula: AM|BC|DE|K.)

It is easy to pose many critical questions and remarks:

- a. Why are the presence of 'Een' in B and C and the fact that 'al' is missing assumed to be informative? Why has no attention been paid to the fact that A, B, C and M have a form of the adjective 'Fray', where D and E have the alternative adjective 'schoon' and E and K lack these adjectives?
- b. Why do Dees *et al.* not mention that 'al' is present in A, K and M?
- c. Why does the fact that D and E present a main clause tell us something about their text-genealogical relation? Why do D, E and K not form one group? After all, they all are main clauses.
- d. Why is the fact that text K does not have the conjunction 'ende' ignored? The absence or occurrence of 'ende' is not just an accidental variant. In Dees *et al.* (1988-89: 101), we read (translated into English): 'Therefore, we must start with composing a list in which all agreements and differences between the text versions ('redacties' in Dutch; BS) are registered. Of course the list to be built must apply to all available manuscripts (...). Furthermore, the list has to be exhaustive (...).' Dees (1975:4) adds that these agreements and differences concern the level of morphemes: 'Dans une première approximation, il suffira de noter les identités et les différences au niveau des vers. Un vers donné sera considéré comme identique au vers correspondant d'un autre manuscrit, si aucun morphème n'a été remplacé par un morphème différent (abstraction est faite de variations graphiques et de différences de déclinaison casuelle).' Considering the questions above, the claim of Dees *et al.* that they treat all variants, putting aside the accidental variants, cannot be maintained.

- e. Furthermore, statements like 'M (after A)' and 'E ... (after D)' give rise to the impression that Dees *et al.* claim that the first lines of text versions M and E are (quite) the same as, respectively, A and D. However, A and M do not show the same opening lines (admittedly, the difference between 'Fraeye' and 'Vraye' is small or accidental), nor do E and D (the difference between 'Dit' and 'Het' cannot be specified as accidental).

More remarks can be found in Duinhoven (1988-89), which offers a review of Dees *et al.* (1988-89).

APPENDIX C: Variation Formulas and Original Verses (see fig. 4)

Formula 1: afghpsv [rat | c(h)at]
303a iex de cuete et nes de rat
303f et nes de kieure et els de rat
303g nes de cuete et nes de chat
303h ialz de cuete et nes de chat
303p le couille noire nes de chat
303s iols de cuette et nes de cat
303v oilz de suette nes de chat

Formula 2: afghpsv [pris | vif/uif]
900a se mort ov pris ne detient
900f se mort ov pris ne le detient
900g se mort ov uif ne le detient
900h se mort ov vif ne le retient
900p se mort ov uif ne le detient
900s se vif u mort ne le detient
900v se uif ov mort ne le detient

Formula 3: asfghpv [preus et uailans/ vaillans | molt/mout auenans/aunenanz/ avenanz] [N.B. I was told that "zz" = "uns"; the two places of variation, showing preus/mout and vaillans/aunenanz] are presented here as one place of variation.]
059a vns cheualiers preus et uailans
059f zz cheualiers mout auenans
059g zz cheualiers mout auenanz
059h uns chevaliers molt avenanz
059p unz chevaliers mout auenans
059s uns cheualiers preus et uailans
059v uns cheualiers mout auenanz

Formula 4: asfghpv [word order "sire yuains ... fu" | "fu ... sire ivens"]
968a si que me sire yuains fu pris
968f ensi fu mes sire yuains pris
968g einsint fu me sire yuains pris
968h ensi fu mes sire yuains pris
968p ainsi fu me sire yuains prins
968s ensi me sire yuains fu prins
968v ensi fu mis sire yvains pris

Formula 5: avfghps [trouai | entrai]
192a trouai et ui zz bretesche
192f entrai et ui vne bretesche
192g entrai et uers une bretesche
192h entrai et vi une bretesche
192p entrai et ui une bretesche
192s entrai et vi une bretesche
192v la trouai ge vne bretesche

Formula 6: fgahpsv [apel. | est. moie]
629a de traison se ele estoit moie
629f de traison m apelerioie
629g de traison l apelerioie
629h de traison s ele estoit moie
629p de trayson s ele estoit moie
629s de traison s elle estoit moie
629v de traison s el estoit moie

Formula 7: fvaghps [murm. | parole]
046a et qui mout grant parole en tinrent
046f et qui mout grant murmure en firent
046g et qui mout grant parole en firent
046h et qui mout grant parole en firent
046p et qui mout grant parole en firent
046s et ki mout grant parole en fisent
046v et qui mout grant murmure en firent

Formula 8: hpafgsv [rois | co(u)r.]
008a la cors fu a cardoil en gales
008f la cors fu a cardoil en gales
008g la cors fu a cardueil en gales
008h li rois fu a carduel en gales
008p li rois fu a cardueil en gales
008s la cours fu a cardueil en gales
008v la court fu a cardueil en gales

Formula 9: hvafgps [gre. | pesa]
045a si ot de teus qui mout pesa
045f si ot de tex qui mout pesa
045g si ot de tex qui en pesa
045h si ot de tex cui mout greva
045p si out de tix qui mout pesa
045s et teus i ot ki mout pesa
045v si ot de tiex qui mout greua

Formula 10: hvafgps [entrez | ven.]
209a par v laiens uenus estoie
209f par onc laiens uenus estoie
209g par ou leenz uenuz estoie
209h par ou leanz entrez estoie
209p par ou laiens uenus estoie
209s par v laiens venus estoie
209v par ou laienz entrez estoie

Formula 11: pvafghs [p and v show no text in (rhyming) verses]
268a petit por mon hoste fessie
268f petit por mon hoste fessie
268g petit por mon oste fessie
268h petit por mon oste fessie
268p
268s petit pour men oste fessie
268v

269a se cest don li escondeisse
269f se cest don li escondeisse
269g se cest don li escondeisse
269h se cest don li escondeisse
269p
269s se ce don li escondeisse
269v

Formula 12: pvafghs [contreal | (et) chiet/ciet aval]
952a descent la porte et ciet aual
952f descent la porte et ciet aual
952g descent la porte et chiet aual
952h descent la porte et chiet aval
952p destent le porte contreal
952s descent la porte et chiet aval
952v descent la porte contreal

Formula 13: afgghpsv [uelues | mo(u)s.]
300a oreilles uelues et grans
300f orilles uelues et grans
300g oreilles uelues et granz
300h oreilles moussues et granz
300p oreilles moussues et grans
300s oreilles moussues et grans,
300v oreilles moussues et granz

Formula 14: afhigpsv [word order "tant ... m - auez2/avez" | "m auez2/ave2s tant"]
588a que uos tant le m auez2s cele2
588f que vous tant le m auez2s cele2
588g que uos le m auez tant cele2
588h quant vos tant le m auez cele2
588p que uous le m auez2s tant chele2
588s que vous le m auez2s tant cele2
588v quant uos le m auez tant cele2

Formula 15: agsfhvp [pleuis | iust/ just.]
342a je les gare si je uos pleuis
342f jeo gart si cestes et justis
342g si les gart ie le uos pleuis
342h je gart si cestes et justis
342p je les gart si uoir et iustis
342s iou les gart se iou te pleuis
342v ge gart ci cestes et iustis

Formula 16: ahpfgsv [word order "puet ... set" | "set ... puet"]
635a car tant puet et set et tant uaut
635f tant set et tant puet et tant vaille
635g tant set et tant puet et tant vaut
635h tant puet et tant set et tant vaut
635p tant puet et tant set et tant vaut
635s tant set et tant puet et tant vaut
635v tant set et tant puet et tant vaut

Formula 17: asvfghp [souf./sof. | ferir]
531a si grant colp que ie poc souffrir
531f si grant fais com io poi ferir
531g si grant cop com ie poi ferir
531h si grant cop con je poi ferir
531p si grant cop com ie poi ferir
531s si grant cop iou poi souffrir
531v si grant cop com ge poi souffrir

Formula 18: fgyahsp [blec. | plai.]
989a car me sire est a mort plaie2s
989f car mes sire est a mort blecie2s
989g que me sire est a mort blecie2
989h que mes sire est a mort plaiez
989p que me sire est a mort plaie2s
989s car me sire est a mort plaie2s
989v car mis sire est a mort blecie2

Formula 19: fgsahpv [change of order of verses (rhyming words)]
056a et kes et me sire gauains
056f et si i fu mes sire yuains
056g et si i fu me sire yueins
056h et kex et mes sire gauvains
056p li rois et me sire gauains
056s et si fu me sire yuains
056v et keus et mes sires gauuains

057a aueuc eus fu me sire yuains
057f et kex et mes sire gauains
057g et keu et me sire gauuains
057h et si i fu mes sire yvains
057p et si fu pres me sire yvains
057s et kex et me sire gauains
057v et ouoec aus mis sire yuains

Formula 20: fhsahpv [dame | sire]
108a "sire" fait il "de la tencon
108f "dame" fait il "de la tencon
108g "sire" fait il "de la tencon
108h "dame" fait il "de la tencon
108p "sire" fait il "de la tencon
108s "dame" dist il "de la tencon
108v "certes sire de la tencon

Formula 21: fpvaghps [word order "li ... rois/roys artus" vs. "artus li ... rois"]
002a artus li boins rois de breitaingne
002f li bons rois artus de breitaingne
002g artus li bons rois de breitaingne
002h artus li boens rois de breitaingne
002p li bons roys artus de breitaingne
002s artus li boins rois de breitaingne
002v li bons rois artus de breitaingne

Formula 22: gpsahfv [word order "mes armes toutes" | "to(u)tes mes armes"]
561a et totes mes armes ius mis
561f et toutes mes armes tramis
561g et mes armes toutes ius mis
561h mes jus totes mes armes mis
561p mais mes armes toutes ius mis
561s et mes armes toutes ius mis
561v et ius totes mes armes mis

Formula 23: hpvafgs [paroit | auoit]
465a qu il n i auoit branche ne foille
465f qu il n i auoit brance ne foille
465g qu il n i auoit branche ne foille
465h qu il n i paroit branche ne fuelle
465p qu il n i paroit branche ne fuelle
465s k il n i auoit brance ne fuelle
465v que n i paroit branche ne foille

Formula 24: hpvafgs [ato(u)r. | monta]
766a ot atorne2 et puis monta
766f ot atorne2 et si monta
766g ot atorne2 et si monta
766h en aporte cil l atorna
766p ot atorne2 si l atoma
766s ot atorne2 puis si monta
766v li a porte2 si l atorna

APPENDIX D: A Translation of Parts of Salemans (1987) on Cladistics*

1. *Introduction*

[...] In this paper the basis of a new text-critical method is formulated, which I would like to call *cladistic textual criticism*. [...] I will demonstrate how biologists investigate the genealogies of living species (plants and animals) using the cladistic method. This method, for which the onset was given in 1950 by the biologist Willy Hennig and which is still in development, appears to be a general method, which can also be applied to other scientific fields, like, for instance, text-genealogical research.¹ [...]

3. *Biological genealogical research: classification, taxonomy, phylogenetics and cladistics*

Taxonomy is the art of classification: the arrangement of different elements into categories, groups, etc. It is not an exclusive part of one particular branch of science, but a general discipline. Taxonomic elements or 'taxa' (the plural of 'taxon') are ordered into a hierarchy, based upon the presence or absence of certain 'characters' (or 'characteristics', 'attributes') occurring in these taxa. Throughout the last years, taxonomy has developed enormously, especially within the field of *biological cladistics*. I am convinced that the cladistic taxonomic methods¹⁶ can be used outside of biological scientific territories, for example, in text genealogy.

* [I would like to thank Josephie Brefeld for bringing to my attention a slight, but disturbing error in the original Gramma paper (Salemans 1987; see step 5 in §3.3.2). Only parts of the complete paper have been translated here: a part of the Introduction (p. 191), and §3 to §3.3.4 (pp. 208-221). Though it was tempting to incorporate in the translation my current, more sophisticated knowledge on text genealogy, I translated the paper as it was. However, I did add comments in few places, recognisable by being placed between square brackets '[' and ending with my initials followed by the current year: BS 1994. Because passages have been left out (marked as '[...])' the numbers of the 'afbeeldingen' (= 'figures') and notes do not follow each other consecutively. From the original notes only notes 1, 16, 17 and 19 to 23 are displayed. I would again like to thank here Dr. W. Ellis to whom the Gramma paper was dedicated. BS 1994.]

¹ The idea that the biological cladistic method can also be applied to the research of the relation of texts or languages, has been voiced by the biologists Platnick and Cameron (1977:380): "The concept that historical interrelationships can be demonstrated only by the presence of shared innovations is fundamental to the fields of textual, linguistic, as well as phylogenetic, reconstruction. ... The ... cladistic analysis is a general comparative method applicable to all studies of historical interrelationships based on ancestor-descendant sequences."

¹⁶ For more of the same mind, see note 1.

3.1. *Building blocks of relationship structures: characters and character states*

In taxonomic orderings, taxa are ordered on the absence or presence of certain characters. The fundamental problem is determining which characters can serve to this purpose. [...] Determining whether a character is adequate or not depends upon the purpose to which it serves. When it is used to determine the genealogical relation of taxa, the character has to possess a *relationship-revealing quality or power*. While choosing ordering characters, the researcher has to ask himself continuously whether his chosen characters possess this quality. Usually, the best characters are those which cannot appear or disappear in a species spontaneously.

A *character* is an abstraction (for example, 'color of eyes'; 'length'; 'presence of wings'). A character state is the way a character manifests itself in a taxon (for example, 'green', 'yellow'; '1 metre', '1 decimetre'; 'having wings'). The *state* is an specification of the character.

Character states can be divided into underived (original) and derived (non-original) character states.¹⁷ The quality of characters is more important than their quantity: with a few characters many taxa can be ordered. For instance, with six characters, each with two character states, theoretically $2 \times 2 \times 2 \times 2 \times 2 \times 2$ or 64 taxa can be sorted out. The relation between the number of characters and the taxa to be ordered is exponential. One single, well-chosen character can often reveal more than thousands of doubtful characters can! [...]

3.2. *An intermediate balance: the significance for cladistic textual criticism*

In my opinion, text versions may be considered as taxa for which the relation pattern has to be determined. The characters to be used for that goal have to possess relationship-revealing qualities. From this point of view, the nature of a *reading*, used in generating textual family trees, can be described as *a character state in which a relationship-revealing character manifests itself in a taxon or text version*. This implies that not every textual difference can be used as a character in building text-genealogical trees. Not every textual variant is a relationship-revealing character state! More often it will rather be an exception than a rule that a textual difference can be used as a character state to build a stemma with. Whether we like it or not, we will always have to determine whether the differences between text versions are relationship-revealing, because only with such differences can the relation pattern ('chain' or 'stemma') of these text versions be determined. [...] From a taxonomic point of view, the fundamental idea of many modern text-critical methods that differences and agreements in readings can be determined easily and objectively, is incorrect. A sound choice of relationship-revealing readings is not as simple and obvious as it may appear at

¹⁷ In the literature on cladistics many concepts are mentioned, which are not of immediate importance for the understanding of cladistic textual criticism. I will explain some of the most frequently employed cladistic concepts. When a derived, non-original character state is occurs uniquely in one single taxon, this character state is called *apomorphic*. An apomorphic character state is known as an *apomorphy*. When two or more taxa share a derived character state, this is entitled as a *synapomorphic* character state (or *synapomorphy*). An original, primitive (= *plesiomorphic*) character state is called a *plesiomorphy*. When several taxa show the same plesiomorphic character state, we have to do with a *sympleiomorphy*.

first sight. The determination of whether an element of a text version is a relationship-revealing character state has to be sustained with arguments. [...]

3.3. *The Wagner network algorithm: the structure of a cladogram*

'Cladograms' or 'relationship structures' or, in text-genealogical terms, 'stemmas', can be generated in several ways. When composing cladograms it is often important to know which character states are original and which are derived. The most important, direct building block of a cladogram is the so-called *monophyletic group*. A group of taxa is monophyletic if all of these taxa go back to one single common ancestor, while it is impossible to point out any single taxon that does not go back to that common ancestor and does not belong to that group of taxa. In other words, the monophyletic group contains the ancestral species and all of the species which descended from that common ancestor. I want to emphasise here explicitly that the 'method of Lachmann' in fact works with monophyletic groups for the development of text genealogies. The Lachmannian monophyletic groups are found with the 'common errors', derived character states. The often so heavily attacked 'method of Lachmann' may be considered as a predecessor of Hennig's method and deserves rehabilitation.

For most philologists the method of building text genealogies from monophyletic groups will hold little fascination, because they claim that usually it will be unclear which character states ('readings' or 'variants') are original and which ones are not. Fortunately, a method exists that enables us to determine the structure ('chain') of cladograms without the interpretation or judgement of character states: *the Wagner network method*.

A Wagner network is the cladistic counterpart of the philological chain. From a Wagner network a cladogram can be created by 'orienting'¹⁹ (or 'rooting', 'polarising') it, exactly like a philological chain can be oriented into a chain. The Wagner network is made up of interlinked nodes, so-called OTUs and HTUs. OTU is the abbreviation for 'Operational Taxonomic Unit', a taxon which must be given its place in a genealogy or relation structure. HTUs are 'Hypothetical Taxonomic Units', the (untransmitted) hypothetical, inside nodes of a Wagner network, comparable with Quentin's intermediate nodes.

The building of a Wagner network can be done with the twelve-steps algorithm,²⁰ presented in the next section. This algorithm must be worked through from top to bottom, step by step. Only when the algorithm explicitly states that one is to skip to another step in the algorithm (see, for example, step 12), is departure from this top-to-bottom procedure allowed. [...]

¹⁹ [...] When orienting a philological chain into a stemma it is 'pulled up' or branched upwards; when orienting a cladistic network, it is derived downwards ('rooted'). In the cladistics the generating of a network (chain) is called *anagenesis*; the polarisation or orienting of the network is called *cladogenesis*.

²⁰ An algorithm is a systematic description of the steps to be taken in a certain order to solve a problem. The Wagner network algorithm, presented here in a simplified version, has been derived from Farris (1970:89-90 [corrected by me in 1993; BS 1994]), Wiley (1981:176-192) and primarily from notes which I wrote down during the lectures 'Numerical Cladistics' (course number 207 of the educational series 'Taxonomy') in 1986, given by Dr. W. Ellis of the University of Amsterdam.

The Wagner network algorithm:

1. Build the table of character states of the OTUs.
2. Build the table of distances of the OTUs.
3. Put the value of the 'HTU counter' at 0.
4. Build the first internode of the Wagner network. Procedure: find two OTUs, X and Y, which mutually have the greatest distance from all (couples of) OTUs, as can be determined from the table of distances; connect or join together both OTUs, X and Y, through a line; the result will be the first internode (XY) of the network.
5. Determine the 'work internode' and the 'work OTU'. Procedure: find an OTU with the smallest distance to the first internode; this must not have been placed in the network as yet; the internode XY is called the 'work internode' and the newly found OTU 'work OTU'.
6. Connect the 'work internode' and the 'work OTU' through a HTU and give this HTU a unique name, for example, with a raised 'HTU counter'. Procedure: draw from the middle of the 'work internode' a line to the 'work OTU'; the point at which the line out of the 'work OTU' connects with the 'work internode' is a HTU; raise the value of the 'HTU counter' by 1 and label the HTU on the 'work internode' with the name of the value of that 'HTU counter' at the moment: for example, HTU(1), HTU(2), etc.
7. Determine the median character states of the HTU(value of 'HTU counter').
8. First, expand the table of character states; then, second, expand the table of distances with the data of the HTU(value of HTU counter).
9. Build a survey of the distances of the OTUs which have not yet been connected to all existing internodes.
10. Find in this survey an unconnected OTU of which the sum of distances to all internodes is largest. Call this OTU the (current) 'work OTU'.
11. Find the internode with the smallest distance to the 'work OTU'. Call this internode the (current) 'work internode'.
12. If any unconnected OTU exists go to step 6. Otherwise stop.

3.3.1. Building Wagner networks with distances between character states

In this paper I will not discuss in depth the principles behind Wagner networks. Primarily, attention will be paid to the way these networks can be built. Nevertheless, I would like to discuss here briefly the central thought behind building Wagner networks: the notion 'distance'.

With the Wagner network method, taxa (from now on I will call them 'OTUs'), are grouped into a chain on the basis of the knowledge of the *distances* between the OTUs.²¹ In a Wagner network, OTUs with many character states in common stand closer to each other than the OTUs with fewer character states in common. When character states are expressed in words, it is difficult to speak about their mutual distance. After all, what is the distance between character states 'black', 'white' and 'grey' of the abstract character 'color'? When using the Wagner network method the character states are expressed by numbers. In this way it is possible to talk about the distance between character states. The

²¹ The idea of determining the distances between nodes in a chain has been formulated, for other purposes, in J. Froger's "La méthode de Dom Quentin, la méthode de distance et le problème de la contamination", in *Ordinateurs et critique des textes* 1979, 13-22.

total distance between two OTUs can be determined by adding the distances between the several characters, or, more precisely, character states.

The distance between two character states is equal to the difference between the two numbers which represent these states. Suppose that we specify the color character states 'black', 'white', and 'grey' with, respectively, the number codes '1', '5' and '3'. The distance between, for example, 5 (white) and 3 (grey) equals 2, for $5-3=2$. The distance between 5 and 1 (black) is larger: 4. Notice, that it would not have made any difference if we would have chosen for these colors the numbers '5', '9' and '7'. Only *the distances between the character states* are important for the development of Wagner networks. Therefore, the choice to represent 'white', 'black' and 'grey' by '1', '5' and '3', or by '5', '9' and '7' is arbitrary, because the distances are equal. All characters to be used for the building of the Wagner networks must have the same importance or weight; we then can decide that the maximum distance between the character states equals, let us say, '6'.²²

The subsequent question concerns how character states can be labelled with the use of numbers. Suppose that we want to assign to all characters the same weight: '6'. When a character has precisely *two states*, it will be clear that the distance between the OTUs with one state and the other OTUs with the other state must be '6'. As long as the difference between the states equals '6', it is not important which numbers represent the character states. When we label one character state with '7', the other state can be represented by '13' or '1'.

If *more than two character states* exist, the choice of the representation of differences will be more complicated. The different states have to be labelled in such way that the differences in distance between more related states is smaller than the distance between less related states. Suppose that the character 'color' occurs in the OTUs with the character states 'white', 'grey' and 'black'. If we take grey to be a transitional color (state) exactly between white and black, we must choose the numbers in such way that the distance between grey and white (or black) equals '3'; the distance between black and white has to be '6'. For instance, for the states white, grey and black we can choose the numbers '1', '4' and '7', or '9', '6', and '3', respectively. It is only important that the transitional color state grey is represented by a number between the other numbers representing the states white and black.

If character states are considered to be transitional, an interpretation will be given of the order of the development of the character states. This order is expressed in a so-called *transformation order or sequence*, like 1--2--4--7: character state '1' developed into '2', '2' into '4' and '4' into '7'. The right-to-left interpretation of this sequence is also allowed. Only the distances between the character states of the OTUs matter. The direction of interpretation of the transformation is not important, but it will often be difficult to build the transformation order, since it is not easy to determine objectively which character states are transitional and which ones are beginning or end states. Such interpretation is not necessary when a character has exactly two states. Therefore, usually it will be convenient and recommendable to build Wagner networks from these *binary characters, with two character states*. [...]

²² It is possible to assign heavier weights to certain characters. The distance (or the difference) between the character states of these heavier characters is larger.

3.3.2. Example of working with the Wagner network algorithm

Step 1: Build the table of character states of the OTUs.

Suppose that we want to determine the relationship of six taxa A, B, C, D, E and Pre, and that we have eight characters s, t, u, v, w, x, y and z. After having determined the character states in these OTUs (or taxa), we build a *table of character states*, in which the six OTUs A, B, C, D, E and Pre are put vertically and the eight characters s to z horizontally. Under and beside the table some space is left for later extensions. In the table we note the character states, represented by numbers. A matrix of character states will be the result, for example:

		s	t	u	v	w	x	y	z	--characters-->
O T U s	A	1	0	0	0	0	0	0	2	'z' is the character 'color';
	B	0	0	0	0	0	1	0	0	'0' represents, e.g., 'black'
	C	1	1	1	1	1	0	0	2	(in B and Pre), '1' repre-
	D	0	0	0	0	0	1	0	2	'grey' (in E) and '2' repre-
	E	0	0	0	0	0	0	1	1	sents 'white' (in A, C, D)
	Pre	0	0	0	0	0	0	0	0	

Fig. 9: first table of character states

We see, for instance, that character x occurs in OTUs B and D as state '1' and in the other OTUs as '0'. 'z' is a character with more than two character states ('0', '1' and '2'). Notice that the distances between the character states are not equal. The distance between the states of character s is '1' and of z '2'. As stated before, usually it is better to choose the same distances between character states to avoid having some character states with a heavier weight. However, we presume that our table is correct.

Step 2: Build the table of distances of the OTUs.

With the table of character states we can build a table of distances between the OTUs. To determine the distance between two OTUs, we add up all the distances between all the character states of the OTUs. When we compare, for example, the character states of OTUs C and D, we see that the distance for character s between these OTUs is '1', because the difference between '1' and '0' equals '1'. For t the distance is 1, for u 1, for v 1, for w 1, for x 1, for y 0, for z 0: 1+1+1+ 1+1+1+0+0=6. The total distance between C and D is 6, which is inserted into the table of distances. After having calculated all distances between the OTUs, the first table of distance will be as follows:

	A	B	C	D	E	Pre
A	*	4	4	2	3	3
B	4	*	8	2	3	1
C	4	8	*	6	7	7
D	2	2	6	*	3	3
E	3	3	7	3	*	2
Pre	3	1	7	3	2	*

Fig. 10: first table of distances

[...]

Step 3: Put the value of the 'HTU counter' at 0.

Thus far, no part of a Wagner network exists, because there are no HTUs. The 'HTU counter', which is used to number or label the HTUs as HTU(1), HTU(2), etc., is zero. Each time we perform step 6, the HTU counter will be raised by 1.

Step 4: Build the first internode of the Wagner network.

The building of the Wagner network can now commence. First, we look in the table of distances to find the two OTUs which differ most and, therefore, have the highest mutual distance. We connect these two OTUs, B and C (distance 8) by a line. In doing so, the first internode, a line between two nodes, of the Wagner network has been established:

B-----C

Fig. 11: the first internode of the Wagner network

Step 5: Determine the 'work internode' and the 'work OTU'; make a connection between the first internode and the OTU with the shortest distance to it.

A third OTU can now be connected to the internode BC. [...] This third OTU must have the smallest distance to the internode BC. To determine the distance from a node to two other nodes, we use the depicted formula.

Formula to determine the distance between node Z and the internode XY: ===== $d(Z/XY) = 0.5 \times \{d(XZ) + d(YZ) - d(XY)\}$ in which: d(Z/XY) = distance from Z to internode XY d(XZ) = distance from X to Z d(YZ) = distance from Y to Z d(XY) = distance from X to Y
--

In other words, the distance from a node Z to an internode XY can be determined as follows: add the distance from X to Z and the distance of Y to Z; subtract from the result the distance from X to Y; divide that amount by 2. With this formula we can determine which new OTU has to be attached to the internode. We will use the distance table. The distance from A to the internode BC equals {the distance from A to B (4), plus the distance of A to C (4), minus the distance from B to C (8)} divided by 2: $0.5 \times (4 + 4 - 8)$. This equals 0. The distance from D to BC is $0.5 \times (2 + 6 - 8)$; also 0. The distance from E to BC is $0.5 \times (3 + 7 - 8) = 1$. The distance from Pre to BC is $0.5 \times (1 + 7 - 8) = 0$ [and not $0.5 \times (3 + 7 - 8) = 1$, as stated in Salemans 1987, which Josephie Brefeld (see footnote * above) correctly noticed; BS 1994].

We conclude that both A and D [and Pre; BS 1994] have the smallest distance to the internode BC. Therefore, these OTUs can be chosen as the nodes to be connected to the internode. Arbitrarily, we choose A. The more characters we can use for building a network, the bigger the distances between the OTUs will be and the smaller the chances will be that we will have to make arbitrary choices between OTUs to be attached to an internode. [...]

Step 6: Connect the 'work internode' and the 'work OTU' through a HTU and give this HTU a unique name, for example with a raised 'HTU counter'.

The procedure for attaching an OTU to an internode is as follows: first, point out a point at the middle of the internode; from this point, draw a line towards the OTU. The new node on the internode is a Hypothetical Taxonomic Unit (HTU). The HTU counter is raised with 1 and the new HTU is labelled as HTU (HTU counter) (for example, HTU(1), HTU(2), etc.). By raising the HTU counter each new HTU will get a unique name. At the moment our Wagner network looks like:

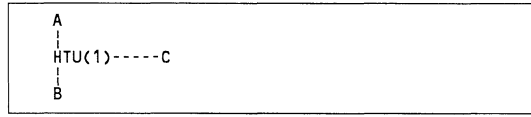


Fig. 12: Wagner network with first HTU

Step 7: Determine the (median) character states of the HTU (value of 'HTU counter').

Since HTU(1) is a hypothetical node its character states cannot be observed, but have to be calculated separately. For this calculation we use the next rule: in a HTU, a character will show the character state which is the median value of the three OTUs connected through the HTU. The *median value* of three values is one of those three values, namely, the value which lies the closest to the mean value of the three values. So the median value of 0, 3 and 5 is '3': the mean value is 2.666 and so '3' is the closest value. In the table of character states we can look up the character states of A, B and C. With these states we can calculate the median character states of HTU(1).

	s	t	u	v	w	x	y	z
A	1	0	0	0	0	0	0	2
B	0	0	0	0	0	0	1	0
C	1	1	1	1	1	0	0	2
HTU(1)	1	0	0	0	0	0	0	2

Fig. 13: Median determination of character states of HTU(1)

Step 8: Expand (first) the table of character states and then (second) the table of distances with the data of the HTU (value of HTU counter).

We add the character states of HTU(1) to the tables of character states and the table of distances:

	s	t	u	v	w	x	y	z
A	1	0	0	0	0	0	0	2
B	0	0	0	0	0	1	0	0
C	1	1	1	1	1	0	0	2
D	0	0	0	0	0	1	0	2
E	0	0	0	0	0	0	1	1
Pre	0	0	0	0	0	0	0	0
HTU(1)	1	0	0	0	0	0	0	2

	A	B	C	D	E	Pre	HTU(1)
A	*	4	4	2	3	3	0
B	4	*	8	2	3	1	4
C	4	8	*	6	7	7	4
D	2	2	6	*	3	3	2
E	3	3	7	3	*	2	3
Pre	3	1	7	3	2	*	
HTU(1)	0	4	4	2	3	3	*

Figure 14.

N.B.: Notice that HTU(1) and A are equal; A is in fact the intermediate HTU between C and D.

Step 9: Build a survey of the distances of the OTUs which have not been connected at the moment to all existing internodes.

[...] The Wagner network now has three internodes: A-HTU(1), B-HTU(1) and C-HTU(1). One unconnected OTU must now be attached to this network. First, we build an OTU-internode distance table with the distance formula (see step 5). For example, the distance from E to the internode C-HTU(1) is $0.5 \times \{\text{distance C to E, plus distance HTU(1) to E, minus distance C to HTU(1)}\}$. Using the extended table of character states, we substitute this into: $0.5 \times \{7+3-4\} = 3$. In this way we calculate all distances from the unconnected OTUs to all internodes. Result:

	A-HTU(1)	B-HTU(1)	C-HTU(1)	total distance to internodes
D	2	0	2	=> 4
E	3	1	3	=> 7
Pre	3	0	3	=> 6

Fig. 15: survey of OTU-internodia distances

Step 10: Find in this survey a unconnected OTU, of which the sum of distances to all internodes is largest. Call this OTU the (current) 'work OTU'.

Looking at the survey (fig. 15), we see that node E (distance '7') must be the new 'work OTU'.

Step 11: Find the internode with the smallest distance to the 'work OTU'. Call this internode the (current) 'work internode'.

Looking at the survey (fig. 15), we see that 'work OTU' E is closest to internode B-HTU(1), which becomes our new 'work internode'.

Step 12: If any unconnected OTUs exists go to step 6. Else stop.

Because the OTUs D, E (the current 'work OTU') and Pre have not been connected to the network, we jump back to step 6. There 'work OTU' E will be connected to the 'work internode' B-HTU(1) through the new HTU(2). The result is pictured in the next network:

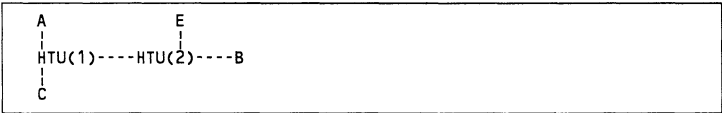


Fig. 16: Wagner network until now with 2 HTUs

End of algorithm; stop:

After the algorithm has been completed and all OTUs have been attached to the Wagner network, the tables of character states and distances will look as pictured in fig. 17. The helping lines within it mark out the original OTUs and the new artificial intermediate nodes, the HTUs.

	s	t	u	v	w	x	y	z		A	B	C	D	E	Pre	i1	i2	i3	i4
A	1	0	0	0	0	0	0	2	A	*	4	4	2	3	3	0	2	3	2
B	0	0	0	0	0	1	0	0	B	4	*	8	2	3	1	4	2	1	2
C	1	1	1	1	1	0	0	2	C	4	8	*	6	7	7	4	6	7	6
D	0	0	0	0	0	1	0	2	D	2	2	6	*	3	3	2	2	3	2
E	0	0	0	0	0	0	1	1	E	3	3	7	3	*	2	3	1	2	1
Pre	0	0	0	0	0	0	0	0	Pre	3	1	7	3	2	*	3	1	0	1
i1	1	0	0	0	0	0	0	2	i1	0	4	4	2	3	3	*	2	3	2
i2	0	0	0	0	0	0	0	1	i2	2	2	6	2	1	1	2	*	1	0
i3	0	0	0	0	0	0	0	0	i3	3	1	7	3	2	0	3	1	*	1
i4	0	0	0	0	0	0	0	1	i4	2	2	6	2	1	1	2	0	1	*
Definitive table of character states									Definitive table of distances										

Figure 17.

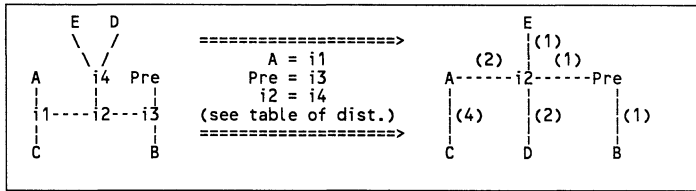


Fig. 18: the definitive Wagner network (right)

In the tables 'HTU()' is replaced by the shorter character 'i' (from 'intermediate') to save space. The Wagner network is pictured in the left part of fig. 18. In the definitive table of distances in fig. 17, we see that the nodes which have a mutual distance of 0 are identical. Therefore, we can cut some HTU-nodes from the network, which results in the definitive Wagner network, pictured in the right part of fig. 18.

Next to the lines of connection in the right network, the distances between the nodes of the network appear as mentioned in the table of distances (fig. 17) between brackets '()'.

3.3.3. Rooting ('orienting') the Wagner network ('chain') into a cladogram ('stemma')

Rooting or orienting a Wagner network is the same as orienting a philological chain. [...] We can pull up the chain at the point where the character states are original. Remember that the character states used until now have not been judged on originality. [...] In fig. 19 the Wagner network has been oriented at three different points [1. at Pre; 2. between i2 and D; 3. at E; BS 1994]. This results in three different cladograms (see fig. 19). [...] Notice that the point of rooting or orientation in the second cladogram lies between i2 and D. Lost text versions may be given a place on or imagined at the lines of connection in a chain or stemma. Although the three cladograms do not look the same at first sight they are derived from one and the same network, as pictured in fig. 18.

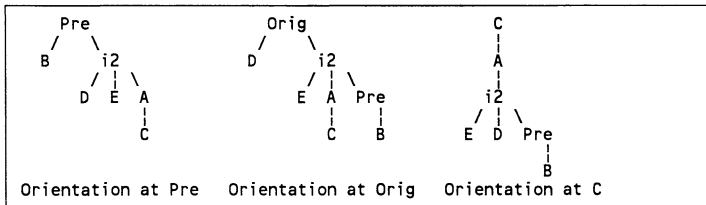


Figure 19.

3.3.4. The length of a Wagner network: criterion for reliability and guard against contamination and parallelism

The shorter the realised network is, and, by inference, the more it approaches the theoretical minimum length of the network, the better a network is. [One can judge the previous sentence as an expression of the 'principle of parsimony', as has been discussed in part 3 of the main paper before the appendices; BS 1994.] To formulate it differently, a network is more reliable or consistent than another network if its *quotient of the minimum length and the realised network length* is smaller and comes closer to the value 1. One of the great advantages of Wagner networks is that we can compare several networks with the use of this quotient. Usually the quotient is called *Consistency Index* (CI). The CI in our example equals 9/11 or 0.8181.

We can now briefly discuss the explanation of step 5 of the Wagner network algorithm in section 3.3.2. Both OTU A and OTU D could be attached to an internode. Had we chosen to attach D to the internode, a different network might have resulted. As a matter of fact, it would have been better if we had drawn all networks which are possible according to the algorithm. Then, with the CI's of these networks we could have picked out the most consistent network. Fortunately, computer programs exist which can calculate quickly the best Wagner network for us, given a certain amount of characters and character states.²³

The difference between the minimum network length and the realised network length is due to the fact that not all character states point in the direction of one single network. Disturbing factors have emerged. This implies either that the chosen characters and/or character states are incorrect or that contamination(s) or parallelism(s) have occurred. The *CI measures the 'static' (or disturbance) which is due to these incorrect, parallelistic or contaminated characters and their states*. The more disturbing these factors are, the smaller the CI will be.

²³ One of these computer programs is 'TreeTools' of Dr. W. Ellis. [= Ellis (n.d.; 1985?). Ellis points out in the first footnote of the first chapter that the PAUP-package at that time of David Swofford comes close to TreeTools: "After the programs were in their final shape, I received the announcement of David Swofford's microcomputer version, for the IBM PC, of his PAUP package, and of Joe Felsenstein's PHYLIP version for MSDOS and CP/M computers. Judging from the description, especially Swofford's program comes close to TreeTools." BS 1994.]

Computer-Assisted Stemmatic Analysis and 'Best-Text' Historical Editing¹

Peter M.W. Robinson

The Canterbury Tales Project

Oxford University Computing Services, Oxford, UK

Imagine that the great German textual scholar Karl Lachmann were alive today. Imagine, too, that he is (as he surely would be) an adept at computing. Lachmann's dream would be that at the flick of a switch, a mighty engine would whirl into action, chewing its way through all the variant texts of some master work, finding variants, sorting, counting, classifying them, and producing, at the other end, by a process faultlessly mechanical, purely scientific, a perfect stemma and, thus, a perfect text.²

¹ Several of the themes in this paper are treated in greater detail in other articles by me: on computer collation (Robinson 1994; cf. Robinson 1989); on cladistic analysis in (Robinson and O'Hara 1992; Robinson and O'Hara Forthcoming) and on this and other computer-assisted stemmatic techniques (Robinson and O'Hara 1993). The concept of best-text historical editing presented here is original to this paper. I am grateful to Professor van Reenen for the opportunity to present this paper at a seminar in Amsterdam, and to the seminar participants for their comments on it.

² The classic statement of the 'scientific' model of editing is Lachmann's declaration (1876: II 276): "Die Feststellung eines Texts nach Ueberlieferung ist ein streng historische Arbeit" [The establishment of a text according to its tradition is a strictly historical undertaking]. See also Timpanaro's discussions of the genesis of the Lachmann 'method' (Timpanaro 1971; 1981). Belief in scientific editing, and attempts to practice it, had a long history before Lachmann: Timpanaro (1963:4-5) points out that in 1489 Politian determined the relationship of several manuscripts of Cicero by recognising a significant shared error and Richard Bentley declared, in a letter dated 1716, his intention to restore the text of the Greek New Testament "exactly as it was in the best exemplars at the time of the Council of Nice" (Wordsworth 1842:503). Among contemporary scholars, Vinton Dearing has advocated the power of the computer to achieve these ends: see particularly his likening of the collator with his computer to John Glenn in his space capsule awaiting lift-off: "I am go; all systems are go" (Dearing

In the last years, developments in the application of computing to textual criticism have shown such promise that one is tempted to proclaim the achievement of Lachmann's dream: we have systems for the computer capture of text; we have systems for computer collation; we have systems, some of them of stunning virtuosity, for generating stemmata from the results of the collation. In the first half of this paper, I will describe some of these systems. On the face of it, it appears that these may fulfill Lachmann's dream of a perfect method producing perfect texts. In the second half, I will report on some of the results of our application of these methods to actual texts: specifically, on the 46 manuscripts of the Old Norse *Svipdagsmál* and the 58 pre-1500 witnesses of Geoffrey Chaucer's Wife of Bath's Prologue. I suggest that far from fulfilling Lachmann's dream, use of these new tools suggests something very different: that the whole stemmatic project is fundamentally flawed. There is no perfect method which will by its mere application guarantee a perfect stemma and, hence, a definitive text. The genie of these new tools offers vast pools of knowledge, but where we might suppose that increasing knowledge and sharper tools might bring us more certainty, they actually bring us only doubt. In place of decision, we have new kinds of indecision.

1. New tools of computer-assisted stemmatics: transcription, collation, analysis

Any process of stemmatic analysis must begin with the collection of information concerning the agreements and disagreements between the witnesses to the text.³ Until a few years ago, all scholars used much the same method. In essence, the editor chose a 'base text', and recorded where witnesses differed from this base text. The base text might be an existing printed edition or manuscript, emended to some degree. Every variation in

1962; 1970). Few scholars have shared this uncritical enthusiasm: compare the explicit abandonment of the historical method by George Kane in his editions of *Piers Plowman* and his forceful criticisms of the Manly-Rickert edition of the *Canterbury Tales* (Manly and Rickert 1940; Kane 1960; 1984).

³ The standard description of the techniques of collation and recension is that of Paul Maas (1958); Martin West (1973) provides a more detailed and practical discussion. Useful historical overviews are given in Reynolds and Wilson 1991, Whittaker 1991 and Kenney 1974. Moorman (1975) suggests techniques for medieval editors and the introduction to Manly and Rickert (1940, volume II), provides an full account of the methods used in a large scale collation.

every witness might be noted, or some selection of variation, according to the type or variant or the status of the witness. The variants might be physically written on the base text, or on carefully-ruled sheets, or on collation cards.⁴ Once these variants had been collected the task of analysis began. By some means of filtering, sorting, and counting, scholars searched for decisive evidence of a stemma: a directory of relationships which might determine the choice of this reading or that. The variants were collected by hand, collated by hand, analysed by hand, with nothing more sophisticated than pencil and paper.

The advent of sophisticated computerised collation tools has (at least potentially) changed this model. Instead of nominating a base collation text in advance and recording a more or less complete selection of variants against that base, one can now transcribe all the witnesses into the computer and then have the computer compare them. Further, one can take the resulting machine-readable record of variation and send that straight to other computer programs which can filter, sort, count and generate stemmata and other analyses.

There are immediate advantages in this procedure. The first is accuracy. When carrying out a collation by recording variants against a base, three things must be done at once: the base text must be read correctly; the correct variant in the text collated must be observed; then the variant must be recorded correctly. The possibility for error multiplies at each stage, as information is sorted, analysed and finally printed. With computer collation, you have to do one thing only: simply transcribe each witness, as accurately as possible, and the computer will do the rest. The second advantage is that the editor need not nominate a base text in advance. At the least, this avoids the inconvenience of having to adjust the collation each time an emendation is introduced into the base: an editor need not be deterred from a necessary emendation by the labour of this adjustment. Further, classical stemmatic theory demands that all agreements by witnesses against the base be agreements in introduced reading. The base must represent at all points the 'original'; only when we have this base can we identify what witnesses have introduced unoriginal readings and so establish a stemma. E. Talbot Donaldson (1970:107) has elegantly pointed out the illogic of this procedure: you cannot establish the original and, hence, the base until you have a

⁴ For collation techniques, see Charles Moorman's description in his handbook for aspiring Middle English editors (Moorman 1975). The introduction to Manly and Rickert (1940, volume II), provides an full account of the methods used in a large scale collation.

stemma (cf. Kane and Donaldson 1975:17, fn. 10), but to discover the stemma, you need to have the original. Donaldson puts this in terms we may refer to as Donaldson's dilemma: if one is able to establish all the original readings in advance, why bother with the tedium of collation and analysis; why not just print the reconstructed archetype and have done with it. As we shall see later, computerised collation and analysis offers a remarkable resolution to this dilemma, as cladistic methods permit the creation of stemmata without prior identification of a putative original.

There are three stages to computerised stemmatic work: the capture of the text of the witnesses themselves into machine-readable form; the collation of the captured text; the analysis of the information generated by the collation. Computerised systems of character recognition have made great advances in the last years and may now be reliably used to capture modern printed texts; however, handwritten medieval manuscripts present a much more difficult case. No computer system has yet come even close to 'reading' these texts, and for some time at least, we will have to transcribe the manuscripts into the computer ourselves. For a large manuscript tradition, this is a daunting prospect: there are some 1.6 million lines and around 8 million words in the 88 pre-1500 witnesses to the *Canterbury Tales*. A complete computerised collation will require transcription of every word of this. Perhaps more than anything, the sheer labour of such a complete transcription has deterred scholars from computerised collation.

This immense labour of transcription offers, however, a signal, and possibly decisive, reward. Once made, the full computer-readable transcripts of each witness may be published in their own right. Though the transcription may be expensive, CD-ROM and network systems make publication cheap. Further, it is a simple matter to link every transcript to digital images, and publish these digital images alongside the transcripts. Other scholars may then read, search and re-collate the transcripts for themselves, as well as easily check the transcript against the image of manuscript itself. Our first *Canterbury Tales* CD-ROM will accordingly include transcripts of all 58 pre-1500 witnesses to the text of the Wife of Bath's Prologue and images of the 1200 pages of these witnesses.

After transcription, the next stage is collation of all the transcribed witnesses. We use *Collate*, a computer program which contains many features specifically designed to fit it for collation of a medieval manuscript tradition.⁵ In particular, this program can collate up to a hundred witnesses

⁵ *Collate* is more fully described in Robinson 1994. Similar features are also offered by Wilhelm Ott's TUSTEP suite of programs (Ott 1992).

at once, so that decisions about what should collate against what can be made in full knowledge of what is in each manuscript. The program also contains powerful spelling regularisation features. Our Chaucer transcripts are rich in information, as they aim to represent the spelling, and much else, of the manuscripts. In our eventual stemmatic analysis we are interested in just those things likely to have been copied from witness to witness: essentially, the 'substantive' variants of different readings, and not the 'accidental' variants of spelling, etc., which might be individual to each scribe. With *Collate* we are able to filter out as we collate the 'substantive' variants from those which appear 'accidental', through the regularisation feature. In traditional collation this filtering of 'accidental' and 'substantive' is done at the first stage of the task. As each witness is read, the scholar selects what will and will not be recorded and there is no means to recall later what has been discarded, short of redoing the entire collation. Not only can we do this filtering as we collate all the witnesses together, when we have complete access to the full text of every witness, but we can alter our decisions, about what and how we level, at any point. Further, the information about spelling, etc., which we filter out is not simply discarded. It is kept in a database, and from this we generate a complete record of every spelling in every witness, with all spellings grouped under appropriate lemmata and sorted by witness. This opens up immense research possibilities, as we can explore the different spellings by dialect, by scribe, by date or by any combination of these and other criteria.

As with the transcripts, the electronic form of the collation makes publication on CD-ROM and network easy and cheap. The collation may be redone, over and over, with different base texts and with different selections of texts; it can be output in various formats as desired. Our first *Canterbury Tales* CD-ROM will publish several such collations in several such formats. It will also publish regularisation databases in which all spellings in all witnesses are grouped under appropriate lemmata. One of these databases will contain all the spellings sorted first by lemmata and then by witness; other databases (one for each witness) will contain all the spellings sorted first by witness and then by lemmata.

Because computer collation will tell us of every variant on every word in every witness, it generates huge quantities of information. In the case of the Wife of Bath's Prologue, the collation gives us around 250,000 items of information. More than anything, it is the capacity of the computer to cope with such floods of data, where manual sorting techniques cannot, which fits them to this work. From *Collate*, we can pass all the information about the agreements and disagreements between the manuscripts directly to a variety

of specialised computer tools for this analysis.⁶ We have found two complementary methods particularly useful. The first of these is cladistic analysis, which we use to provide a 'road-map' of the witness relations: that is, a preliminary picture of how the witnesses might be related. Our second method is database analysis, by which we examine in fine detail exactly what readings are found in what witnesses, in order to refine as exactly as we can our understanding of how the readings have flowed from one witness to another, and hence how they are related.

Cladistic analysis has been developed over the last thirty years by researchers in the field of systematics, the branch of evolutionary biology which specialises in the reconstruction of the evolutionary tree of life. There is a fundamental identity between the assumptions, aims and methods of cladistic systematics and stemmatics. Both disciplines assume that the existence of a varied collection of objects (manuscripts for stemmaticists, organisms for systematists) results from a sequence of branching descents over time from a common ancestor: what Darwin (1859:117) calls "descent with modification". Both disciplines aim to represent this evolutionary history for any group of related objects (whether manuscripts or organisms) in a tree of ancestry and descent. Both disciplines seek to construct this tree by study of the characteristics the objects share and do not share. Because the concepts and methods of cladistic analysis are explicitly historical in character (O'Hara 1988), they can be readily adapted to the reconstruction of manuscript stemmata.

A simple, actual example should illustrate the closeness of the disciplines, and hence the ease with which cladistic methods may be adapted to stemmatics. There are 210 known species of woodpeckers. Presume that sixty species have four toes, while 150 have only three toes. Cladistics operates on two central presumptions. The first is that one of these two different types is ancestral, with the other descended from it. The second assumption is that the simplest explanation of descent is to be preferred: the principle of parsimony. Let us concentrate, for the moment, on this second assumption, of the preference for the simplest explanation. In cladistic terms, the

⁶ A number of authors have suggested that computer-assisted techniques might prove valuable in pointing quickly to possible relationships which could then be thoroughly examined by other means. The first scholar to have tried such a computer-assisted approach appears to have been John Griffith (1968; 1984), who applied cluster analysis to some variant readings in the manuscripts of Juvenal. Most subsequent studies of this type have also used either cluster analysis or multivariate analysis (see the reviews in Hockey 1980 and Pierce 1988). For reasons that will become apparent, we have found these statistical clustering techniques less satisfactory than cladistics.

'simplest explanation' is the one that posits the fewest 'changes of state'. In this case, the shift from the four-toed species to the three-toed species (or vice versa) is a change of state. One could suppose that every one of the (supposed) sixty four-toed woodpeckers evolved quite independently of all the others: that is, there are sixty 'changes of state'. Or, one could suppose that at some point in time one, and only one, four-toed species evolved from a three-toed species and that every other four-toed species is descended from this single ancestor. In this second case, there is only one 'change of state', with every one of the four-toed species descending from this intermediate ancestor. In cladistic terms, an ancestral tree which suggests only one 'change of state' is 'more parsimonious' than a tree which suggests sixty changes of state.

In essence, cladistic method surveys the possible ancestral trees, spreading the information about agreement or disagreement across the trees until it finds the tree which accounts for the observed distribution of character states among the descendants in the simplest manner. The number of possible trees that must be evaluated for their fit to the data can rapidly become enormous. The task of analysis will also be vastly complicated when the characters conflict with one another owing to evolutionary convergence. It is here that computer programs are of assistance, and several tree evaluation programs have been written and are in wide use in the systematics community. In our work, we have used David Swofford's *PAUP* program: Phylogenetic Analysis Using Parsimony. This program is flexible and fast, offering a comprehensive range of output options.⁷

In the same way that cladistics accounts for shared agreement by presuming common ancestry, a manuscript scholar accounts for shared readings by presuming common ancestry. In the first line of the Wife of Bath's Prologue 42 manuscripts read 'Experience' and 10 read 'Experiment'. We may presume that the ten manuscripts which read 'Experiment' are all descended quite independently of one another, and that they all happened on this reading quite separately: that is, there are ten 'changes of state' from the reading 'Experience' to 'Experiment'. Or, we may presume that the ten manuscripts are descended from a single hyparchetype which introduced the reading 'Experiment'. Cladistics would regard the latter explanation as simpler and, therefore, preferable; so, too, would a stemmaticist.

Let us now consider more closely the first assumption cladistic method

⁷ See Mayr and Ashlock 1991:320-321 for a recent listing; for comprehensive introductions to cladistic analysis see Sober 1988; Brooks and McLennan 1991; Maddison and Maddison 1992.

would make concerning the three- and four-toed woodpeckers: that one of these types is 'ancestral', and the other 'derived'. The ancestral or primitive state will have been present in the ancestor of the whole group originally, and could potentially retained unmodified anywhere. Presume that the ancestral state is three-toed: the fact that a woodpecker has three toes shows only that it is descended from the ancestor of all woodpeckers, which also happened to have three toes. Thus, it tells us nothing. On the other hand, if the ancestral state is to have three-toes then a four-toed woodpecker is likely to be closely related to other four-toed woodpeckers: they would share a common ancestor which introduced the four-toed state.

Stemmaticists will immediately recognise this cladistic principle, that only the derived states of characters identify tree branches, as identical to the stemmatic principle that only 'introduced', 'unoriginal' readings are evidence of manuscript relations. This was spelt out by Lejay in 1888 (reported in Kenney 1974:135), and has been repeated by many others (West 1973:32-33; Kane 1984:209; cf. Quentin 1926:61-96). Cladistics faces exactly the same problem as stemmaticists: how does one determine what is 'ancestral', what is 'introduced'? Is it the four-toed or the three-toed woodpecker which is the ancestor? Did Chaucer originally write 'Experience' or 'Experiment'?

It may be the most remarkable contribution of cladistic analysis to stemmatics that it has thought through this problem and offers a way past it. Traditional stemmatics tries to identify which readings are ancestral at only one point of the manuscript family tree: the single root or ultimate ancestor. Cladistic analysis, by contrast, and in keeping with the goals of systematic biologists, aims not simply to reconstruct the ultimate ancestor but the whole history of the tradition, including the attributes of each ancestor throughout the tree. This has a remarkable and most powerful consequence. It allows systematists who are uncertain about which of their character states are ancestral and which are derived to regard their trees initially as 'unrooted', and then to orient those trees in any particular direction based on whatever partial information on the ancestral conditions they may have available. As a result of this practice all groupings (clades) will be defined by introduced variants no matter which way the tree is rooted, and there is no need to specify beforehand just which variants are ancestral to the tree as a whole.

To see how this works, consider the two possible trees cladistic analysis would give us for the three- and four-toed woodpeckers. In this first tree, we decide that the ancestor is the four-toed species:

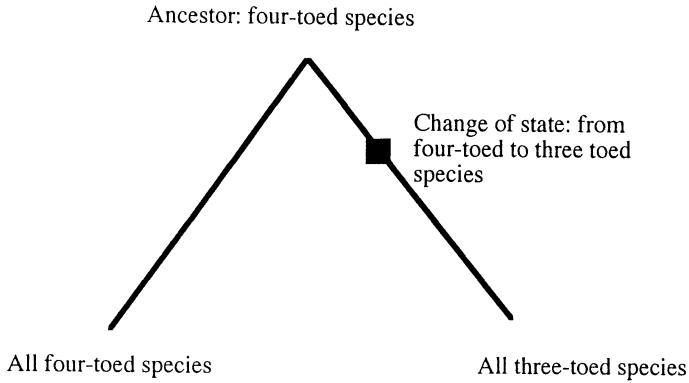


Figure 1: Change of state from four-toed to three-toed species

In this second tree, we presume that the ancestor is the three-toed species:

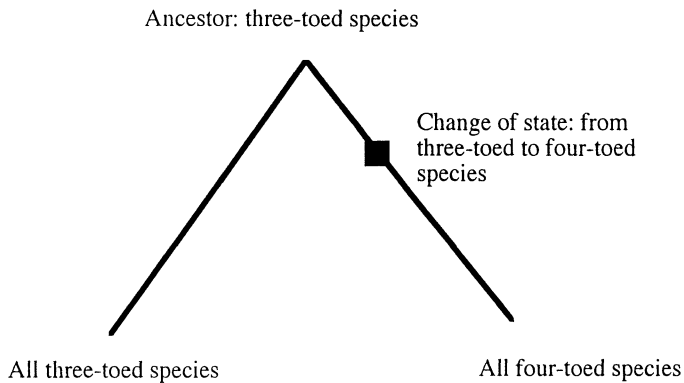


Figure 2: Change of state from three-toed to four-toed species

Now, it is apparent that in cladistic terms both trees are *equally parsimonious*. Both trees presume that there was one, and only one, change of state: the change between the three-toed and the four-toed state. In effect, both trees are identical in that both see the change between the three- and four-toed state as occurring at a single sub-ancestor. The only difference between the trees is their orientation. One tree places the three-toed type at the apex, and sees the change of state as occurring from three-toed to four-

toed. The other places the four-toed type at the apex, and sees the change of state as occurring from four-toed to three-toed. In both trees, parsimony fixes the change between three toes and four toes as occurring at a single point. Regardless of how we orient this tree, that relationship, of the three- and four-toed types being linked by a single common sub-ancestor, remains fixed.

This gives the clue as to how stemmaticists can use cladistic methods as a way past Donaldson's dilemma. We can use cladistics to find the most parsimonious tree for us, without having to specify in advance which readings are ancestral, and hence which way this tree is oriented: in cladistic terms, this is an 'unrooted tree'. All groupings in this tree will be defined by introduced variants no matter which way the tree is rooted, and there is no need to specify beforehand just what variants are ancestral to the tree as a whole. Observe that this tree may contain many groupings defined by descent from sub-ancestors. We can then use whatever information we have to determine which of these sub-ancestors is nearest to the putative original, and then re-root the whole tree with that sub-ancestor as ancestor to the whole tree.

This is the method so far explored and proposed for the *Canterbury Tales* Project. First, one makes an unrooted tree on the basis of a table of the manuscripts and their variants, deferring judgement on just what variant readings are ancestral to the whole tree until one has this unrooted tree. Then, one can decide which of the branches of the tree lies closest to the archetype and root the whole tree near this branch. Especially, one can scrutinise the variants introduced at each hypothetical node. Where there appears a possibility that a particular group of variants introduced at a particular node might be authorial, this group can be isolated and studied in further detail.

The second computer tool we use for stemmatic analysis is database analysis of the corpus of variants. This analysis is necessary for two reasons. First, cladistic methods, powerful as they are, may only be able to reconstruct the general outlines of the relations of the manuscripts, and may not be able to determine all the exact details. Second, we know from our work with the *Svipdagsmál* tradition that cladistic methods may be defective in their treatment of contamination and convergent variation (see below). Database analysis allows us to answer, in an instant, such questions as what readings are found in manuscripts A, B and C but not in D, or in any two of A, B and C, for all or any nominated section of any text. This information might take hours, even days, to extract from a manual collation record; it is available within seconds from a database. Further, given an

accurate transcription, the answers returned from the database will be complete and accurate.

2. The use of these methods: their success

The test of any method is whether it works. In stemmatics, the test of a successful method is that it should produce results which can be independently confirmed. The best kind of confirmation is from external, documentary evidence. For example, the method might, from pure analysis of the patterns of agreements and disagreements within the manuscripts, indicate relationships which independent, external evidence (information about the provenance of manuscripts; their owners, copyists, exemplars, etc.) might confirm or deny. The use of these methods, cladistic and database analysis, with the manuscript tradition of the Old Norse poetic narrative sequence *Svipdagsmál* provided remarkable evidence of their potential.

Svipdagsmál consists of the poems *Gróugaldr* and *Fjölsvinnsmál*, together about 1500 words in length, and is extant in forty-six manuscripts written between 1650 and 1830 (see Robinson 1991 for an edited version of *Svipdagsmál* as a doctoral thesis at Oxford). Two articles in *Literary and Linguistic Computing* discuss the computer methods (chiefly collation and database analysis) developed to assist this edition. These methods were the predecessors of those now developed within *Collate* and described in this article. The *Svipdagsmál* tradition is notable because of the full external evidence we have for the provenance of, and relationships between, the manuscripts. The manuscripts are late, and the scribes and the owners of the manuscripts in Iceland, Denmark and Sweden were in the habit of writing into the manuscripts information about their owners, exemplars and history. For sixteen of these manuscripts, we know from this external evidence alone exactly how they are linked. Figure 3 shows the stemma of the manuscripts I made for these manuscripts using traditional philological methods, notably collection of external evidence about the relationships among the manuscripts. The manuscripts which are linked by arrows are the sixteen manuscripts for which there is clear external evidence that they are related as given. This external evidence provides the opportunity to judge decisively the validity or invalidity of these methods. Observe particularly the group at the left of the stemma: St, 34, 1870, Gu, 682 and 11. The arrows between every one of these manuscripts (except the pair St/Gu) are based on clear statements in the manuscripts themselves. Indeed, there is strong external support for the supposed descent of Gu from St: we know that Gunnar

Ólafsson brought St (Stockholm Papp 15 oct.) to Stockholm from Denmark in 1681, and that Gu was copied by Gunnar in Stockholm between 1681 and 1686. For the pair Gu/St, however, we do not have a clear, unequivocal statement in the manuscripts themselves that Gu is copied directly from St. For every other manuscript in this group, we do have statements in the manuscripts themselves about exactly what was copied from what, and so these are linked by arrows in the stemma.

Evidence of this nature provided an ideal testing ground for computer-assisted stemmatic methods. Two instances will show how this tradition was able to verify the results of computer methods. The first instance concerns the manuscript designated He/X5 on the stemma, there shown as descending from Gu. In the early stages of my work with this tradition, I was puzzled by this manuscript, MS 1562 Oct. in the National Library of Iceland. My first impressions of it, based only on a microfilm of part of it and on a cryptic catalogue dating '1660', was that it might be the oldest manuscript of *Svipdagsmál*, and possibly the source of all other manuscripts. Database analysis, based on simple counts of agreements among the manuscripts, however, appeared to suggest that He/X5 was most closely related to four other manuscripts: to Gu and 682 (in the group immediately above He/X5 in the stemma) and to L and 47 (in the group below He/X5 in the stemma). This seemed to contradict the apparent evidence that this might be the oldest manuscript, and that it was written in Iceland around 1660: Gu and 682 were both written in Sweden between 1681 and 1686, and L and 47 in Iceland about 1750. When I examined He/X5 itself in the National Library of Iceland it became apparent that the database was exactly right: He/X5 is indeed closer to 46 and 682 than any other manuscripts. I found from examination of the whole manuscript what had not been apparent from the microfilms: He/X5 was actually written by Helgi Ólafsson, the brother of Gunnar Ólafsson. From external evidence I knew that Helgi was in Sweden between 1682 and 1686, that he and his brother had copied many manuscripts, and that 682 was actually Helgi's copy of Gu. It appears that before Helgi returned to Iceland, he made yet another copy of Gu and took that copy with him: He/X5 is that copy. L and 47, and indeed several other manuscripts written in Iceland around 1750, were all descended from this copy. Hence the database was correct in suggesting that He/X5 lay between Gu and 682, written in Sweden in the 1680s, and L/47, written in Iceland in the 1750s. In fact, He/X5 forms a crucial link in the stemma and explains what otherwise appears inexplicable: the descent of manuscripts written in Iceland after 1750 from manuscripts written in Sweden before 1690. It is possible that I might have arrived at this conclusion in any case, once I had

examined He/X5, but the database analysis was invaluable as it suggested a line of inquiry which proved immediately fruitful, and which might otherwise have taken some time to discover.

The second instance of the use of this tradition to establish the success of computer-assisted stemmatic methods has an unusual history. By mid-1991, I was satisfied that the stemma shown in figure 3 was substantially correct, or as correct as traditional analysis supplemented by the database methods I described in my 1989 *Literary and Linguistic Computing* articles could make it. Nonetheless, I was concerned that while both database and traditional analysis could yield excellent results once one knew where to look, it could take a long time to find out where to look. The He/X5 example discussed above was a spectacular success of just how illuminating the combinations of these methods could be. I had spent much time exploring and eliminating other possibilities, but there was always the suspicion that, in so complex a tradition, there were possible relationships I had overlooked altogether. Thus, I began to seek other methods: in particular, I wanted some method capable of building hypothetical models of relationships which could point out exactly which manuscript groupings needed further analysis.

In August 1991 I placed a challenge on the computer networks: I would make available all the raw information about all the agreements and disagreements among the manuscripts of *Svipdagsmál*, and I would buy lunch for anyone who could produce, using computer methods alone on this raw data, something close to the stemma I had spend five years making. There were nine responses to this challenge, with all but one using some variety of statistical grouping technique. The ninth method, applied by Dr. Robert O'Hara (then of the University of Wisconsin at Madison), used cladistic analysis and was far the most successful (Robinson and O'Hara 1992; Robinson and O'Hara Forthcoming).

The table in Figure 4 gives the family tree, or cladogram, for the manuscripts created by Dr. O'Hara using the cladistic program *PAUP* in its run over the raw collation data. Comparison of this with the stemma in Figure 3 shows the accuracy with which *PAUP* replicated the outlines of my stemma. First, the sixteen manuscripts which external evidence showed as directly related to one another: each of these manuscripts is placed close to its known relative, usually adjacent to it. Note especially its success with the three manuscripts St, 34 and 1870, and the three manuscripts 1689, 5 and 329, all written by one scribe: these are placed directly adjacent in the cladogram just as they are in my stemma. Second, there are major groupings of manuscripts having relationships with one another and with individual key

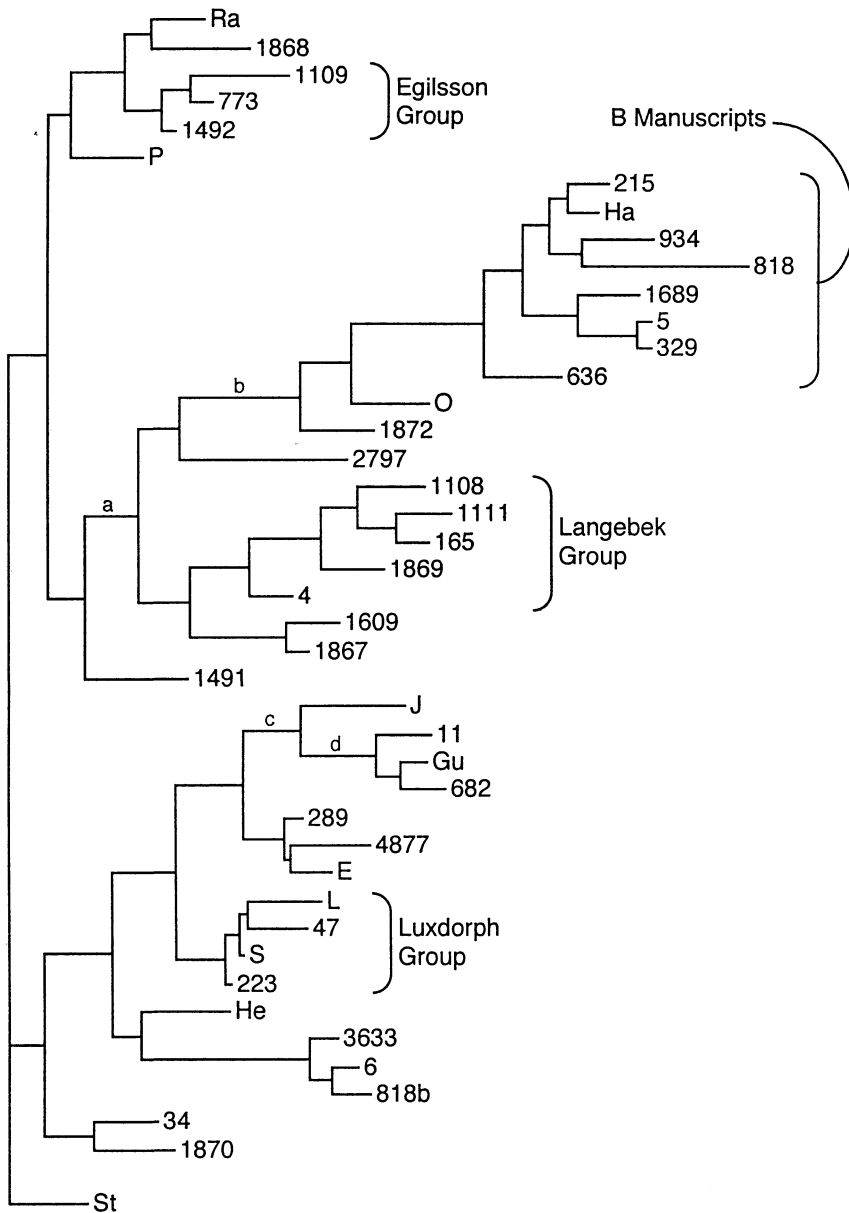


Figure 4.

manuscripts. The cladistic analysis identified all these correctly. For example, the Egilsson group manuscripts are placed close to Ra in the cladogram, with one of them, 1868, separated by just one node. Without *PAUP*, establishing the closeness of these manuscripts to Ra took considerable effort. Another example can be seen in the B manuscripts, the group on the right in my stemma. After much effort without *PAUP*, I had decided that the B manuscripts all descended from O, with another manuscript, 1872, descending on a different branch from O. That is much how *PAUP*'s analysis places them in the cladogram, with O and the B manuscripts appearing as coordinate branches (called 'sister clades' in systematics), and with 1872 sister to O and the B manuscripts taken together. Consider, too, the three manuscripts 818b, 3633, and 6: Figure 4 shows these forming a subgroup of their own within the larger Stockholm group, and that is just how they appear in the cladogram.

My study of the manuscripts St and Ra, summarised in the stemma in Figure 3, revealed their fundamental importance in the evolution of the *Svipdagsmál* tradition. Some two-thirds of all the manuscripts, thirty-one of the total forty-six, appear to have descended either from St or Ra, or a manuscript (the hypothetical X3) close to Ra. Thus, although St and Ra are similar in absolute terms, in evolutionary terms they are far apart: they represent the twin roots from which most of the manuscripts derive. The cladistic analysis manages to show this: Figure 4 places five nodes between St and Ra, and from these five nodes all the other manuscripts descend.

3. The limits of computer methods: contamination and cladistic analyses

Although the cladistic analysis of the *Svipdagsmál* material was successful in discriminating all the major manuscript groups and in fixing the relations of some of these groups to others and to individual manuscripts, it was not correct in all details. Its greatest difficulty was caused by manuscript contamination, the deliberate importation of readings from one manuscript into another that is not its copy. Contamination takes place 'horizontally' across a stemma, rather than 'vertically' from ancestor to descendant, and cladistic analysis effectively assumes that instances of horizontal transmission will be outnumbered by instances of vertical transmission. Though this is broadly true of the mass of variants in most manuscript traditions and is hence generally successful with the *Svipdagsmál* material, there may be subgroups of variants in subgroups of manuscripts that have been much influenced by horizontal transmission. There are, for example, a

large number of variants found as marginalia in several groups of *Svipdagsmál* manuscripts which appear to have been borrowed from the text of other distinct groups, and the inclusion of these variants led to some deformation in the tree produced by the cladistic analysis. As a case in point, because of large scale contamination of the Langebek manuscripts by B manuscript readings, the Langebek manuscripts appear far closer to the B manuscripts in the cladogram in Figure 4 than they should. This incorrect placement of the Langebek manuscripts had other, potentially serious, consequences. My analysis of the manuscripts suggested that the B group had descended from Ra, or a manuscript close to Ra, probably via manuscript O. The evidence for this is a set of some twenty-six readings found in Ra, also in O, and thence characteristically in the B manuscripts. In order to accommodate the Langebek manuscripts (none of which have any of these twenty-six readings) somewhere between Ra and O in the cladogram, *PAUP* had to suppose that these twenty-six readings were first removed along the branch marked *a* (hence their absence from the manuscripts below that point, including the Langebek group), and then restored along the branch marked *b* (hence their presence in the manuscripts below that point, including O and the B manuscripts). This obscures the most likely flow of readings and makes Ra, O and the B manuscripts appear rather more distantly related than they actually are.

The cladogram also places the Gu 11 682 group rather further from their immediate ancestor St than it should. Indeed, it actually suggests that these manuscripts might descend from He/X5 (here simply He), where it appears that the truth is the reverse: He is copied from Gu, not Gu from He. The reason for this error has to do with manuscript J. My stemma shows that J descends quite separately from either St or Ra, and appears among a group of manuscripts shown at the centre of the table which lack any strong affiliations with any other manuscripts. The cladistic analysis, however, has grouped J with the three manuscripts Gu, 11 and 682; not only this, it has moved Gu, 11 and 682 much further away from their direct ancestor St than is correct. Examination of the variants that *PAUP* judged to have been introduced along the branch marked *c* in the cladogram revealed the reason for this error. In the second poem of the *Svipdagsmál* sequence, a question formula is repeated eighteen times over. Most manuscripts abbreviate this formula in one way or other, some giving the initial letter of each word, some just giving the first one or two words, and so on. Four manuscripts alone spell out every word of the whole question sequence on each repetition: they are the four manuscripts J, Gu, 11 and 682. Clearly, 11 and 682 have simply inherited this from Gu. Clearly, too, in view of the lack of

any other evidence linking J and the three manuscripts Gu, 11 and 682 — J has only six of the twenty-eight variants which characterise the manuscripts descended from St, while Gu, 11 and 682 have, respectively, twenty-five, sixteen, and twenty-five — it is simple accident that the scribe of J happened to spell out every instance of the formula just as the scribes of the other three manuscripts did. Nonetheless, this accident has caused the group Gu, 11 and 682 to be placed next to J and much further away from their direct ancestor St than is correct. Once more, this distorts the flow of readings: it requires us to suppose that most of the St variants present in Gu, 11 and 682 were removed before point *c* and then restored along the branch marked *d*.

4. The limits of stemmatics: manuscripts close to the original

The difficulty cladistic analysis had with manuscript J points towards a fundamental problem not just with cladistics, but also with stemmatics. As I remarked above, the problem of classifying this manuscript, and these other so-called ‘ungrouped’ manuscripts, is that they lack any strong affiliation with any other manuscripts. In cladistic terms: they share few ‘derived’ character-states with other objects. It is revealing to see what the cladistic analysis did with these manuscripts: it scattered them, almost at random, across the whole cladogram. Thus, J is here with Gu 682 11; 1491 1609 1867 are in the group immediately above; 2797 elsewhere in this same group; P is at the top of the table. We have seen that in placing J here, the cladogram was misled by the coincidence of identical treatment of a question formula in J and these three otherwise unrelated manuscripts. The placement of the other ‘ungrouped’ manuscripts is likely to be similarly artificial.

In traditional stemmatic terms: these manuscripts share few unoriginal readings with any other manuscripts. We saw how cladistics tried to cope with these. In traditional stemmatics, where do we put such manuscripts in a stemma; how do we use them when we edit? In my stemma, I had no better solution than to place them in a separate category, labelled ‘manuscripts not included in any group’. Insofar as this suggests that these manuscripts form a single line of descent, united by some common unoriginal readings, this is misleading. They form no such single line of descent; they share no unoriginal readings. In fact, every one of these six manuscripts ought to be regarded as a single line of descent from the original, and potentially of equal authority with any other line of descent. With this in mind, when we

look at my stemma it is difficult to see how we could use this in any kind of mechanical manner to determine the originality of any one reading. In traditional stemmatics, one would determine that there were (say) three lines of descent, and establish at every point the reading in each of these three. When two of the three agree against the third, we simply accept the majority reading. If the notion of 'scientific', 'objective' editing has any validity at all, then this is what it must mean (though I have yet to discover any editor who actually does this). Nonetheless, it is difficult to determine, from my stemma, even how many lines of descent there are. There are at least three, represented by the three major branches of the stemma, but the line headed by manuscript O, at the very right of the table, may actually represent a fourth distinct line of descent, and every one of the six 'ungrouped' manuscripts may also represent a distinct line of descent: a total, then, of some ten lines of descent. While a model of simple majority verdict might apply when there are just three lines of descent, there will be many cases with ten lines of descent where the witnesses simply do not divide so easily. In this case, knowledge leads not to resolution but irresolution: 'scientific' editing is incapacitated by too many possibilities.

5. The manuscripts of the Wife of Bath's Prologue

In some respects, the *Svipdagsmál* tradition is an unusually simple one. The two poems are minor and self-contained, and scribes felt little inclination (with dramatic exceptions) to do more than copy them. All the extant manuscripts appear to descend from a single lost exemplar, probably a seventeenth century copy of a medieval exemplar. There is no question of competing authorial versions and the quantity of text and variation is not extensive.

None of these considerations apply to the manuscripts of Geoffrey Chaucer's *Canterbury Tales*. There are eighty-four manuscripts and four pre-1500 printed editions of a collection of tales amounting to around 20,000 decasyllabic lines in a full manuscript: some 130,000 words. As well as many variants within the text, scribes reordered and annotated the tales in many different ways. Further, Chaucer left the tales unfinished, and it is quite uncertain as to what degree (if any) the differences among the witnesses reflect differing stages of Chaucer's composition. In quantity and type of variation, the *Canterbury Tales* tradition represents a problem several orders of magnitude greater than *Svipdagsmál*. We could not, then, be certain that the computer methods which seemed to have served well on

the *Svipdagsmál* material would work well on the larger and more difficult *Canterbury Tales* tradition. The first stage of our work, then, was to try out these methods on a part of the *Canterbury Tales*. We chose for our experiment the Wife of Bath's Prologue, which contains around 850 lines and exists in 58 fifteenth-century witnesses: 54 manuscripts and 4 printed editions. This is a manageable amount to transcribe and collate for a test project. Further, there had already been an attempt by John Manly and Edith Rickert to establish a stemma for the Wife of Bath's Prologue on the basis of a collation of all the manuscripts, as part of their massive *The Text of the Canterbury Tales*. Comparison of our results with theirs would provide a useful check. Further, serious questions had been raised by scholars regarding the adequacy of Manly and Rickert's methods and the utility of their results (Kane 1984). We might discover from this sample if their work had been well-done. If we found that it were not and that our methods gave promise of better results, then we would have reason to carry on our efforts with the Wife of Bath's Prologue and then on into other parts of the *Canterbury Tales*.

By early 1993, we had completed first transcriptions of 46 of the 58 Wife of Bath's Prologue witnesses, sufficient to commence serious work with collation and the cladistic and database methods. The results of this analysis are described in a paper in *Computer-Based Chaucer Studies* (Robinson 1993), and I will here briefly summarise these results.

First, cladistic analysis showed remarkable inconsistencies in the affiliation of certain key manuscripts over the 850 lines of the Wife of Bath's Prologue. Any explication of the textual history of the *Canterbury Tales* must rest on successful explanation of the textual relations of four vital early manuscripts:

Hg (the Hengwrt manuscript, now Peniarth 392D in the National Library of Wales);

El (the Ellesmere manuscript, now Huntington Library 26.C.9);

Ha4 (British Library Harleian 7334); and

Cp (Corpus Christi Oxford MS 198).

Scholars this century have divided between those who would base their text on Hg (Ruggiers 1979; Blake 1980) and those who would base their text on El (Robinson 1957; Benson 1987). The two texts differ in some four thousand places, with El containing a whole tale and many lines not in Hg and also having the tales in a different order. There has been no convincing analysis of the relationship between El and Hg, or why they differ so. Further, most scholars believe that both manuscripts are written by the one

scribe, known as 'hand b', though this has been recently doubted (Parkes and Doyle 1978; Samuels 1988): if the two were indeed written by one scribe this makes the differences between the two rather more extraordinary. Further, when El differs from Hg it often, but not always, agrees with Ha4. Cp is possibly the earliest manuscript of the *Canterbury Tales* but actually differs quite radically from the other three. Yet Cp and Ha4 were also, most scholars believe, written by the one scribe, the so-called 'hand d' (Parkes and Doyle 1978). Again, this attribution of the two manuscripts to the one scribe has been doubted by Samuels. Unravelling the links between these four manuscripts, probably written by just two scribes, might provide a key to unlock the whole tradition.

Thus, we began by looking closely at what cladistic analysis showed us for these four manuscripts. Figure 5 gives our cladogram for the 46 manuscripts for the first half of the poem (by 'poem', I mean the Wife of Bath's Prologue). Figure 6 gives the cladogram for the same 46 manuscripts for the second half of the poem.

What is remarkable here is the shift in position of El between the two halves of the poem. In the first half, El is close to Gg and Ha4 and some distance away from Hg, but in the second half, El moves right next to Hg (so close, indeed, as to be almost indistinguishable from Hg). What has happened? When we came to carry out database analysis of exactly what readings occurred in what manuscripts, we discovered that in the first half of the poem Ha4, El and the third manuscript Gg shared 22 readings against Hg. In the second half of the poem, these same three manuscripts (Ha4 El Gg) shared only two readings against Hg. Indeed, it appears that Ha4 and Gg stay quite close together throughout the second half, as indeed through the whole poem. After around line 400, El leaves them and agrees with Hg. Meanwhile, we see that Cp does not appear to move at all: it stays closely related to the same manuscripts, La Sl2 Pw Mm, etc., in both halves.

Database analysis also produced striking evidence that the attribution of the two pairs of manuscripts, Hg/El and Cp/Ha4, to two separate scribes is sound. We are able to collate the original-spelling transcripts against each other and from this calculate the degree of agreement in spelling. We did this for these four manuscripts, and also (for control purposes) included two other early manuscripts, Dd and Gg (Cambridge University Library Dd.4.24 and Gg.4.27) in the collation and comparison. Over the 856 lines of WBP we found some 958 places in which Hg and El agreed against the other four MSS and 777 places where Cp and Ha4 agreed against the other four: a far higher proportion of agreement than that achieved by any other pair of these

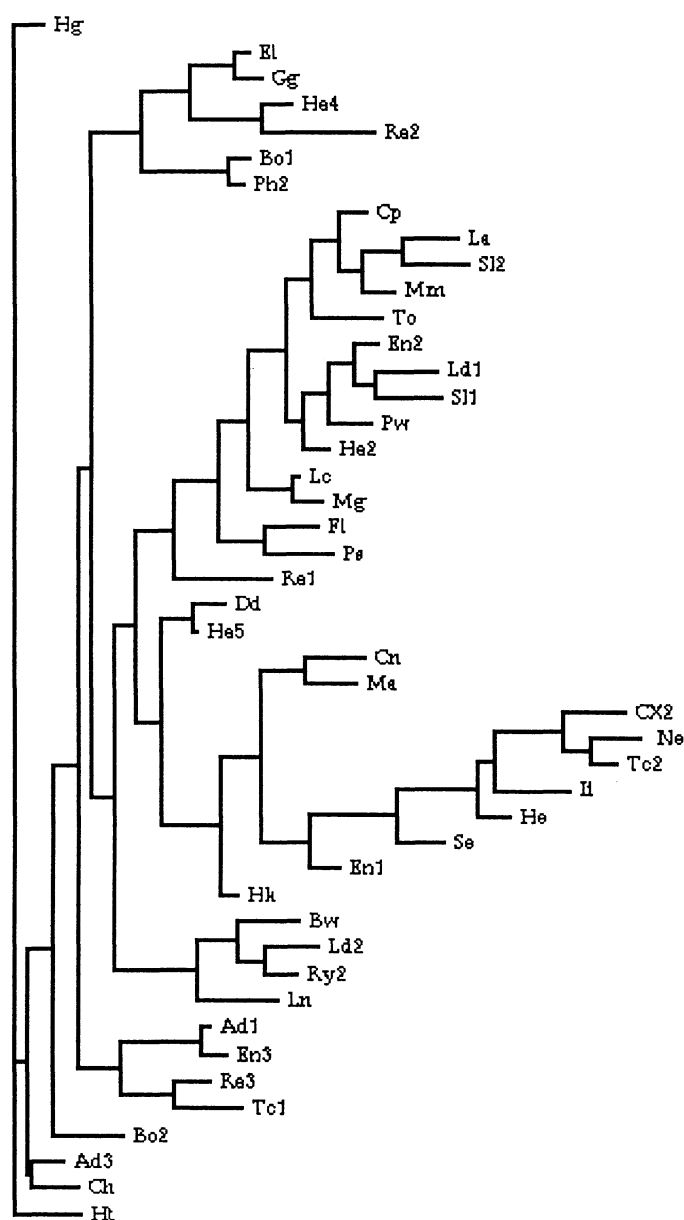


Figure 5: Cladogram for 46 manuscripts of the *Wife of Bath's Prologue* for the first half of the poem

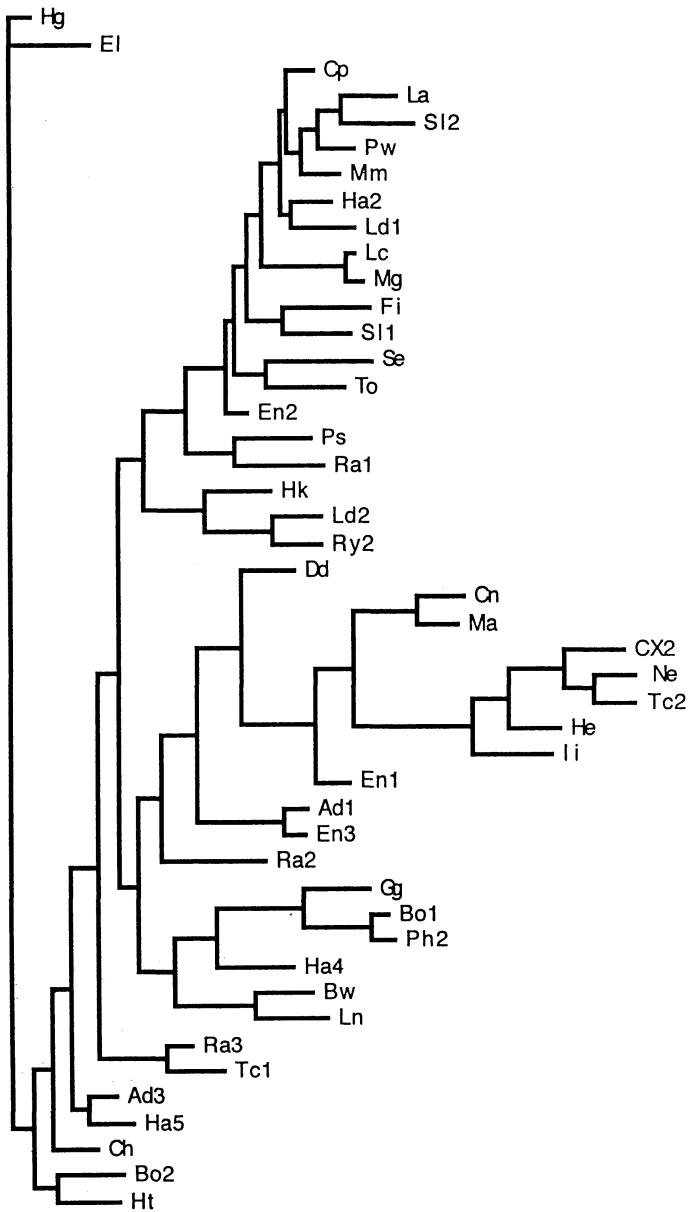


Figure 6: Cladogram for the same 46 manuscripts for the second half of the poem

six MSS.⁸ These agreements are virtually all in spelling or punctuation: of the first fifty agreements of Hg/El against the other four, forty-seven are spelling, three punctuation. This remarkable and persistent agreement in these spellings of these manuscript pairs against the other manuscripts, and especially against these closest to them in date, must weigh in favour of the argument for the two pairs of manuscripts being written by just one scribe.

This returns us to the central problem: what is the relationship between these four manuscripts, and, especially, between Hg/El/Ha4? Simple numerical counting of agreements suggests strongly (as we saw above) that Hg and El are written by the same scribe, and that El is close to Ha4 in substantive readings in the first half and then leaves it to come close to Hg in the second half. To go further, we have to look at the variants themselves: at their kind, at their distribution across the whole text. Here, a discovery by my colleague, Elizabeth Solopova, provides a vital clue. In an article for the next *Canterbury Tales* Project *Occasional Papers* volume, she suggests that the ‘hand d’ scribe responsible for writing both Ha4 and Cp was metrically gifted, and inclined to intervene in the text to produce metrically smoother lines (Solopova Forthcoming). In contrast, it appears that the two manuscripts Gg and El shared an exemplar which contained many unmetrical lines which are not found in Hg. The text of Ha4 also appears to be based on this same exemplar, but ‘hand d’ frequently ‘improves’ the metre where it is defective in El and Gg. An excellent instance is line 308, where Hg reads the metrically correct:

But tel me this , why hidestow with sorwe

El and Gg both omit the ‘this’, creating the metrically irregular:

But tel me , why hidestow with sorwe

It appears that ‘hand d’ saw this in his exemplar, and restored the metre thus in Ha4:

But tel me wherfor hidestow with sorwe

The supposition that the ‘hand d’ scribe was a skilled metrist unafraid to intervene in the text has, since Dr. Solopova’s article, gained further support from our transcription of the full text of the *Canterbury Tales* in Cp and Ha4. We have found that on several occasions the ‘hand d’ scribe writes the

⁸ Thus 203 for Gg/Dd, 154 for Gg/Cp, 49 for Hg/Gg, 47 for Cp/Hg, 46 for Cp/El, etc. These figures, and all other figures in this article relating to counts of agreements between manuscripts, are derived from the database facility within *Collate*.

second line of a couplet first, and then discovers his error before he writes the next line of the text. Most scribes mend this by simply writing the first line out of order, with marginal 'a' and 'b' signs indicating the correct order in which the lines should be read; however, 'hand d' is so confident of his metrical ability, and apparently so loath to spoil the appearance of his text, that he does not do this. Instead, he invents a new line to finish the couplet. Usually, these lines do not make very good sense, but they are in most (if not all) cases metrically regular.

All this has given us what we think may be the first outlines of a convincing and clear account of the relationships among the crucial early manuscripts of the Wife of Bath's Prologue. If we can achieve that, then the way is open to an account of the relationships of all the manuscripts of all the *Canterbury Tales*. From this alone, it appears that 'hand b' used different exemplars for the two copies (Hg and El) he made of the Wife of Bath's Prologue. The exemplar he used for El contained many unmetrical readings in the first half of the Wife of Bath's Prologue; these he copied into his text, and they were copied, too, into Gg. The scribe of Ha4, using this same exemplar, corrected many of these unmetrical readings.

The cladograms, which first directed our attention to the manuscript grouping El/Ha4/Gg, suggesting that these share a different exemplar to that of Hg, may also provide clues to manuscripts which may (unlike El, in the first half of the Wife of Bath's Prologue) have the same exemplar as Hg. These are the group of some eleven manuscripts which appear close to Hg in both cladograms: Ht Bo2 Ch Ha5 Ad3 Ra3 Tc1 Ad1 En3 Bw Ln. As we made different cladograms over different regions of text, we discovered that these manuscripts did not settle in any particular grouping. They would usually appear quite close to Hg, but in no particular pattern of relationship with other, settled groupings (for example, the grouping containing Cp). In this, these manuscripts resembled the behaviour of the 'ungrouped manuscripts' of *Svipdagsmál* discussed above: neither they nor Hg fall into any clear grouping simply because they share few unoriginal readings with other manuscripts. This is consistent with all these manuscripts, Hg included, being close to the common ultimate exemplar of the whole tradition. Accordingly, it is a further strong argument for regarding Hg as the manuscript nearest this exemplar. The evidence cited above that El is at least in part based on an exemplar (shared with Gg and Ha4) which contains many readings which are scribal in origin, as evidenced by the metrical impossibility of several of them and their generally flat and unpoetic quality, suggests that El is at least one degree further removed from this exemplar than Hg, and that Hg is, therefore, the best choice as the base for an edition.

6. An editor's use of this: best-text historical editing

I remarked above that the stemma I constructed for the manuscripts of *Svipdagsmál* appeared to suggest some ten lines of descent, and that it did not seem possible to use a stemma with so many lines of descent in any kind of 'majority verdict' stemmatic editing. This is even more true of the Wife of Bath's Prologue. If each of the manuscripts which cannot be grouped forms a separate line of descent and there are other lines of descent among the manuscripts which can be grouped, then there may be as many as twenty different lines of descent. In such circumstances, to choose a reading by counting the lines of descent is absurd.

If we cannot use the stemma to arbitrate between readings in this manner, how then can we use it? Indeed, what point is there in making a stemma at all? Why not just do as Donaldson suggests, and as he and George Kane actually did, and use our taste, knowledge and instinct to make the best text we can on the basis of the evidence of all the manuscripts and whatever else we can muster, and leave it at that? An example may help answer this question. The 'B' text group of manuscripts, at the bottom right of my *Svipdagsmál* stemma, differs so radically from that in the other manuscripts as to amount, at times, to a different poem. Two verses are omitted in *Fjölsvinnsmál*; the overall text differs in some 250 places from that in other manuscripts; around one word in five is omitted, added or altered in the B text relative to other manuscripts. Many of these alterations demonstrate an inclination to meddle in the text, without understanding of meaning, grammar or metre. *Svipdagsmál* was probably composed around 1250. By 1700, its matter, manner and language were strange to its scribes, and the B-text alterations, apparently introduced by a redactor some time between 1680 and 1730 (note on these dates), demonstrate this ignorance amply.

Three readings in the B text, however, make excellent sense.⁹ At all three points, the text in all other manuscripts (including those which otherwise are closest to the original) appears defective and at all three points I accept the B text reading into my edition. In strict stemmatics, I have no warrant for this procedure: the B manuscripts are derived from a manuscript (O; Codex Oblongus, AM 738 octavo, now in the Stofnun Árna Magnússonar, Reykjavík) which stands closer to the original than the B manuscripts. I should include no readings introduced into manuscripts below O; all three of these readings appear first in the B manuscripts and are not

⁹ The three readings are: *Gróugaldur* 8 *fjörðlátum* for *fjörðlotum*; *Gróugaldur* 11/6 *ljái* for *ljá*; *Fjölsvinnsmál* 4/6 *af* for *at*.

in O. I justify the inclusion of these readings as follows: at all three points, the text in all other manuscripts, including manuscripts nearer to the lost hyparchetype, is defective. At all three points, therefore, the text of the hyparchetype was defective. If it were my duty as an editor to do no more than try to recover the hyparchetype (the avowed aim of traditional stemmatics), I should not repair the text at these three points, but leave it as it is, defective though it be. I, however, do not consider my duty to be finished with reconstruction of the hyparchetype: as an editor, I believe I should use all the information at my disposal to reconstruct not the hyparchetype, but, as nearly as I can, the text as it was when it left the author's hand. Indeed, there are 37 places in *Svipdagsmál* where I believe that the hyparchetype is incorrect, and where I emend the text accordingly.

This may seem no different from what Donaldson advocates: I have simply ignored the stemma in choosing these three readings. In the same way, Donaldson himself advocates the reading 'wright' in line 117 of the Wife of Bath's Prologue, although it appears in just two manuscripts which on any stemmatic analysis are far from the original. There is good reason to think that Donaldson is correct here, and that Chaucer indeed wrote 'wright': this reading corresponds to 'conditor' in Chaucer's likely source of this passage. It does not follow, however, that because one ignores the stemma at this point, therefore the stemma has no use at any point. I have introduced these three readings into my text of *Svipdagsmál* because the base text is defective at these three places. Emendation is necessary, and these three readings are introduced as emendations, exactly as if they were suggested by a modern editor, which, in a sense, is exactly what the B text redactor was.

A stemma, then, does have real use, though not quite the use proposed by traditional stemmatics. It gives a picture of the relationships among the manuscripts, and especially of the likely relative proximity of each to the archetype. The editor may use this information for two reasons. First, it may be used to guide the choice of a 'best-text': that is a single manuscript which forms the base of the text of the edition. For *Svipdagsmál*, the stemma suggests that St and Ra are equally close to the archetype. Internal evidence (chiefly orthographic) suggested that St was actually the nearer of the two: the role of the stemma here was to narrow the choice to just these two manuscripts. Similarly, our explorations of the Wife of Bath tradition suggest that for this part of the *Canterbury Tales* Hg is the best choice for a base text. For both *Svipdagsmál* and the Wife of Bath's Prologue, this best text, or base text, is the copy text of the edition, and thus its reading is to be followed at every point where it is not clearly defective.

The second use of the stemma is to advise on the weight to be given to be alternative readings at those points where the chosen 'best-text' is defective. There are less than ten lines in the Wife of Bath's Prologue where we believe Hg is defective: examples are line 117 (as discussed above) and line 484, where the scribe writes 'troce' for 'croce'. Almost all other manuscripts read 'croce': it is an easy slip and easy to repair. In line 117, we choose the reading 'wright', although the stemma tells us that the manuscripts which have this reading are far from the archetype. This reading has precisely the same status as the three B-text readings introduced into *Svipdagsmál*: it is an emendation which happens to occur for the first time in manuscripts.

St, the chosen best-text for *Svipdagsmál*, is much more corrupt than Hg among the Wife of Bath manuscripts, and actually requires emendation no less than sixty-one times in its 400 lines. From the stemma, the first place to look for an improvement must be in Ra, which appears nearly as close to the archetype as St. Ra provides ten corrections to the text of St. Because Ra is, according to the stemma, nearly as close to the archetype as is St, these ten corrections have considerable weight, rather more than might be accorded, for example, the three corrections from the B-text manuscripts. Besides their own excellence, these ten corrections have a historical weight as well. Another six corrections derive from manuscripts in the centre of the stemma, and it is possible that these corrections, too, have historical force, as these manuscripts in the centre of the table appear to derive from a line of descent distinct from St and Ra. Eight other corrections, including the three B-text corrections we have already discussed, come from manuscripts which the stemma suggest are derivative copies. These corrections (like 'wright' in the Wife of Bath's Prologue) are to be regarded as successful scribal emendations. They have no historical force and have only their own excellence to recommend them. Finally, in addition to all these I introduce a total of 37 emendations at places where every manuscript is corrupt, many of these suggested by other editors.

This procedure may be called 'best-text historical editing'. It is best-text editing for two reasons: first, in the Bédierist sense of a best-text edition, it uses the single witness which appears to have the highest historical claim to authority as the copy text. Second, it is best-text editing in that the editor's responsibility is to emend this copy text where it is defective; indeed, to create a best text. It is historical editing for two reasons also. First, it is historical editing because it seeks to identify the copy text on the basis of a historical description of the manuscript tradition. That is: one uses all the techniques described in this paper, all that one knows of the manuscripts and

their copying, to reconstruct the history of the tradition and hence to identify the manuscript closest to the archetype. Second, it is historical editing because where the copy text is defective, the weight of alternative readings is assessed with reference to the historical authority of the manuscripts which contain them. This weight may be considerable, as in the readings introduced from Ra into *Svipdagsmál*, or the manuscripts may have no weight, and the reading must stand or fall by its quality alone, exactly as if it had been proposed by a recent editor.

In this context, it is misleading to call my representation of the *Svipdagsmál* tradition a 'stemma'. This implies that it has a definitive and prescriptive role in the determination of readings. In fact, it has a guiding and advisory role. It may help the editor, but not substitute for the editor. It is best described not as a 'stemma' but as a 'table of relationships'.

7. How 'definitive' is a best-text historical edition?

This model of editing here proposed is far from the scientific, objective model of Lachmann's dream with which I began this paper. It permits, even demands, editorial judgement at every point, as the editor weighs the mass of evidence these methods produce and travels through it towards an understanding of the development of the tradition. The aim is not to obviate editorial judgement but to enable it: to make ourselves better informed about all the processes of transmission and better able to judge the historical force as well as the merit of this reading in this manuscript. Clearly, there is room for different scholars to interpret the same evidence differently. In this world of multiple texts, in multiple uncertain relationships to one another, final decision will always elude us. The concept of a definitive text has no meaning, though there will certainly be better texts and worse texts.

There is a second reason why the definitive text is a chimera. In five years transcription of the *Canterbury Tales* manuscripts, we have become profoundly aware that transcription is a fundamentally interpretative activity. It is not at all a simple recording of what is on the page. It is forever an act of translation: at every point, we have had to decide just what we think the marks on the page mean and how best to represent them, using the means at our disposal. To illustrate this briefly: in the second line of Hg we find three minims, followed by the words 'this world'. If we view these minims in isolation, we would see them as a clear representation of the letter m, but from the context it is clear that these three minims must represent the word 'in' and that is how we transcribe it.

Our transcription is not a mechanical representation of marks on paper. As here, we make a judgement on what we believe the scribe intended, based on our assessment of the scribe's usual practice, of the meaning at that point, of every aspect of the context. Transcription is also fundamentally incomplete. We can represent the spellings of the manuscripts. We can represent some elements of the layout of the manuscripts, for example, their use of ornamental capitals, but there is much we simply cannot record. In the words themselves, our transcription ignores a whole repertoire of scribal signs (different letter shapes in different positions; apparently decorative strokes through, above or after letters which hover on the edge of meaning). We cannot represent all the subtleties of the disposition of the words upon the page: the use of different scripts, of degrees of boldening or ornamentation, or marginal markers, of variously coloured letters.¹⁰ Some of these might be merely decorative, but some appear to have a real semantic function, serving to highlight this word or that, or to mark distinct structural sections.

This raises an immediate difficulty. Every aspect of computer-assisted stemmatic analysis — the collation, and the analysis based on the collation — flows from the transcription. Yet, our transcription is fundamentally interpretative and incomplete. How then can we aspire to create, from this flawed base, a 'definitive text'? Whitaker (1991:128-129) argues that this is a weakness in the technique of computer collation: the imperfection of the transcriptions must carry through into the collation. The argument can be turned on its head. Traditional collation proceeds by abstraction: the scholar, in pursuit of 'real', 'significant' variation commonly ignores all these elements we find so difficult to represent in our transcription. This assumes that it is possible to identify, at every point in every witness, just what is 'significant' variation and what is not. It assumes, too, that this can be done a single witness at a time, as the scholar compares them one by one with the putative base. These are far-reaching assumptions. In contrast, our recognition of the provisional nature of our transcription, and hence of every aspect of our analysis, appears more realistic and may ultimately prove more productive. For example: crucial to our historical analysis of the *Wife of Bath's Prologue* tradition is the belief that just two scribes, hands b and d, between them wrote the four early manuscripts El Hg Ha4 Cp, and crucial to this argument is the evidence of our collation of the spellings of these four manuscripts outlined above. Traditional collation would have levelled

¹⁰ A full description of our system of transcription, and what it records and does not record, is contained in Robinson and Solopova 1993; see also Robinson 1994.

all this spelling evidence away, and hence lost a critical piece of evidence about the relationships of these manuscripts.

By including all this material in our transcripts — on spellings, and increasingly in future, on the orthography and layout of the manuscripts — we greatly complicate our task as editors. But we also greatly enrich the texts we produce. Paradoxically, we also may create texts which can be read with greater confidence than the proclaimed definitive texts of stemmatic editing. Because we do not claim so much for them, because we base them on a fuller survey of the evidence, our text may actually be more convincing than a text which achieves an improverished perfection. The means now at our disposal (and advances in computer imaging and pattern recognition promise yet more) permit us to confront the full range of variation in the traditions we work in. We can use these means, and our own editorial judgement, to let the editions grow from this mass of evidence. Finally, we have the means, too, through electronic publication, to make all this evidence, and all we do with it, available to our readers. From this, they can judge better than ever before what we have done, and redo it for themselves if they wish.

References

- Benson, L.D., ed. 1987. *The Riverside Chaucer*. Boston: Houghton Mifflin.
- Blake, N.F. 1980. *The Canterbury Tales by Geoffrey Chaucer. Edited from the Hengwrt Manuscript*. (= *York Medieval Texts*, 2d ser.) London: Arnold.
- Darwin, C. 1859. *On the Origin of Species*. London: John Murray.
- Dearing, V. 1962. *Methods of Textual Editing*. Los Angeles: Clark Library.
- Dearing, V. 1970. "Computer Aids to Editing the Text of Dryden". *Art and Error: Modern Textual Editing*, ed. by R. Gottesman and S. Bennett, 254-278. London: Methuen.
- Donaldson, E.T. 1970. *Speaking of Chaucer*. London: Athlone.
- Hockey, S. 1980. *A Guide to Computing Applications in the Humanities*. London: Duckworth.
- Kane, G. 1960. *Piers Plowman: The A Version*. London: Athlone.
- Kane, G. 1984. "John M. Manly and Edith Rickert". *Editing Chaucer: The Great Tradition*, ed. by P.G. Ruggiers, 207-229. Norman, Oklahoma: Pilgrim Books.
- Kane, G. and E.T. Donaldson. 1975. *Piers Plowman: The B Version*. London: Athlone.
- Kenney, E.J. 1974. *The Classical Text: Aspects of Editing in the Age of the Printed Book*. Berkeley: University of California Press.
- Lachmann, K. 1876. *Kleinere Schriften*, 2. Berlin: Mouton.
- Maas, P. 1958. *Textual Criticism*. Oxford: Clarendon Press.

- Manly, J.M. and E. Rickert, eds. 1940. *The Text of the Canterbury Tales: Studied on the Basis of All Known Manuscripts*, 8 vols. Chicago: Chicago University Press.
- Mayr, E. and P.D. Ashlock. 1991. *Principles of Systematic Zoology*. Chicago: University of Chicago Press.
- Moorman, C. 1975. *Editing the Middle English Manuscript*. Jackson: University Press of Mississippi.
- O'Hara, R.J. 1988. "Homage to Clio, or, Toward an Historical Philosophy for Evolutionary Biology". *Systematic Zoology* 37:142-155.
- Ott, W. 1992. "Computers and Textual Editing". *Computers and Written Texts*, ed. by C.S. Butler. Oxford: Blackwell.
- Parkes, M.B. and A.I. Doyle. 1978. "The Production of Copies of the Canterbury Tales". *Scribes, Scripts and Readers: Studies in the Communication, Presentation and Dissemination of Medieval Texts*, ed. by M.B. Parkes, 210-248. London and Rio Grande: Hambledon Press.
- Pierce, R.H. 1988. "Multivariate Numerical Techniques Applied to the Study of Manuscript Traditions". *Tekst Kritisk Teori og Praksis*, ed. by B.F. et al. Oslo: Novus Forlag.
- Quentin, H. 1926. *Essais de Critique Textuelle*. Paris.
- Reynolds, L.D. and N.G. Wilson. 1991. *Scribes and scholars: a Guide to the Transmission of Greek and Latin Literature*. Oxford: Clarendon Press.
- Robinson, F.N., ed. 1957. *The Complete Works of Geoffrey Chaucer*. Boston: Houghton Mifflin.
- Robinson, P.M.W. 1989. "The Collation and Textual Criticism of Icelandic Manuscripts (2): Textual Criticism". *Literary and Linguistic Computing* 4:174-181.
- Robinson, P.M.W. 1991. *An Edition of Svipdagsmál*. Faculty of English. Oxford: University of Oxford.
- Robinson, P.M.W. (1993). "An Approach to the Manuscripts of the Wife of Bath's Prologue". *Computer-based Chaucer Studies*, ed. by I. Lancashire (= *CCH Working Papers* 3), 17-47. Toronto: University of Toronto Press.
- Robinson, P.M.W. 1994a. "Collate: A Program for Interactive Collation of Large Textual Traditions". *Research in Humanities Computing* 4, ed. by N. Ide and S. Hockey, 32-45. Oxford: Oxford University Press.
- Robinson, P.M.W. 1994b. *The Transcription of Primary Textual Sources Using SGML*. (= *Office for Humanities Communication Publications* 5). Oxford: Office for Humanities Communication.
- Robinson, P.M.W. and R.J. O'Hara. 1992. "Report on the Textual Criticism Challenge 1991". *Bryn Mawr Classical Review* 3:331-337.
- Robinson, P.M.W. and R.J. O'Hara. 1993. "Computer-Assisted Methods of Stemmatic Analysis". *The Canterbury Tales Project: Occasional Papers* I, ed. by N.F. Blake and P.M.W. Robinson, 53-74. Oxford: Office for Humanities Communication.

- Robinson, P.M.W. and R.J. O'Hara. Forthcoming. "Cladistic Analysis of an Old Norse Manuscript Tradition". *Research in Humanities Computing* 4, ed. by S. Hockey and N. Ide. Oxford: Oxford University Press.
- Robinson, P.M.W. and E. Solopova. 1993. "Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue". *The Canterbury Tales Project Occasional Papers* I, ed. by N.F. Blake and P.M.W. Robinson, 19-52. Oxford: Office for Humanities Communication.
- Ruggiers, P.G. 1979. *The Canterbury Tales, Geoffrey Chaucer. A Facsimile and Transcription of the Hengwrt Manuscript, with variants from the Ellesmere Manuscript.* (= *A Variorum Edition of the Works of Geoffrey Chaucer*, I). Norman, Oklahoma: University of Oklahoma Press.
- Samuels, M.L. 1988. "The Scribe of the Hengwrt and Ellesmere MSS". *The English of Chaucer and his Contemporaries: Essays by M.L. Samuels and J.J. Smith*, ed. by J.J. Smith, 51-69. Aberdeen: Aberdeen University Press.
- Sober, E. 1988. *Reconstructing the Past: Parsimony, Evolution and Inference.* Cambridge, Mass.: MIT Press.
- Solopova, E. Forthcoming. "Chaucer's Metre and Scribal Editing in the Early Manuscripts of the *Canterbury Tales*". *The Canterbury Tales Project: Occasional Papers* 2, ed. by N.F. Blake and P.M.W. Robinson. Oxford: Office for Humanities Communication.
- Swofford, D.L. and J. Olsen. 1990. "Phylogenetic Reconstruction". *Molecular Systematics*, ed. by D.M. Hillis and C. Moritz, 411-501. Sunderland, Mass.: Sinauer Associates.
- Timpanaro, S. 1971. *Die Entstehung der Lachmannschen Methode.* Hamburg: Helmut Buske.
- Timpanaro, S. 1981. *La Genesi del Metodo del Lachmann.* Padua.
- West, M.L. 1973. *Textual Criticism and Editorial Technique Applicable to Greek and Latin Texts.* Stuttgart: Teubner.
- Whittaker, J. 1991. "The Practice of Manuscript Collation". *TEXT* 5:121-130.

Shock Waves in Text Traditions

Cardiograms of the Medieval Literature

E. Wattel

Department of Mathematics and Computer Science
Vrije Universiteit, Amsterdam, NL

and

M.J.P. van Mulken

Department of Language
Katholieke Universiteit Nijmegen, Nijmegen, NL

0. Introduction

In the past decades several methods for developing logically motivated pedigrees of manuscript traditions have been devised. Lachmannian, Quentinian, Cladistic and other statistical approaches have arisen, adding scientific relevance to the philological practice of scholars who wish to edit a sample of a MS tradition with more than one descendant. In contrast to former practices, scholars now prefer to refer to a more or less objective criterion in their selection of the text to be edited. In their search for objectivity they have resorted to mathematicians, logicians and statisticians who were more than happy to adapt their habitual methods to the needs of the philological world.

Because these adaptations have been borrowed from other disciplines, the specific problems inherent in manuscript traditions often remain unacknowledged or are simply discarded because they usually can not be accounted for or, even worse, because they are considered irrelevant. In other words, the intrinsic complexity of MS traditions is passed over and disregarded.

One of the major problems of MS traditions is precisely the often entangled character of the tradition. Practically all - and especially vernacular medieval - MS traditions are known as "contaminated": the transmission cannot be considered straightforward. Instead of being copied from one single older exemplar, manuscript texts have been conflated, i.e., composed from several predecessors.

This "contamination" can be simultaneous (the copyist borrowing from several exemplars at his disposal at the same time) or successive (the copyist using one exemplar for the one part of his transcription and another for the second part and the first or even a third exemplar for yet another part of his copy, etc.); a third possibility is that of incidental contamination (the copyist using one exemplar to complete his transcription and other exemplars to verify or to improve it afterwards by erasing former readings or by interpolating new ones).

If a philologist wishes to use the help of other disciplines to justify his choices in editing texts, he can best turn to a "machinery" that **can** account for these specific problems: the stemmatological apparatus should be devised so that it handles the intrinsic problems of text transmission. In the present paper we will propose a "marriage" of computer science and stemmatology. The marriage is embarked upon in order to solve a typical type of contamination: the successive contamination. Just as in all modern marriages, the contract is stipulated: whenever a design choice has been made - and design choices are inevitable - it has been made meticulously. Though this is an arranged marriage, it is, nevertheless, a happy one.

One of the tasks of modern philology is to unravel the interrelations of the different versions of a manuscript tradition and to construct a graphical representation of those interrelations. The production of a manuscript in medieval times was often done by copying and adapting available material from more than one source. In such a case the manuscript tradition is said to be *contaminated*. One of the simpler types of contamination occurs when one part of the manuscript was copied from one source and another part from another source.

If such a split in sources has occurred several times, thus, if the manuscript tradition cannot be considered straightforward, then it may be advisable to split the manuscript into *sections* and to construct a tree representation for each section. Once these trees are known, then from the comparison of the trees it should be possible to attain insight into the reproduction anomalies of the manuscript tradition considered. In this paper we wish to supply a technique which is able to separate the sections when this type of irregularities clearly occurred during the copying process.

The usual way to express differences in manuscript versions is to use a list of variants based on differences and similarities between the text versions, commonly known as difference formulas. We will use this type of formulas to express the only information which will be available for the constructions and algorithms. The contents and the weight choices of the formulas will be the responsibility of the philologists.

The steps to follow are: one has to verify whether a manuscript tradition is contaminated or not. A manuscript tradition is contaminated if the amount of contradicting difference formulas ($ab/cd \leftrightarrow ac/bd$) surpasses a previously fixed threshold determined by the philologist and by the characteristics of the tradition itself. If a manuscript tradition is not contaminated, one can proceed in the usual way by constructing a pedigree (see our other joint article and the article by Wattel in this volume). If the tradition is contaminated, one should verify whether contamination is produced by a split of sources. If so, one should proceed by the separations of the sections. Since the section separation process is completely isolated from the tree construction, we will concentrate on this matter. The general underlying idea in this respect is:

Suppose different sources (x and y) have been used to produce a new copy A . The first part of A has used x , the second part y ; suppose that the corresponding tree of the manuscript tradition of A is as displayed in *Fig. 1*.

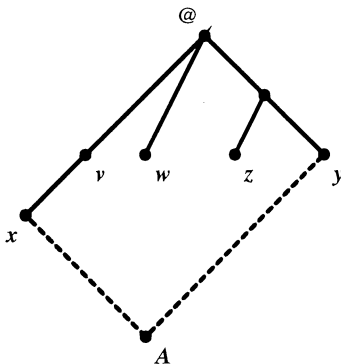


Fig.1 Copying from two sources.

Then in the first part of the tradition the distance between A and v will be relatively smaller than in the second part of the manuscript tradition, since the copyists of A switched sources during the transcription. In order to find out if such a sudden switch in distances between manuscripts occurs, a distance distribution function, which enables one to measure the differences in distances, is designed. We proceed as follows:

Construct a distance distribution function on every pair of manuscripts in the manuscript tradition and every position of the text. If during the copying process one source has been changed for another, then there is a dramatic change in the distance distribution occurring at the boundary of the sections coinciding with the switch. If those shocks in the interrelations of the manuscripts are made visible in cardiograms (shock waves or encephalograms), then a justified choice could be made as to how to divide up the manuscripts.

We will define a distance distribution function which will generate a similarity score for pairs of manuscripts at variant places in the manuscript tradition. Thereafter, we will discuss the possible functions which measure the scope of the similarity scores found in the variant places. Once we have decided on which levelling function corresponds best to the needs of a specific manuscript tradition, we can check whether sudden changes in distances between manuscripts coincide with a former switch of sources of copyists.

Before the details of the algorithms are discussed, it is necessary to supply the data representations and the distance distribution functions which are used by the algorithms.

N.B. In this paper we do not supply *proofs*, but develop a tool which can be helpful in studying a manuscript tradition. We also have had to make many design choices: thus in fact we have a paper which shows how things *could* be done, rather than how things *should* be done.

1. Material

In order to study the various versions of a manuscript tradition by means of automatic procedures, it is necessary to make a clear decision about the way in which the material is presented, the so-called *data structure*. This *data structure* supplies a code reflecting an abstract model of the tradition and of the manuscripts themselves; if well chosen, it admits automatic treatment without prejudice. Should this treatment yield unrealistic answers, even though it is internally consistent and correct, then the problems are to be sought in the assumptions and choices made by the linguist in preparing the material. Thus a flexible data structure admits a test for design choices, conjectures and properties of the tradition. This structure should also be clear enough to admit algorithmic treatment with a low complexity in order to make inferences by means of a computer feasible.

The data structure used here is to represent the manuscript tradition as a *heading line* and *difference formulas* (see the example below). To facilitate search procedures, line numbers are given in correspondence to a reference standard - one of the versions of the manuscript tradition. The *heading line* contains the number of lines of the total as its first field and a coding letter (preferably the usual siglum) for each manuscript in the tradition as its second field. The general structure of a *difference formula* is a line containing the line number of a version difference as its first field, the coding letters for the involved manuscripts as its second field and a weight factor for the difference as its third field. The coding letters of different versions are separated by the sign "/", while manuscripts which have an equal version are not separated by the "/" in the coding.

Line numbers are not necessarily different and are not necessarily integers, thus admitting more than one difference in one line and separating observations at the beginning and at the end of a line. (It is common practice to assume that difference formulas are supplied in non-decreasing order with respect to line numbers.)

To give an example of a very small manuscript tradition:

5	<i>abcdefg</i>		(headering line)
1.1	<i>acd/be</i>	7.5	(first formula)
1.6	<i>abdg/ce</i>	0.5	(second formula)
2	<i>ad/bc/fg</i>	2	(third formula)
4	<i>ab/cef/g</i>	5	(fourth formula)

In this example the story has 5 lines, with an important variant and an unimportant variant in line 1, a difference in line 2 and one in line 4. Moreover, the differences in line 1 are not applicable to *f*, perhaps because *f* is lacunary, or for any other reason which makes the formula unapplicable. The same is true for *g* in 1.1, for *e* in 2 and for *d* in 4. N.B. These formulas can nevertheless be used in our procedure: formulas need not be complete! For instance in line 4, *a* and *b* share a reading, *c*, *e* and *f* share a different reading and manuscript *g* has yet another different reading, which is in this case a single version. This means we do **not** restrict ourselves to "type 2 differences" in which there are at least two corresponding manuscripts in every part of the formula, and the formula has precisely two parts cf., e.g., Salemans (1990).

In the construction of pedigrees and trees we will only use this type of difference formulas. If some information is highly important and occurs throughout the entire story (for example, the occurrence of a different name for the protagonist), a priori it does not fit in such a type of difference formulas. In order to take it into account we could, e.g., introduce it as an extra formula, which is inserted in the database for every line number divisible by, e.g., 25. If the original version of some formulas is known, this could be coded by using a fictive manuscript @ inserted into the favoured section of those formulas.

2. A Zero-Centred Distance Distribution Function

One of the most easy and useful ways to study a manuscript tradition is to make a distance function which indicates which manuscripts are proximate and which are distant. The precise definition of such a distance, however, is not a priori obvious, and poses some problems if many formulas and manuscripts are lacunary. In this section we develop a *similarity score* which is intended as a type of distance function, even though it is not necessarily positive, and which handles lacunary information unproblematically.

Since a distance function is defined on pairs of points, we develop in this paragraph a function on *pairs* of manuscripts which should replace the idea of a distance. Moreover, this can be constructed in such a way that it is concentrated on a couple of lines and their neighbourhood. The necessary variation to use it is supplied as a function in which the line number is the independent variable.

Consider, e.g., the formula for line 2 of the example. There are three different readings, each with two manuscripts. With the six manuscripts in the formula we have 15 pairs, 12 in which the two members of the pair have a different reading and 3 in which the reading is equal. The weight of the formula is 2.

In this formula the event that two readings are different is common and the event that two readings are equal is rare; this means that the *information content* of the message when two readings coincide is larger than the in case that two readings differ. In fact, one could well maintain that the *information content factor* for equality is 4 times the *information content factor* for inequality, since 12 divided by 3 makes 4. Next we require that the value of the *information content* over all pairs should reflect the weight 2 of the formula, that equality should give a positive contribution and that inequality should give a negative contribution to the *similarity score*. By means of the formulas given below, we obtain the result that the pairs with a similar reading get a contribution of +10 and the pairs with a different reading get a contribution of -2.5.

An important observation is that the sum of the contributions over the fifteen pairs equals 0. The fact that manuscript *e* does not occur in the formula means that no contributions are made to pairs of manuscripts with *e*, either in positive or in negative sense. So if the contributions for any pair of manuscripts over many formulas are taken, a resulting positive score means that these manuscripts belong together in a substantial part of the cases that matter; a negative score means that the two manuscripts differ in a majority of the formulas in which both occur. *There is no bias built in for lacunary manuscripts.*

In order to give an explicit definition: Let x and y be two manuscripts and let Φ be the collection of all difference formulas. For every $\varphi \in \Phi$ we compute the number E_φ for similar pairs and the number D_φ for different pairs in the formula. Let W_φ be the *weight* of φ . We obtain:

$$\sigma_\varphi(x, y) = W_\varphi \cdot \frac{D_\varphi + E_\varphi}{E_\varphi} \text{ for } x \text{ and } y \text{ similar in } \varphi,$$

$$\text{resp.} \quad \sigma_\varphi(x, y) = -W_\varphi \cdot \frac{D_\varphi + E_\varphi}{D_\varphi} \text{ for } x \text{ and } y \text{ different in } \varphi,$$

$$\text{resp.} \quad 0 \text{ if } x \text{ or } y \text{ is missing in } \varphi.$$

At the moment the contribution $\sigma_{\varphi}(x, y)$ is localised in the formula φ , and hence on the line number of φ . So for each variant the similarity score for the pairs of manuscripts can be computed. For a variant to obtain genealogical value, it is assumed that the variant is a witness of a difference which extends over a larger domain and is not restricted to the position of the variant only. We must be able to spread the influence of the variant over the neighbourhood of its line number. This part of the construction, the *levelling process*, is the subject of the next section.

3. Spreading and Levelling the Similarity Score

The occurrence of different readings at a certain text position is not in itself a set of relations over the manuscripts, but should be considered as a witness of a grouping of manuscripts. The scope of this grouping is usually much larger than the single line with the different readings. It is well possible that part of the grouping goes unattended in some part of the text, because there was no information about version differences. So if a formula occurs, it occurs in the domain of some grouping. Our task is to estimate this domain as accurately as possible from the information which the formulas supply, in order to be able to use this information in a shock wave setting.

The algorithmic *levelling* technique used for this purpose is sometimes called *splashing*. We refer to the type of artist who makes a painting by throwing blotches of paint at his canvas from a considerable distance. The position he is aiming at with each blotch is precise (in our case the line number); the amount of paint in each blotch is variable but well defined (in our case the weight of a formula). So the influence of aiming points spreads with the paint and in the end the canvas will show an accurate image of the actions of the painter.

In the above example, in the way the paint will spread is of importance. In our case as well, we have to construct a sensible spreading distribution. There are several possible design choices to be made here which we will treat in succession:

a) *The design choice involving the independent variable.*

We need an independent reference variable, a touchstone. In essence a measuring function μ must be defined from which distances can be derived and which runs sequentially through the text. We define a **grid** which is based on a row of incrementing values at fixed intervals. Fast changes in relationship should yield large differences of the spreading intensity at two adjacent points of the grid derived from the independent variable μ .

- a1) The most obvious option is to use the line number *in the text* as the independent variable μ . Implementation of this option is not difficult, since the line numbers are supplied at the first field of every formula. Moreover, it is the obvious independent variable to be used in the graphical representation of the changes in the distance distributions.
- a2) Another option is to use the line number *in the database* of the formula as the independent variable μ . Again this is easy to implement in the algorithms, but it needs another conversion step afterwards in the graphical representation, since there the line number is needed anyway. The option has the advantage that if a significant change in interrelation occurs between two adjacent formulas, with a lot of agreeing lines between them, then it is invisible in method a1) and visible within this design option.
- a3) The next option is to use the *cumulative formula weight*. Note that the attention should go to those parts of the text where the important differences are registered (cf. the amount of paint in the blotch is taken into account). We use a variable which concentrates on such parts. The collection Φ is ordered in a natural way because the database itself is an ordered collection of lines, thus making sense to say for two formulas that $\psi < \phi$. We define $\Omega = \sum_{\phi \in \Phi} W_{\phi}$, which is the sum of all formula weights. Next we can define a new function μ_{ϕ} , which tells what part of Ω is already encountered before formula ϕ , and obtain:

$$\mu_{\phi} = \Omega^{-1} \cdot \left(\frac{1}{2} \cdot W_{\phi} + \sum_{\psi < \phi} W_{\psi} \right).$$

This variable runs from 0 to 1, and for every real number in this range the difference formula in which this number is passed can be found immediately from the database. This option has the advantage over a2) that even if a formula is split up in many parts, i.e., if more than one formula is used to express small differences in influence between the readings in one line, the effect for the similarity score still does not change. A disadvantage could be that a single high weight difference formula dominates its neighbourhood, effectively screening its predecessors from its successors. Again it is necessary to recompute the line numbers from the μ_{ϕ} in order to make the representation of the shock waves visible.

- a4) Since each of the three alternatives has its advantages and its disadvantages and they are not really mutually exclusive, it can be sensible to define a combined measure of the three.

In that case a weighted average should be constructed over the three μ functions; although the weights can be chosen freely, we would suggest to give alternative $a3$) a larger and $a2$) a smaller influence. We could posit, e.g.: $a1) \leftrightarrow 30\%$, $a2) \leftrightarrow 15\%$, $a3) \leftrightarrow 55\%$. If we choose N as the number of lines, M as the number of difference formulas and μ_1 , resp. μ_2 , resp. μ_3 , for the variable of $a1$), resp. $a2$), resp. $a3$), then we obtain:

$$\mu = 0.3 \cdot \frac{\mu_1}{N} + 0.15 \cdot \frac{\mu_2}{M} + 0.55 \cdot \mu_3.$$

b) *The design choice concerning the scope of the spread*

This choice depends on what could be expected in the manuscript tradition. A visible stable section in a figure should cover at least about a twentieth of the text. In order to obtain a stable region inside these minimally sized sections, the influence of a formula should be bounded by a fiftieth part of the text, resp. formulas, resp. weight, i.e., whichever choice is made at a). In order to express this influence, we will define a parameter λ in each of the cases of the next design choice.

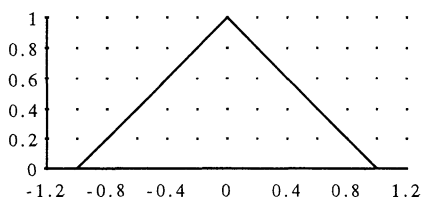


Fig. 2.a Linear Distribution

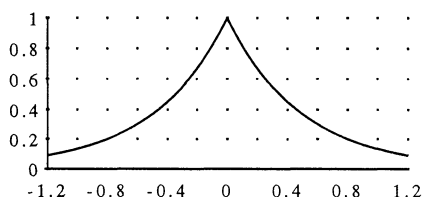


Fig. 2.b Negative exponential

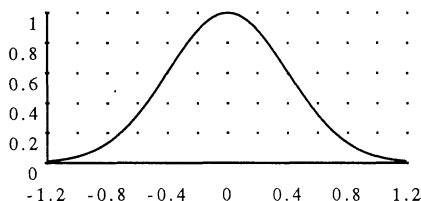


Fig. 2.c Gaussian distribution

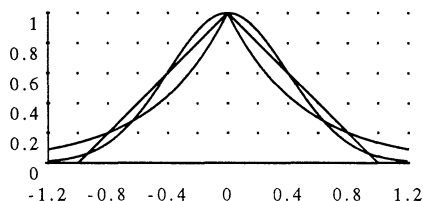


Fig. 2.d Three distributions

c) *The design choice of the function for the spread of the influence*

- c1) The linear spread distribution (see in Fig. 2.a.). The influence decreases from its maximum in the position of the formula as a straight line segment to 0, at a certain defined distance. This distribution is easy to compute, but has the disadvantage that if some position is moderately far from all applicable formulas, it does not allow drawing a conclusion about the grouping.

The formula for the influence at distance δ and width parameter λ in a score σ is :

$$\rho(\phi, \delta) = (1 - \delta \cdot \lambda^{-1}) \cdot \sigma_{\phi} \text{ if } \lambda \cdot \delta < 1, \text{ and } 0 \text{ otherwise.}$$

The influence of a formula is now clearly restricted to a distance λ .

- c2) The negative exponential spread distribution (see in *Fig. 2.b.*). The influence of a formula decreases exponentially with the distance. If a position is far from the formulas, a conclusion can still be drawn from the closest formulas available. The formula for the influence at distance δ and width parameter λ is :

$$\rho(\phi, \delta) = e^{-2 \cdot \delta \cdot \lambda^{-1}} \cdot \sigma_{\phi}.$$

In this formula we use a power of the number e , the base of the natural logarithms. The influence of a formula is now clearly restricted to about $2.3 \cdot \lambda$, because for larger distances the influence is less than 1 %.

- c3) The *Gaussian* spread distribution (see in *Fig. 2.c.*). The influence of a formula decreases with the distance, and even faster than exponential in the distant regions, but it has a level part close to the formula itself. Though this is probably the better formula to describe the real influence, it has a high computational complexity; since the advantages over b2) are only minor we have not used it thus far. The formula for the influence at distance δ and width parameter λ is :

$$\rho(\phi, \delta) = e^{-\pi \cdot (\delta \cdot \lambda^{-1})^2} \cdot \sigma_{\phi}.$$

As usual, π is the proportion between the circumference and the diameter of a circle, and the number e is again the base of the natural logarithms. The influence of a formula is now clearly restricted to about $1.2 \cdot \lambda$, because for larger distances the influence is less than 1 %.

In *Fig. 2.d* the three different distribution functions with the same weight and influence are shown in a single graph. Whichever choice is made, for every position ξ in the text, and especially for each position γ of the *grid*, we can define the total influence of all of the formulas on that particular position. We posit:

$$\rho(\xi) = \sum_{\phi \in \Phi} \rho(\phi, |\mu(\xi) - \mu(\phi)|).$$

We now have determined the influence of the formulas on a continuous function over the entire text. We obtain the most significant case when we put the similarity score $\sigma_{\phi}(x, y)$ in the position of the score function. In that case we have a way to display the probable similarity of two manuscripts as a continuous function over the entire text.

4. Score Comparison and Shock Waves

From the techniques derived in the previous section, we now have the similarity score for every two manuscripts x, y on every position of the text. In order to indicate the manuscripts in the scores, we will denote them in the subscripts of the functions, e.g., $\rho_{x,y}(\xi)$. In order to compare those scores in a reasonable way, we compute how large the score could have been if all influence were positive. That means that we could take the *possible weight*:

$$\omega_{x,y}(\xi) = \sum_{\phi \in \Phi} |\rho_{x,y}(\phi) \cdot (\mu(\xi) - \mu(\phi))|.$$

as the factor to which the score $\rho_{x,y}(\xi)$ should be compared. If we posit $\eta_{x,y}(\xi) = \frac{\rho_{x,y}(\xi)}{\omega_{x,y}(\xi)}$, then the range of the function η is between 1 and -1 .

It is completely clear that this way $\eta_{x,y}(\xi)$ could very well be displayed in a graph of limited range. Moreover, the minimal, the maximal and the mean value of $\eta_{x,y}$ could be used as an indication of similarity which could be compared to other pairs of manuscripts in the tradition.

Even more important than the question of how the interrelation of a pair of manuscripts behaves throughout the text, is the question concerning some particular manuscript, say z , which changes from one group to another group within the manuscript tradition. In this case there is a type of contamination which could have occurred when, during the history of the manuscript, one of the scribes changed his exemplar in the copying process. This means that the scores $\eta_{x,z}(\xi)$ change violently for several manuscripts x which are compared to z in that text position. Since change rates should be defined as differences, we should try to estimate score differences of the functions $\eta_{x,z}(\xi)$ by means of an effective algorithm. At this point it is sensible to use the *grid* defined in the previous section; the difference between two grid points is used as indication of the size of the change.

Next we observe that small changes between the η occur continually, and we are not particularly interested in them. Moreover, the changes should always count positively and equality should count as 0 . We define:

Let $\gamma_1, \gamma_2, \gamma_3$ be three consecutive grid positions, then we define the *change score* τ_z at position γ_2 by:

$$\tau_z(\gamma_2) = \sum_x \omega_{x,z}(\gamma_2) \cdot (\eta_{x,z}(\gamma_3) - \eta_{x,z}(\gamma_1))^2.$$

This definition is motivated by the following observations: It is sensible to use an interval that is not too small to define the differences and to use squares in order to enhance large differences over small differences; an additional advantage is that the results are always positive.

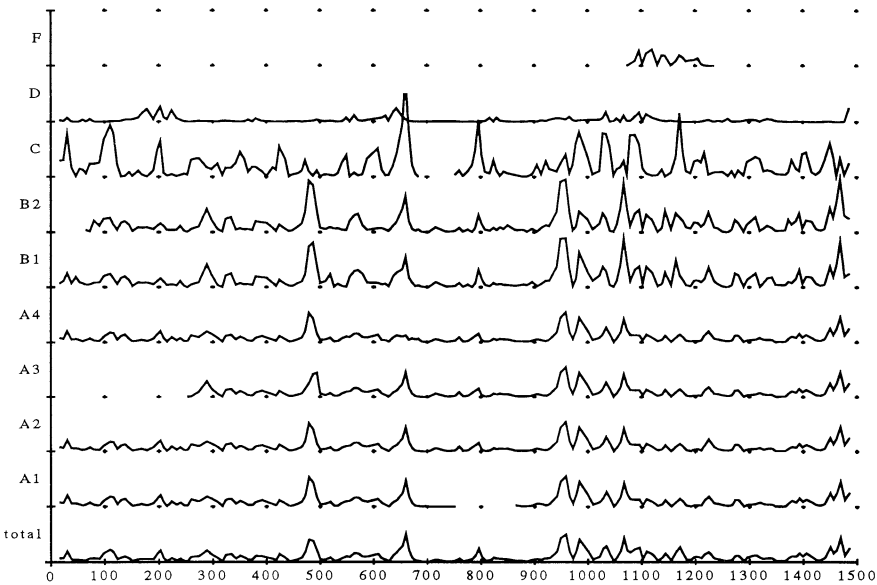


Fig. 3. Charroi shock waves over all manuscripts

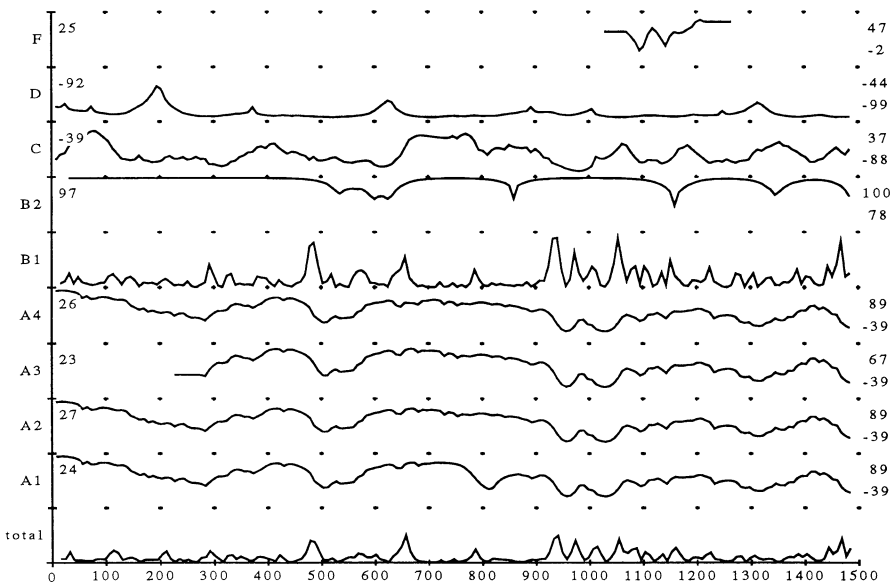


Fig. 4. Charroi comparison with manuscript B1

Finally, it is more important to have changes between manuscripts at positions where the possible weights $\omega_{x,z}(\gamma)$ are large, than with manuscripts and at positions where $\omega_{x,z}(\gamma)$ are small. We only wish to work on grid positions from this point on.

Before $\tau_z(\gamma)$ is displayed in a graphical plot, it is advisable to divide it by a factor in order to keep its range bounded. We could choose, e.g., $\max_{\gamma} \sum_x \omega_{x,z}(\gamma)$. Now it is certain that the maximum value of the displayed change score is less than **4**, and it is safe to assume that it is bounded by the range **0–1**, because differences in η over the value **1** with high weight are extraordinary.

So we use $\omega_{x,z}(\gamma_2)$ as the weight estimation, and the squares of the differences of the η in the predecessor and the successor as the change score in the point γ_2 itself.

Finally, it is useful to display the mean value τ of the τ_z functions as a separate graph. This graph displays the positions in the text where large contaminations or even changes of the pedigree can be expected. We have made several combination figures of the mean value τ function and all the functions τ_z in one plot. Such a plot is often called a *cardiogram* (encephalogram) of the manuscript tradition, while the individual curves are usually called *shock waves* (cf. Fig. 3).

5. Some Final Remarks

In order to make the cardiograms and score figures themselves, some extra choices have to be made on how to display the material. If a graph displays some interesting behaviour, it is necessary to spot the position in the text where the behaviour occurs; thus in all graphs we have chosen for the line number as the independent variable.

In order to attain a reasonable plot the number of grid points is usually taken to be **250**, which means that a stable region in the manuscript tradition should be at least about **10** grid points, and it covers over half a centimetre if displayed in a graph.

In order to obtain a reasonable difference in three adjacent grid points, it should be clear that most of the influence of a formula at γ_1 should have vanished at γ_3 and vice versa. That means that λ should be chosen at about **1** in case of the negative exponential weight function of section 4, c2), and at about **0.5** in the case of the linear weight function c1) and the Gaussian weight function c3).

We usually put several curves in one plot. The plot is divided into strips, a bottom strip for the mean behaviour of the manuscript tradition and a separate strip for each of the manuscripts. Two types of figures are usually made: the first type is the total shock wave $\tau(\gamma)$ on the bottom strip and the $\tau_z(\gamma)$ for each manuscript of the other strips, (cf. *Fig. 3* and *Fig. 5*); the second type is the difference curves for a single manuscript z , with respect to all others. Since the bottom strip and the z -strip are not defined in this way, we could use them for $\tau(\gamma)$, resp. $\tau_z(\gamma)$ (cf. *Fig. 4*). In this second type of plot we also display the mean value of the similarity score between the displayed pairs of manuscripts at the left side of each strip. On the right side of the strip we display the attained maximum and minimum values. In order to make the behaviour of the similarity score visible, the range of the similarity score is enhanced slightly.

With this type of plots we can also see which manuscripts are proximate, because related manuscripts display a similar behaviour. The technique can efficiently handle **31** manuscripts. If more manuscripts are considered in the manuscript tradition, Wattel supplies methods to restrict the computations to a part of the tradition (see this volume).

6. Application of the Algorithms on Manuscript Traditions - Results

After having defined the distance and levelling functions, we were anxious to see whether they allowed us to detect changes in sources in manuscript traditions. It is important to know whether the sections that may have been isolated with the help of these functions correspond to sections of historical entities. We have tested the techniques described here on several manuscript traditions. *Fig. 3* and *Fig. 4* deal with two plots of the manuscript tradition of *Charroi de Nîmes* (Schøsler 1993). A close consideration of *Fig. 3* shows that there are instabilities around line **475**, around line **650** and between **950** and **1050**. The first two instabilities seem to be local, if *Fig. 4* is taken into account, but we concluded that the versions **A1–A4** are closer to **B1–B2** before line **1000** than afterwards. This is a highly remarkable result: the change in the neighbourhood of line **1000** coincides with the length of a quire of 8 pages and 2 columns of 30 lines per page. The change in relationships implies that at a former stage in the genealogy of the *Charroi the Nîmes* quires have been switched.

The switch implies that the manuscript tradition should be separated into two sections and that at least two different pedigrees should be drawn to account for the relationships between the manuscripts, one representing the relationships in the first part (vss. 1-1000) and the second representing those in the second part. These results will be further elaborated in our other joint article in this volume.

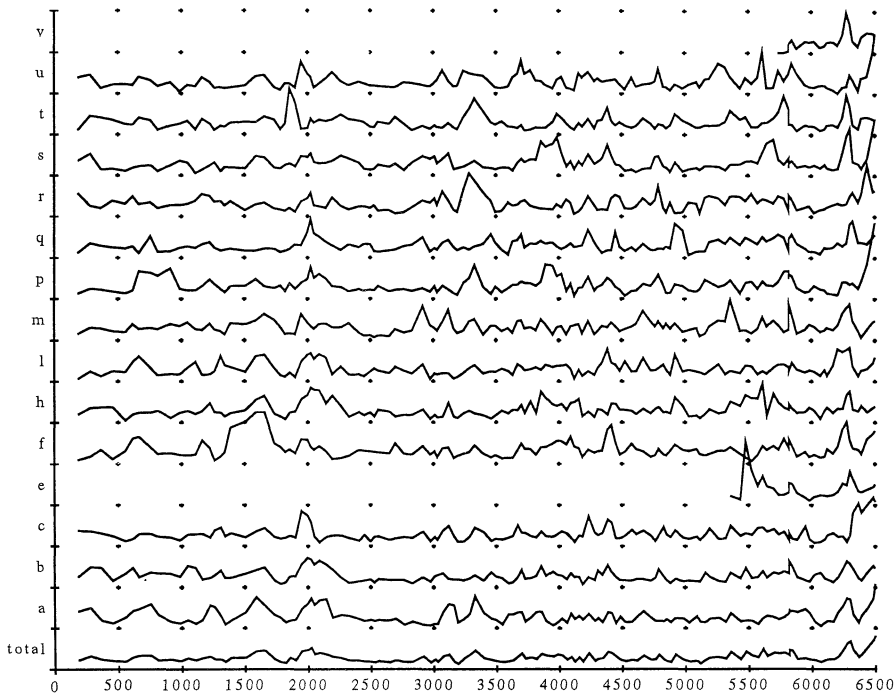


Fig. 5. *Perceval shock waves over all manuscripts*

The study of *Perceval* of *Chrétien de Troyes* (cf. van Mulken 1993) was less successful. In Fig. 5 the shock wave plot of the first 6500 lines is displayed. It does not appear reasonable to separate the manuscript tradition into sections on the basis of this plot, nor is there a clear indication which manuscripts go together. The shock waves seem merely to display some type of "white noise"; according to van Mulken (1993) the contamination in this manuscript tradition is so heavy that no clear pedigree emerged, no matter what kinds of techniques were used.

The most pure type of "white noise" was obtained in studying the structure of a simulated contaminated manuscript tradition, supplied by Onno Huber. This is to be expected since the simulation methods he used in constructing his tradition were stochastically independent of the text position.

Since the main goal of this paper is to develop a method for making a part of the internal structure of the relationship in a manuscript tradition visible, we did not discuss the pedigree itself, the orientation of the manuscripts with respect to the archetype, etc. We have not dealt with the tree construction algorithms either, which should be used after the separation into sections is finished. Those algorithms and their complexity are the main subject in our other joint article in this volume.

This paper is a part of a big machinery dealing with the coding and processing stemmatological information. In Salemans (1990) some remarks are made with respect to the classical results of Lachmann. The research in this area recieved an extra stimulus from the ideas of Dees (cf. references Dees, Dekker & Hoge hout Mulder 1988-1989; Dees 1992). Several members of his research group obtained satisfactory results following his lines of thought (cf. references Dekker 1985; van Mulken 1992); additional techniques used in the construction of the pedigree are found in van Mulken (1993) and in this volume in the articles by Wattel and by Wattel and van Mulken. The ideas for this particular paper evolved from discussions with Marcel Dekker and several others of the stemmatology colloquia.

We tested our methods on a simulated manuscript tradition designed by Onno Huber. In this test we obtained the most pure type of "white noise". From the obtained distance distributions we could reasonably reconstruct the original pedigree of Huber's tradition. Contrary to our experiences with real traditions, there were no indications of switches in relationship. This could be expected since the simulation methods he used in constructing his tradition were stochastically independent of the text position.

References

- Dees, A., M. Dekker and M. Hoge hout Mulder. 1988-1989. "Een voorbeeld van stamboomreconstructie: Karel en de Elegast". *Spektator* 18-2. 96-118.
- Dees, A. 1992. "Les chartes dans la recherche linguistique et philologique". *Le médiéviste et l'ordinateur* 25. 23-27.
- Dekker, M. 1985. "Stambomen". *Corpusgebaseerde Woordanalyse*. 34-39.
- Mulken, M. van. 1992. "Gerd Simon: Ein neues Kraut?". *Corpusgebaseerde Woordanalyse*. 113-127.

- Mulken, M. van. 1993. *The manuscript tradition of Chrétien de Troyes, a Stemmatological and Dialectological Approach*, Dissertation, Faculty of Language. Amsterdam: Vrije Universiteit.
- Salemans, B.J.P. 1990. "Text genealogical remarks on Lachmann, Bédier, Greg and Dearing". *Leuense Bijdragen* 79. 27-468.
- Schøsler, L. 1993. "New Methods in Textual Criticism: The Case of *Charroi de Nîmes*". *Medieval Dialectology*, ed. by J. Fisiak, 225-276. Berlin: Mouton de Gruyters.

Clustering in Stemmatological Trees

How to Handle a Large Number of Versions

E. Wattel

*Department of Mathematics and Computer Science
Vrije Universiteit, Amsterdam, NL*

0. Introduction

It is difficult to describe the interrelation of several versions of a text tradition. The production of a manuscript in medieval times was often done by copying and adapting available material from more than one source. In such a case the text tradition is said to be *contaminated*. One of the tasks of modern philology is to unravel the interrelations of the different versions and manuscripts, to construct a graph representation of those interrelations and, if possible, get a sufficient amount of information to formulate sensible ideas about the most original version of the text, the *archetype*.

The main purpose of this type of research is to construct a *pedigree* of a text tradition in which it can be seen which manuscript is copied from which. We use the terminology of *graph theory*, in which manuscripts are represented as *nodes*, i.e., dots in a figure, and relations are represented as *edges*, i.e., lines in the figure which connect the related *nodes*. Usually the information is too contaminated to find a pedigree directly. Therefore, we first look for a cycle-free unordered graph structure, which we call a tree. If in such a tree a most original version can be spotted, this structure gets a natural ordering and we have a pedigree.

The usual way to express differences in versions is to use difference formulas, i.e., formulas of the type $abc/de/f$, which means that the manuscripts a , b and c share a common reading different from the reading in the manuscripts d and e , and again different from the reading in f . Together with the text positions where differences are located and an indication of the philological importance of the differences, this is the only available information which will be used in the construction of the pedigree.

The importance of a certain formula will be expressed by some real number, called *weight* of the formula. The contents and the weight choices of the formulas will be the responsibility of the linguists.

The aim of the constructions is to generate the best possible tree, in spite of heavy contamination of the text tradition. In fact, we are looking for a kind of tree which is a common denominator of the available information, which contains much information and few contradictions. We are thoroughly aware that this does not supply the final answers in the interrelations of the material, but it will be a starting point from which adaptations could be made if required. Moreover, from a successful tree a possible pedigree could be constructed if there are enough clues which point to the original readings, etc.

The construction of the best possible tree from the material is often a complex task which should be subdivided into stages. Moreover, often in the collection of manuscripts there are clusters, and the construction of (i) a tree for every cluster and (ii) a tree for the collection of clusters could reduce the complexity of the overall problem. The aim of this paper is to describe some algorithms for isolating clusters and to find a suspension point where a cluster could be connected to another part of the tree. Before the details of the algorithms are discussed, it is necessary to supply the data representations and the general setting in which those algorithms have their place.

1. Material

In order to study the various versions of a text tradition by means of automatic procedures, clear decisions must be made as to how the material is presented, the so-called *data structure*. This *data structure* supplies a code which reflects an abstract model of the tradition and of the manuscripts themselves, and, if it is well chosen, it admits automatic treatment without prejudice. Should this treatment yield unrealistic answers, in spite of its being internally consistent and correct, then the problems are to be sought in the assumptions and choices made by the linguist in preparing the material. Thus a flexible data structure admits a test for design choices, conjectures and properties of the tradition. This structure should also be clear enough to admit algorithmic treatment with a low complexity, making inferences by means of a computer feasible.

The data structure used here is to represent the text tradition as a *heading line* and *version formulas* (see the example below). The *heading line* contains the number of verses of the total manuscript as its first field and a coding letter for each manuscript in the tradition as its second field.

The general structure of a *version formula* is a line containing the line number of a version difference as its first field, the coding letters for the involved manuscripts as its second field and a weight factor for the difference as its third field. The coding letters of different versions are separated by the sign "/", while manuscripts which have an equal version are not thus separated in the coding. Line numbers are not necessarily different and are not necessarily integers. This admits more than one difference in one verse and allows for separating observations at the start and at the end of a line. Usually it is assumed that version formulas are supplied in non-decreasing order in relation to verse numbers. To give an example of a very small text tradition:

5	abcdefg		(header)
1.1	acd/be	7.5	(first formula)
1.6	abdg/ce	0.5	(second formula)
2	ad/bc/fg	2	(third formula)
4	ab/cef/g	5	(fourth formula)

In this example the story has 5 lines, with an important and an unimportant variant in line 1, and a difference in line 2 and one in line 4. Moreover, the differences in line 1 are not applicable to *f*, perhaps because *f* is lacunary, or perhaps for some other reason which makes the formula inapplicable. The same is true for *g* in 1.1, for *e* in 2 and for *d* in 4. For instance, in line 4 *a* and *b* share a reading, *c*, *e* and *f* share a different reading, and manuscript *g* has yet another different reading, which is in this case a single version. (This means we do *not* restrict ourselves to "type 2 differences" like some other authors do, a method which requires that there are at least two corresponding manuscripts in every part of the formula, and that the formula has precisely two parts.)

In the construction of pedigrees and trees we will only use this type of version formulas. If some information is very important and occurs throughout the entire story, it does not fit a priori in such a type of version formulas. In order to take it into account we could, e.g., introduce it as a extra formula inserted into the database for every verse number divisible by, e.g., 25. If the original version of some formulas is known, this could be coded by using a fictive manuscript @ inserted among the preferred formulas. Sometimes it is necessary to postpone the decisions on relative weight while preparing the version formulas. In that case one could make another type of code in the weight field, which could be replaced by a real weight just before the algorithms start. This is a sensible way to test the influence of certain weight decisions.

2. Zero-Centred Evaluation Functions

One of the most useful ways to study a text tradition is to make a distance function, which shows which manuscripts are closely related and which are remotely related. The precise definition of such a distance is not a priori obvious, and poses some problems if many formulas and manuscripts are lacunary. In this section we develop a *similarity score* which is meant as a type of distance function, but which is not necessarily positive, and which handles lacunary information without problems.

2.1. *Similarity score*

Since a distance function is defined between pairs of points, we develop in this paragraph a function on *pairs* of manuscripts, which should replace the idea of distance.

Consider, e.g., the formula for line 2 of the example, where there are three different readings each with two manuscripts. With the six manuscripts in the formula we have fifteen pairs, twelve in which the two members of the pair have a different reading, and three in which the reading is equal. The weight of the formula is 2.

In this formula the possibility that two readings are different occurs frequently, and the possibility that two readings are equal occurs rarely. In such a case the *information content* of the message when two readings coincide is larger than when two readings differ. In fact, one could very well contend that the *information content factor* for equality is 4 times the *information content factor* for inequality, because 12 divided by 3 makes 4. In the definition of the *similarity score*, we require that the value of the *information content* over all pairs should reflect the weight of the formula, that equality should give a positive contribution, and that inequality should give a negative contribution. With the defining formulas given below, we obtain that the pairs with a similar reading get a contribution of +10, and pairs with a different reading get a contribution of -2.5.

An important observation is that the sum of the contributions over the fifteen pairs equals 0. The fact that manuscript *e* does not occur in the formula means that no contributions are made to pairs of manuscripts with *e*, either in a positive or in a negative sense. So if the contributions for any pair of manuscripts over many formulas are taken, a positive score indicates that these manuscripts go together in a substantial part of the cases that matter, and a negative score indicates that the two manuscripts differ in the majority of the formulas in which both occur. *There is no bias built in for lacunary manuscripts.*

In order to give an explicit definition: Let x and y be two manuscripts, and let Φ be the collection of all version formulas. For every $\varphi \in \Phi$ we compute the number E_φ of similar pairs and the number D_φ of different pairs in the formula. Let W_φ be the *weight* of φ . Now the contribution $\sigma_\varphi(x, y)$ to the similarity score Σ of the pair (x, y) is:

$$\sigma_\varphi(x, y) = W_\varphi \cdot \frac{D_\varphi + E_\varphi}{E_\varphi} \text{ for } x \text{ and } y \text{ similar in } \varphi,$$

$$\text{resp.} \quad \sigma_\varphi(x, y) = -W_\varphi \cdot \frac{D_\varphi + E_\varphi}{D_\varphi} \text{ for } x \text{ and } y \text{ different in } \varphi,$$

$$\text{resp.} \quad 0 \text{ if } x \text{ or } y \text{ is missing in } \varphi.$$

Now the similarity score $\Sigma(x, y)$ can be defined as:

$$\Sigma(x, y) = \sum_{\varphi \in \Phi} \sigma_\varphi(x, y).$$

With this Σ we obtain a clear idea which manuscripts are close to each other, and which manuscripts are far apart. It is rather easy to make a pilot tree for the pedigree, (cf. Wattel and van Mulken 1995b). We successively replace the two manuscripts x and y , with the highest similarity score to a virtual intermediate manuscript m , and then state that the $\Sigma(m, z)$ for some other manuscript z is the mean of $\Sigma(x, z)$ and $\Sigma(y, z)$. The pilot tree is then found by connecting every x and y with the intermediate m that replaced it. For more details we refer to Wattel and van Mulken (1995b).

2.2. Centrality score

In order to get some structure in studying a text tradition we could pose the following question: which manuscripts belong to the main stream of the tradition and which manuscripts should be considered peripheral? Because a database yields only abstract information, we get no more than a formal algorithmic solution to this problem, which does not necessarily have consequences for the position of the *archetype*. To solve this, we develop a function on the manuscripts, which again has a mean of 0, in such a way that a manuscript which is central gets a positive score and a peripheral manuscript gets a negative score. Again there is no bias for lacunary manuscripts.

The idea behind this score is that manuscripts are central when they share the most common form in as many formulas as possible. Compared with the group behaviour often seen in youngsters, "in" is what "everybody" does, and the most popular member of the group is the one who can registers positive for as many aspects as possible of that which is "in".

The score for a manuscript x in a particular formula φ will be denoted by $\gamma_{\varphi}(x)$, and the sum of scores for x over all formulas is $\gamma(x)$.

Let us consider the formula of line 4 in section 1. Here the three manuscripts which should get a positive score γ_{φ} are c , e and f , because they share a reading with three manuscripts, while the other readings have a smaller number of manuscripts.

The way in which a score should be computed is not as obvious as in the case of the similarity score. The difficulty lies in determining how significant it is, that the three member reading cef is the largest. The two member reading ab is a close second, and if from some obscure place another manuscript turns up with this reading, then the inference that c , e and f is most common is no longer valid. So we must compensate in the formula for the fact that there is a runner up for the most common reading. This small difference yields less certain results if there are many manuscripts in the formula, and this could be used in the definition of the score γ_{φ} .

We define N_{φ} to be the number of manuscripts in the formula, C_{φ} to be the number of manuscripts in the most common reading, and S_{φ} the number of manuscripts in the next common reading. We then define the influence of the formula to be: $W_{\varphi} \cdot (C_{\varphi} - S_{\varphi})$, and use this number as a basis for the other computations. In our example $W_{\varphi} = 5$, $C_{\varphi} = 3$, $S_{\varphi} = 2$, and $N_{\varphi} = 6$, so our influence will be $5 \cdot 1 = 5$. Since there are three manuscripts in the common reading, and three in the other readings the application of the formulas below yield +10 for the manuscripts c , e and f , and -10 for a , b and g .

Next we take the formula with line number 1.6 as an example. We obtain $W_{\varphi} = 0.5$, $C_{\varphi} = 4$, $S_{\varphi} = 2$, and $N_{\varphi} = 6$, so our influence will be $0.5 \cdot 2 = 1$.

Now there are four manuscripts in the common class and two in the rest, and we have to multiply the scores in the common class by $1\frac{1}{2}$ and the others by -3 to obtain the scores. We get:

$$\gamma_{\varphi}(a) = \gamma_{\varphi}(b) = \gamma_{\varphi}(d) = \gamma_{\varphi}(g) = 0.75; \quad \gamma_{\varphi}(c) = \gamma_{\varphi}(e) = -1.5.$$

In order to define this score for arbitrary formulas we put:

$$\gamma_{\varphi}(x) = \frac{W_{\varphi} \cdot (C_{\varphi} - S_{\varphi}) \cdot N_{\varphi}}{(N_{\varphi} - C_{\varphi})} \quad \text{for } x \text{ in the common reading in } \varphi, \quad \text{resp.}$$

$$\gamma_{\varphi}(x) = \frac{-W_{\varphi} \cdot (C_{\varphi} - S_{\varphi}) \cdot N_{\varphi}}{C_{\varphi}} \quad \text{for } x \text{ not in the common reading in } \varphi,$$

resp. 0 if x is missing in φ .

The centrality score $\gamma(x)$ can be defined as:

$$\gamma(x) = \sum_{\varphi \in \Phi} \gamma_{\varphi}(x).$$

3. Centrality and Periphery

At the moment we have two score functions and from them it is possible to get a clear idea which manuscripts are central and which are peripheral. In any case it is possible to declare the manuscript with the highest centrality score as the ultimate centre, and to declare everything with a very negative centrality score as peripheral. From the similarity score it is also possible to recognise a extremely peripheral manuscript if we take one of the manuscripts in which the similarity score reaches its minimal value.

In order to find a collection of peripherals which is pairwise remote from each other and from the centre, we could take primarily the manuscript whose centrality score, plus the similarity score in relation to the other manuscripts, is heavily negative. Let p be the manuscript for which this minimum is reached. Other peripheral manuscripts, which should be considered separately, could be found by requiring that they should be far from p and far from the centre. We could require for the next peripheral manuscript q that $\gamma(q)$ and $\sigma(p, q)$ are both negative and that their sum is minimal.

The underlying idea here is that another manuscript which is peripheral in another branch of the tree should be remote both from the centre and from the first peripheral branch. If the structure of the tree was almost linear, i.e., there is a centre and two long branches extending from it, then we would have determined the entire structure when we have determined p and q and the centre. If the structure has another branch which extends far from the centre in yet a different direction, then it can be found by requiring that its most extreme end be far from p , from q and from the centre.

This procedure can easily be extended by requiring that the next peripheral r have a negative σ with respect to all a priori defined peripherals, and a negative centrality score γ . Again the choice of r from among the possible candidates is made by the requirement that the sum of those scores be minimal. Another iteration yields a manuscript s , far from p , q , r and the centre. The iteration stops automatically when no other manuscript exists which has all negative scores.

There now exists a simple clustering. We take the most central manuscript and the peripherals found thus far which will be the representatives of the clusters. Every other manuscript is compared to these representatives for its classification among the clusters. A manuscript is put in the cluster of the representative for which the score σ is maximal. This score is always positive, because the iteration has stopped.

If the number of manuscripts itself is too high to be handled, we could start to find a general structure by trying to deal with the cases in which only the cluster representatives exist and recompute the formulas for this restricted set of manuscripts. Recomputation of formulas for a restricted number of manuscripts and for the lumping together of a collection of manuscripts to a fictive combination manuscript is treated in the next section.

The next step is to recompute the collection of formulas for every cluster. Our problem is subdivided in (i) finding the tree for every cluster and (ii) finding the tree for the general structure of the cluster representatives. The complete tree structure should be found by overlaying the clusters on the positions of the representatives.

If this procedure is chosen there is one problem: *where in a particular cluster G is the attachment of the connecting edge of the general structure tree?* The easiest way to solve this is to make of the other clusters a fictive combination manuscript z , by means of the announced lumping procedure given in the next section, and solve the tree problem for the manuscripts of G with z . The attachment of z will be chosen as the attachment of the general tree. For the peripheral clusters the attachment found in this procedure is often a reasonable one. For the central cluster, it is often necessary to take a representative or a fictive combination manuscript for the different peripheral clusters, and see where the attachments are located in the central cluster itself. A better tree is then obtained if the area around the centre in the general structure tree is replaced by the central cluster with attachments obtained in this way.

It is also possible to make a construction of centre and periphery without using the centrality score function defined in 2.2. The central part of the text tradition can then be found as being close to zero in the similarity score for every peripheral. The difficulty with this choice is that it automatically puts highly lacunary manuscripts centrally. From the linguist's point of view this might be reasonable, since extremely old and often used manuscripts will be more worn out than newer and less used ones, but from algorithmic point of view this is disastrous, since those lacunary manuscripts are bad representatives for clusters.

Using both centrality and similarity has a small bias in favour of non-lacunary manuscripts as representatives, a bias caused by the fact that the representatives are chosen from the extremes; since lacunary manuscripts will often have scores of 0 , these will not be extreme.

4. Lumping Manuscripts

Since tree constructing algorithms can be highly complex, at many stages of the construction process it is necessary to restrict an algorithm to a small number of manuscripts. This means that some manuscripts should plainly be omitted from these algorithms and that some clusters of other manuscripts should be squeezed together in such a way that they are considered to be one single manuscript. Such a fictive combination manuscript will be called a *lump* in the sequel.

It is necessary to reorganise the set of formulas Φ for cases where some manuscripts are omitted and others are lumped. The easiest way to do so is to maintain the following rules:

- a* First, skip all omitted manuscripts and all lumped manuscripts from the head line, and add a code letter for each of the introduced lumps in the head line.
- b* For every formula, skip the omitted manuscripts.
- c* If in a formula all members of a lump are in the same reading, then replace these members of the lump by the lump code letter.
- d* If in a formula some members of a lump are in different readings, then split the formula up into parts. In each part every manuscript of the lump is deleted and the code letter of the lump is put in each of the readings in succession. The *weight* of the parts is computed in such a way that (i) the sum of all weights equals the original weight of the formula, and (ii) the weights of the parts are proportional to the *squares* of the number of occurrences of lump manuscripts in the readings.

For example, if the original formula was

$$6 \quad abcd/efghij/klmn/pqrst \quad 3.5$$

and the manuscripts to be lumped are *b, c, g, q, r, t* and $\&$ is the code letter for this lump, then the formula is split up according to the relative weights $2^2=4$, $1^2=1$ and $3^2=9$.

This means that the weights will have to be *1.0*, *0.25* and *2.25* and we obtain:

<i>6</i>	<i>a&d/efhij/klmn/ps</i>	<i>1.0</i>
<i>6</i>	<i>ad/ef&hij/klmn/ps</i>	<i>0.25</i>
<i>6</i>	<i>ad/efhij/klmn/p&s</i>	<i>2.25</i>

- e* If in a derived formula, after all omitting and lump-splitting is completed, all manuscripts have different readings, then the information content of the formula has vanished, and so the split formula itself is omitted. Moreover, if in a derived formula all manuscripts have the same reading, then this formula is omitted for the same reason.
- f* The database is treated on a line by line basis. So all skipping, lumping and splitting and omitting as defined in the steps **b** through **e** is done on one line of the database before proceeding to the next.
- g* The omission and lumping algorithm should always be started from the *original* database, even when it is done in several stages of the tree constructing process. This is necessary since the recomputation of the weights has some unwanted side effects, especially for the similarity scores, which should not cumulate in an iterative procedure.

It is now possible to lump each of the peripheral sections and establish a general structure by taking the central manuscripts and a fictive combination for each peripheral. The tree constructed from this setting gives a clear indication in which way the periphery is attached to the centre. Since this method also has the possibility that different peripheral parts are attached to different parts of the centre, it is more general and more flexible than the construction of the general structure by means of representatives as in the previous section.

5. Conclusions

In this report we have not discussed the tree-construction algorithms which should be used after the number of manuscripts is reduced to a manageable amount (cf. Wattel and van Mulken 1995b). The numerous possible geneological trees become a problem if the directions in which a construction can go are not clear.

The computational complexity of the number of trees in the general case is in the so-called "Nondeterministic Polynomial Complete" class, which in fact means that the best possible tree can only be found by the brute force method of comparing all trees, unless a method can be found to exploit the underlying structure of the problem. We hope that the knowledge that our trees emerged from a text tradition, and that the different formulas representing transcription errors, redactional adaptations and (alas!) contamination, will help to find a tree which represents the transcription process. Hopefully this supplies a sufficient amount of knowledge about the underlying structures in order to keep the tree construction process manageable. In any case, this paper emerged from the necessity of reducing the number of manuscripts in the tree construction processes.

Another consideration which follows from this high computational complexity is that we should be extremely cautious in reducing the manuscripts. In order to construct a peripheral part of a tree, it is advisable, e.g., to take the immediate vicinity of the representative together with a central manuscript. The central manuscript will indicate the attachment to the rest of the tree. The structure of the immediate vicinity of the representative is now known and need not be taken into account in the next step. In the next step, we lump the vicinity of the peripheral representative to one fictive manuscript and take a little larger environment of the representative in another tree construction process. In this way the peripheral cluster can be solved iteratively by always taking lumps as the new representatives, and a somewhat larger part of the neighbourhood of the representative is taken into account.

In a way, this paper is only a small part of a big machinery. The ideas how stemmatological information should be coded and processed evolved over several years. In Salemans (1990) some remarks are made on the classical results of Lachmann. The research in this area received an extra stimulance from the ideas of Dees (cf. Dees et al. 1988-1989; Dees 1992). Several papers in the annuals of "*Corpusgebaseerde Woordanalyse*" show part of this evolution (cf. Dekker 1985; van Mulken 1992; Schøsler 1989); further techniques used in the construction of the pedigree are found in van Mulken 1993; and in the articles in this volume by Wattel and van Mulken. The ideas of this particular paper evolved from discussions with Marcel Dekker and several others of the stemmatology colloquia.

One could doubt whether the construction of periphery and centre represents a part of the transcription process which actually took place. An affirmation has been that the described algorithm faultlessly spotted the structure of a simulated contaminated text tradition supplied by Onno Huber.

References

- Dees, A., M. Dekker and M. Hogehout Mulder. 1988-1989. "Een voorbeeld van stamboomreconstructie: Karel en de Elegast". *Spektator* 18-2. 96-118.
- Dees, A. 1992. "Les chartes dans la recherche linguistique et philologique". *Le médiéviste et l'ordinateur* 25. 23-27.
- Dekker, M. 1985. "Stambomen". *Corpusgebaseerde Woordanalyse*. 34-39.
- Mulken, M. van. 1992. "Gerd Simon: Ein neues Kraut?". *Corpusgebaseerde Woordanalyse*. 113-127.
- Mulken, M. van. 1993. *The Manuscript Tradition of Chrétien de Troyes, a Stemmatological and Dialectological Approach*, Dissertation, Faculty of Language. Amsterdam: Vrije Universiteit.
- Salemans, B.J.P. 1990. "Text genealogical remarks on Lachmann, Bédier, Greg and Dearing". *Leuvense Bijdragen* 79. 27-468.
- Schøsler, L. 1989. "Problèmes de Stemmatologie". *Corpusgebaseerde Woordanalyse*. 167-174.
- Schøsler, L. 1993. "New Methods in Textual Criticism: The Case of *Charroi de Nîmes*". *Medieval Dialectology*, ed. by J. Fisiak, 225-276. Berlin: Mouton de Gruyters.

Weighted Formal Support of a Pedigree

E. Wattel

*Department of Mathematics and Computer Science
Vrije Universiteit, Amsterdam, NL*

and

M.J.P. van Mulken

*Department of Language
Katholieke Universiteit Nijmegen, Nijmegen, NL*

1. Introduction

One of the main interests of the research into a medieval text tradition is the establishment of the relations between the manuscripts. If reliable copies of all manuscripts that have ever existed of one text tradition would be available the task of the philologist would be to describe a pedigree showing which manuscript had been copied from which, and which manuscript should be considered the ultimate original of the text, or the *archetype*, the common ancestor of all of the others. If every manuscript were copied from a single exemplar, then the differences between the manuscripts would occur in one copying step and all changes in readings and adaptations would either have a common origin or be independent. It would not be difficult to reconstruct the history of the manuscripts from the available material, since the transmission would have been a straightforward linear process.

In practice things go quite differently. It is well known that only a fraction of the medieval manuscripts still remains today. Contrary to the aspirations of other scholars, we consider it naive to pretend that pedigree-building can be used for reconstructing the real, historical transmission of the text. Furthermore, evidence has shown that many traditions are far from straightforward, even more so than has been assumed thus far. Linearity should not be inferred in such cases. The reason for the lack of success of the traditional stemma-building procedures is precisely that the possibility of non-linearity is not allowed for. Traditional procedures are often confronted with contradictive evidence: not all genealogical material points in the same direction. This is only natural, since noise is likely to appear, even in manuscript traditions.

Sometimes, however, the degree of noise appears to be so important that the traditional genealogist is in distress: he has but two choices, either to reject all contradicting formulas, i.e., to compromise his own input data, or to renounce building any pedigree, claiming the tradition to be "too contaminated". The traditional procedures simply do not foresee multi-exemplar transmissions.

In this paper we will demonstrate that a new method, incorporating the possibility of non-linearity, increases the probability of the construction of a reliable pedigree.

First, an (approximative) pedigree is constructed on the basis of the collection of surviving manuscripts, fragments or whatever may be available; it is hoped that this construction reflects the *mainstream* ancestral relationships as exactly as possible. In view of the general complications and lacunary information in medieval manuscript traditions, the mainstream tendencies are sketched first hoping for ulterior refinement. There may be substream relationships, contaminations, crossing influences and other accidental common readings, but we hope, that these "disasters" do not obscure the overall structure.

Because of all of the contradictory information in the relationships between the text, the best we can hope for is to establish a kind of majority rule. We construct a pedigree which "fits best" with all of the available information and which minimises the contradictions.¹

In speaking of "goodness of fit", we usually introduce some design choices, e.g., which contradictions are really bad and which deviations are only minor. It is well possible that these design choices influence the final result, i.e., the constructed pedigree. If the final result is not satisfactory or has serious deficiencies, this could mean that our design choices or perhaps even our original perception of the text tradition is misguided. In such a case an automatic procedure to render the best possible results from the available material could help us understand the consequences of our conjectures.

Even when all abstractions are made, all differences are evaluated and the "goodness of fit" is well defined, it can still be a difficult job to construct the pedigree which optimises this "goodness of fit" function in the text tradition.

¹ It goes without saying that if a manuscript tradition is simple (some classical traditions are known to be straightforward), then our procedure works just as the other methods and leads to a direct, simple pedigree.

In spite of the aid of modern computers, it is far from clear how to achieve this. The problem has moved into the area of "computational complexity", and finding the ultimate "best fit" cannot be guaranteed. It may be possible to design automatic algorithms which come close to the best fit and spot the main clusters.

The primary aim of this paper is to describe a set of algorithms for constructing a pedigree or at least an unordered graphical tree, which fits as exactly as possible the available information from the different readings of manuscripts of a text tradition. The simplest type of tree in this respect is the type in which (i) every surviving manuscript is represented as a terminal node and (ii) every other node in the tree represents a fictive (lost) manuscript, with precisely three connections to other nodes. Initial restriction to this type of trees is applied to reduce the computational complexity. In the final stage of the algorithms we can get trees without these type restrictions by contracting some of the connections. If, e.g., a connection between two fictive internal nodes is contracted, we obtain an internal node with a larger number of connections (cf. Fig. 1.a); if a connection between a surviving manuscript and an internal node is contracted, we get an intermediate available manuscript (cf. Fig. 1.b).

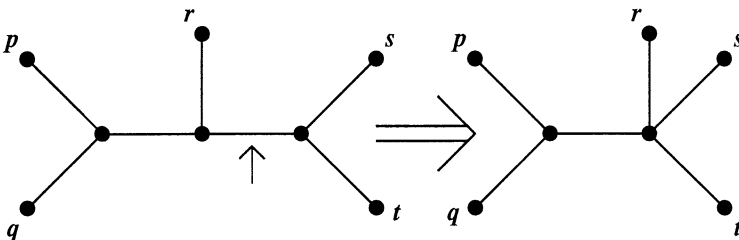


Fig. 1.a Contracting an internal connection

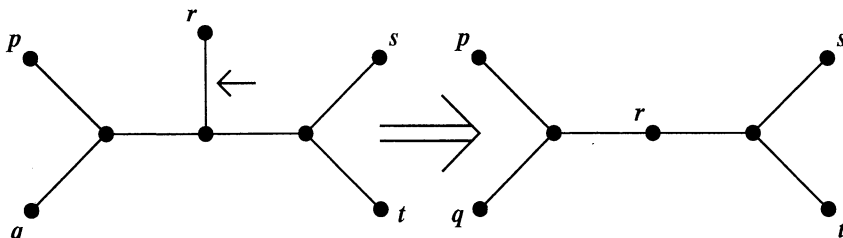


Fig. 1.b Contracting a terminal connection

The described method fits into the framework of the Three Level Method, developed by A. Dees (cf. Dees, Dekker & Hogeboom Mulder 1988-1989), which defines deep structures, intermediate structures and oriented stemmata.

The most important algorithm in the tree construction will be a "stepwise refinement" algorithm. If we start out from some tree satisfying (i) and (ii) then we evaluate how well this tree fits with respect to our evaluation functions; the fit can be improved upon by making small changes in the internal connections. A better tree is constructed by taking the change with the largest improvement and reevaluating the result. These steps are repeated until no further improvement can be made.

From the above it is clear, that the following steps are to be taken :

- a) define an evaluation function,
- b) define an initial tree,
- c) define the set of applicable changes and
- d) finish with the contraction of certain edges.

Preliminarily, we will describe in detail the *database*, i.e., the abstract material from which the trees are constructed (see also the other articles by Wattel and Wattel and van Mulken in this volume).

2. Material

The terminology of *graph theory* will at times be used in the description of the investigated structure. A graph is an abstract mathematical structure to describe certain entities, the *nodes*, and the interrelations between them. To visualise the structure, the graph is often represented as a figure in which the *nodes* are dots and the relations are line segments connecting the dots. A connection between two *nodes* is called an *edge*. A graph is called *connected* if there exists a *path* using *edges* of the graph between every pair of *nodes*. A connected graph is called a *tree* if it is cycle free, that is, if it is impossible to make a list of alternating nodes and edges which starts and ends at the same node and which uses every edge at most once. The *degree* of a node is the number of edges touching it. In our model the nodes are the manuscripts. Nodes with degree greater than 2 are often hypothetical intermediate manuscripts, while the surviving manuscripts are usually *terminal* nodes, i.e., those nodes with degree 1.

In a tree we can define *branches*. Take a *node* and an incident *edge*. Imagine that the other *edges* incident with this *node* are removed. The connected part of the graph which contains the *node* and the original *edge* is a subtree. This subtree will be called a *branch*. For every *node* and incident *edge* a *branch* can be defined in this way.

(The rest of this section and the first part of the next section is also presented in the other articles in this volume by Wattel and by Wattel and van Mulken.)

In order to study the several versions of a text tradition by means of automatic procedures, it is necessary to make a clear decision as to which way the material is to be presented, the so-called *data structure*. This *data structure* supplies a code reflecting an abstract model of the tradition and of the manuscripts themselves; if well chosen it admits automatic treatment without prejudice. Should this treatment yield unrealistic answers, even although it is internally consistent and correct, then the problems are to be sought in the assumptions and choices made by the philologist in preparing the material.

A flexible data structure admits a test for design choices, conjectures and properties of the tradition. This structure should also be clear enough to admit algorithmic treatment with a low complexity, thus making inferences by means of a computer feasible.

The data structure used here is to represent the text tradition as a *heading line* and *version formulas* (see the example below). The *heading line* contains the number of verses of the total manuscript as its first field and a coding letter for each manuscript in the tradition as its second field. The general structure of a *version formula* is a line containing the *line number* of a version difference as its first field, the *coding letters* for the involved manuscripts as its second field and a *weight factor* for the difference as its third field. The coding letters of different versions are separated by the sign "/" in the coding, while manuscripts which have an equal version are not. Line numbers are not necessarily different and are not necessarily integers. This admits more than one difference in one verse and separates observations at the beginning and at the end of a line. Usually it is assumed that version formulas are supplied in non-decreasing order with respect to verse numbers. To give an example of a small text tradition:

5	abcdefg		(heading line)
1.1	acd/be	7.5	(first formula)
1.6	abdg/ce	0.5	(second formula)
2	ad/bc/fg	2	(third formula)
4	ab/cef/g	5	(fourth formula)

In this example the story has 5 lines, with an important variant and an unimportant variant in line 1, a difference in line 2 as well as in line 4.

Furthermore, the differences in line *1* are not applicable to *f*, because *f* is lacunary or for any other reason which makes the formula inapplicable. The same is true for *g* in *1.1*, for *e* in *2* and for *d* in *4*. For instance, in line *4* *a* and *b* share a reading, *c*, *e* and *f* share a different reading and manuscript *g* has yet another different reading, which is in this case a single version.²

In the construction of pedigrees and trees we will use only this type of version formulas. Furthermore, there are some extra possibilities. For example, if some information is very important and occurs throughout the entire story, it does not fit a priori in such a type of version formulas. In order to take it into account we could introduce this information in an extra formula, which is inserted in the database for every verse number divisible by, e.g., *25*. If the original version of some formulas is known, this could be coded by using a fictive manuscript @ which is put in the favoured section of those formulas. If there is a sufficient amount of information of this type, then @ will give the position of the *archetype*. We then would automatically have a natural ordering on the tree, and we obtain a complete pedigree.

3. Construction of an Initial Tree

One of the most simple and useful ways to study a text tradition is to make a distance function which indicates which manuscripts are close and which are distant. The precise definition of such a distance, however, is not a priori obvious, and poses some problems if many formulas and manuscripts are lacunary. In this section we develop a *similarity score* which is meant as a type of distance function, but which is not necessary positive and which handles lacunary information without problems.³

² This means we do **not** restrict ourselves to "type-2 differences" like some other authors, which means that there are at least two corresponding manuscripts in every part of the formula and the formula has precisely two parts.

³ The main problem with lacunary manuscripts and ordinary distance functions is that a distance cannot be well defined if the necessary information is not available. If, e.g., *a*, *b* and *c* are tree manuscripts and all three go together whenever *b* is available, but if *a* and *c* differ on some places where *b* is lacunary, then there is no difference between *a* and *b*, and also no difference between *b* and *c*. So the distance *a-b* and the distance *a-c* are clearly *0*. In case of an ordinary distance function we now are forced to conclude that there is no distance between *a* and *c* and that they are equal. This conclusion is certainly not in agreement with the observations.

3.1. Similarity score

Since a distance function is defined on pairs of points, a function is developed using *pairs* of manuscripts in this paragraph which should replace the idea of a distance.

Consider, e.g., the formula for line 2 of the example where there are three different readings each with two manuscripts. With the six manuscripts in the formula we have 15 pairs, 12 in which the two members of the pair have a different reading and 3 in which the reading is equal ($a/b, a/c, a/f, a/g, b/d, b/f, b/g, c/d, c/f, c/g, d/f, d/g$ versus $a/d, b/c, f/g$). The weight of the formula is 2.

In this formula the event that two readings are different occurs often and the event that two readings are equal is rare; this means that the *information content* of the message in the case that two readings coincide is larger than in the case that two readings differ. In fact, one could defend the statement that the *information content factor* for equality is 4 times the *information content factor* for inequality, because 12 divided by 3 makes 4. In the definition of the *similarity score*, we require that the value of the *information content* over all pairs should reflect the weight of the formula, that equality should give a positive contribution and that inequality should give a negative contribution. With the defining formulas given below, we obtain the result that the pairs with a similar reading get a contribution of +10 and pairs with a different reading get a contribution of -2.5.

An important observation is that the sum of the contributions over the fifteen pairs equals 0. The fact that manuscript *e* does not occur in the formula means that no contributions are made to pairs of manuscripts with *e*, either in a positive or in a negative sense. Thus if the contributions for any pair of manuscripts over many formulas are taken, then a positive score means that these manuscripts go together in a substantial number of the cases that matter and a negative score means that the two manuscripts differ in the majority of the formulas in which both occur. *There is no bias built in for lacunary manuscripts* (cf. footnote previous page).

Let us give an explicit definition: Let x and y be two manuscripts, and let Φ be the collection of all version formulas. For every $\varphi \in \Phi$ we compute the number E_φ of similar pairs and the number D_φ of different pairs in the formula. Let W_φ be the *weight* of φ . Now the contribution $\sigma_\varphi(x, y)$ to the similarity score σ of the pair (x, y) is:

$$\sigma_{\varphi}(x, y) = W_{\varphi} \cdot \frac{D_{\varphi} + E_{\varphi}}{E_{\varphi}} \quad \text{resp.} \quad \sigma_{\varphi}(x, y) = -W_{\varphi} \cdot \frac{D_{\varphi} + E_{\varphi}}{D_{\varphi}} \quad \text{resp.} \quad 0$$

for x and y similar in φ resp. for x and y different in φ resp. if x or y is missing in φ .

Now the similarity score $\mathbf{\sigma}(x, y)$ can be defined as:

$$\mathbf{\sigma}(x, y) = \sum_{\varphi \in \Phi} \sigma_{\varphi}(x, y).$$

3.2. The initial tree

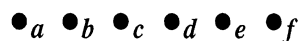
With this $\mathbf{\sigma}$ we obtain a clear idea which manuscripts are proximate and which manuscripts are distant. From this it is rather easy to make a tree for the pedigree as follows (cf. Fig. 2 and the example below):

- a) Look for the two manuscripts x and y with the maximal similarity score on the table.
- b) To replace this pair in the similarity score we introduce a virtual intermediate manuscript, m , and then state that the $\mathbf{\sigma}(m, z)$ for some other manuscript z is the mean of $\mathbf{\sigma}(x, z)$ and $\mathbf{\sigma}(y, z)$. Now the table of similarity scores is reduced with one row and one column.
- c) A part of the tree is then found by connecting x and y with the intermediate m that replaced it in the similarity score.
- d) If the similarity score table contains more than three rows and columns, we repeat the steps a) through c) to reduce it once more and to find another intermediate manuscript.
- e) The three surviving manuscripts in the similarity score table are connected to some new intermediate manuscript to complete the tree.

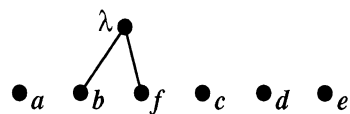
The tree constructed in this way will be called the *pilot tree* which will serve as a starting point for future algorithms.

In the example in Fig. 2 the similarity scores are in the tables at the left side of the figure and the graphs are at the right side. We start with six manuscripts, a, b, c, d, e and f . There are six rows and six columns in the similarity score table and six isolated nodes in the graph thus far.

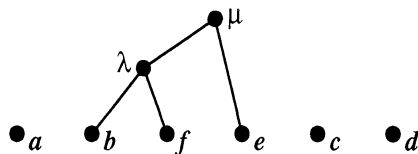
	a	b	c	d	e	f
a	*	0	4	-5	-2	0
b	0	*	-4	-1	5	7
c	4	-4	*	-3	-1	-2
d	-5	-1	-3	*	-4	1
e	-2	5	-1	-4	*	5
f	0	7	-2	1	5	*



	a	λ	c	d	e
a	*	0	4	-5	-2
λ	0	*	-3	0	5
c	4	-3	*	-3	-1
d	-5	0	-3	*	-4
e	-2	5	-1	-4	*



	a	μ	c	d
a	*	-1	4	-5
μ	-1	*	-2	-2
c	4	-2	*	-3
d	-5	-2	-3	*



	ρ	μ	d
ρ	*	-1.5	-4
μ	-1.5	*	-2
d	-4	-2	*

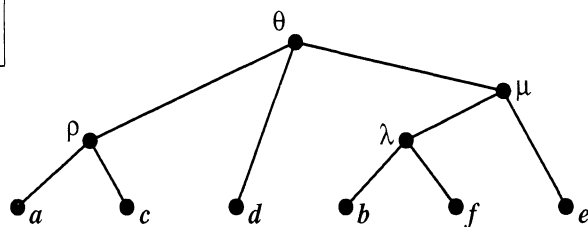
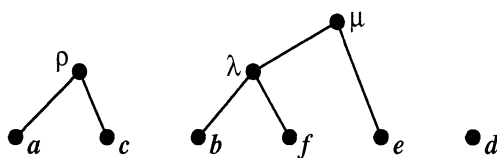


Fig. 2. The construction of a pilot tree from the similarity score

In the first step the manuscripts *b* and *f* have the maximum similarity score and are therefore replaced in the table by a fictive intermediate manuscript called λ . In the graph we take an additional node λ and connect it to *b* and *f*. The score of λ with respect to *a* will be θ , with respect to *c* it is -3 because it is the mean of -4 and -2 , with respect to *d* we obtain θ and with respect to *e* we get 5 . So we find the second table with 5 rows and columns. Next we find that the largest score is 5 in the entry belonging to λ and *e*. We introduce another node μ to connect them and replace the corresponding rows and columns in the table. Next the optimal score is 4 for *a* with *c* and we introduce another node ρ to connect them. Finally, ρ , μ and *d* are connected to yet another virtual manuscript θ in order to complete the tree.

Often it is convenient to use a representation of the pilot tree in a two line table. The first line lists the nodes; under each node is the connecting fictive additional node in the second line. In the example the representation will be:

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	λ	μ	ρ
ρ	λ	ρ	θ	μ	λ	μ	θ	θ

It seems important that λ was the first virtual node to be introduced because that corresponds to the maximum similarity score. One could use the positive integers instead of Greek letters to introduce fictive nodes. If we do that, then we obtain $\lambda=1$, $\mu=2$, $\rho=3$ and $\theta=4$, yielding the representation:

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>1</i>	<i>2</i>	<i>3</i>
<i>3</i>	<i>1</i>	<i>3</i>	<i>4</i>	<i>2</i>	<i>1</i>	<i>2</i>	<i>4</i>	<i>4</i>

From this representation we can obtain a single line tree representation with a minimum of redundant information which is highly suitable for automatic manipulation of trees. This particular representation is discussed in **Appendix A** and is used in the design of the programs.

In Fig. 3 the basic possibilities of pilot trees can be seen for four manuscripts. In these cases we have two internal nodes which should be numbered *1* and *2*. So there are precisely **6** possible tree representations for a pilot tree on four manuscripts. Each tree corresponds to two line representations since there are two possibilities of numbering the intermediate nodes. Therefore, those six representations correspond to the three different trees. In the sequel we will use this particular number because it means that for every internal edge of a tree there are precisely two alternatives.

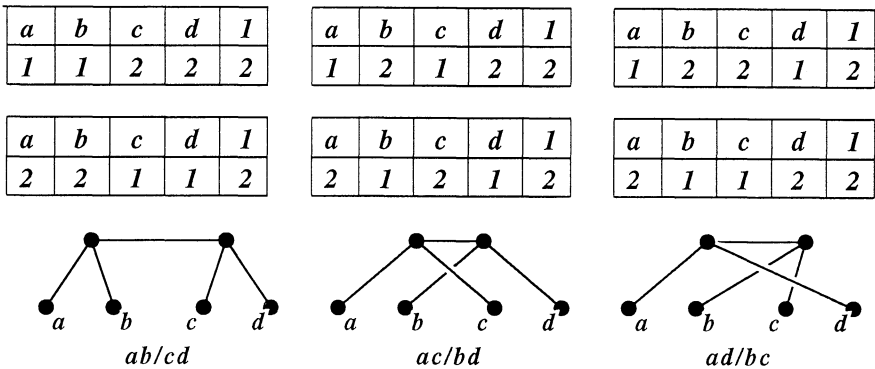


Fig. 3. The three different groupings on *abcd*

The problems which we encounter in case of four basic manuscripts are in some sense generic for all numbers of manuscripts. In fact, we call those configurations with four external and two internal nodes *quadruples*, and, as can be seen in Fig. 3, all quadruples allow essentially three different groupings. In the next chapter we will extensively use these observations. In order to help recognise the quadruples there are representations of the three possibilities in this figure.

4. Stepwise Refinement

Consider the following example : we have four manuscripts *a*, *b*, *c*, *e*, and four formulas: *ab/c/e*, *abc/e*, *ae/bc*, and *abc/e*. The construction the pilot tree (cf. section 3.2) first puts *ab* under a node *1*, then *1c* under a node *2* and, finally, draws an edge *2-e*, while the most sensible tree would be *ae* under *1* and *bc* under *2*. This is the only quadruple which agrees with all formulas. In two line representation this would mean:



This problem is not due to the fact that some formulas contain single versions. Consider the case that in an imaginary text tradition every intermediate has been reproduced in two (almost) identical copies.

For example, take $aa'/bb'/cc'/ee'$, $aa'bb'/cc'/ee'$, $aa'bb'cc'/ee'$, $aa'ee'/bb'cc'$, $aa'bb'cc'/ee'$, then we can make the same observations for the construction of the pilot tree. This example illustrates that the pilot tree need not be the best possibility; the main topic of this chapter is how to construct a better pedigree when the results are not yet optimal.

Starting from the pilot tree we try to make improvements in such a way that the resulting tree is in better agreement with the formulas of the database. In order to do this safely it is necessary to do it in small steps, because it is clear that the pilot tree already uses a lot of the information of the database and the necessary adjustments are probably minor. In the first part the combinatoric aspects of these possible minimal adjustments are discussed, and in the second part the support function which should govern the fitting procedure.

4.1. *Exchanging attachments*

The refinement consists principally of detaching a branch of the pilot tree and attaching it elsewhere. This implies that we split up internal nodes of the pedigree.

First we remark that all internal nodes of the pilot tree have degree 3. So if we split a node, we have one part of degree 2 and the other part will have degree 1 (see Fig. 4). Structurally a node of degree 2 is totally uninteresting, since if some variation occurs between its two neighbouring nodes, it does not contribute to the understanding of the copying process to assume that there was an extra intermediate fictive manuscript. Therefore, without essentially changing the displayed information, nodes of degree 2 can freely be omitted if the two incident edges are identified. If necessary, they can as easily be created in the middle of an edge in order to attach something else to it later on.

The splitting process and the reorganisation of the tree afterwards can be seen in Fig. 4. The top left tree represents the original situation. The top right graph shows the stage in which node 8 is split up, so that the node between 6 and 7 can vanish.

The minimal alteration which can be made from the original we obtain by putting the node 8 of Fig 4, which has degree 1 after the split, on one of the adjacent edges. Naturally we could put it on any other edge in the tree, but in that case we would not consider it a minimal exchange, and to consider all changes of that type would multiply the complexity of the program with the number of manuscripts as factor. In this way we obtain four possible minimal exchanges for each internal node.

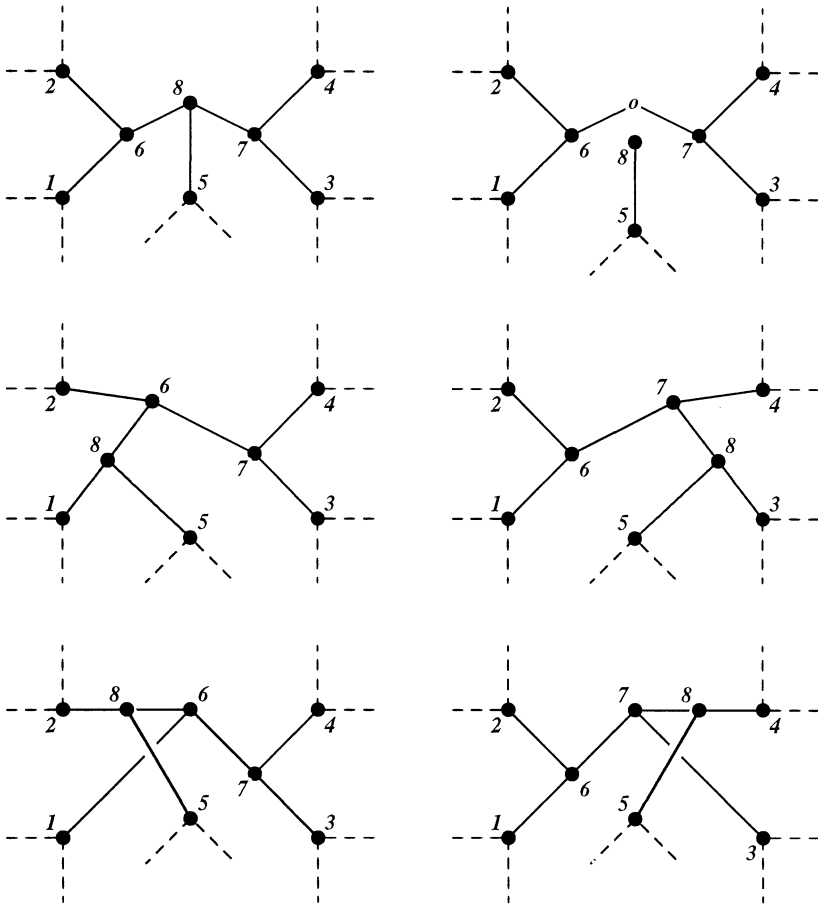


Fig. 4. There are four possible exchanges for each attachment

In Fig 4, 8 can be put between 6 and 1 (middle left), between 6 and 2 (bottom left), between 7 and 3 (middle right) or between 7 and 4 (bottom right).

The exchanges in this figure represent all of the possibilities. If one of the nodes adjacent to the split is terminal, then some cases will vanish. If in Fig. 4 the node 6 should be made terminal by erasing the nodes 1 and 2 and every edge connected to it, then we would only have the two possible exchanges on the right side of the figure.

Next we show that this process of exchanging attachments in a minimal way is functionally equivalent to the exchanging of quadruples. If we consider the left part of Fig. 4, we could consider the nodes **6** and **8** as the internal nodes of the quadruple with **1**, **2**, **5** and **7** as externals. The nodes **3** and **4** are not involved, because their links to **7** remain untouched. The original situation, the middle situation and the bottom situation then can be expressed by the following double line tree representations:

1	2	5	7
6	6	8	8

Original situation.

1	2	5	7
8	6	8	6

Middle left exchange.

1	2	5	7
6	8	8	6

Bottom left exchange.

It is obvious that the two possibilities at the right side of Fig. 4. can be obtained by considering **7** and **8** as internal nodes, and **3**, **4**, **5** and **6** as the externals of a quadruple. This exchange can be expressed by the following two line representations :

3	4	5	6
7	7	8	8

Original situation.

3	4	5	6
8	7	8	7

Middle right exchange.

3	4	5	6
7	8	8	7

Bottom right exchange.

If either **6** or **7** is a terminal node, then there is no possibility of exchange with that node, and so there is also no quadruple involved.

We have shown that detaching a branch and attaching it to adjacent edges implies the evaluation of five quadruples, on the condition that none of the adjacent nodes is terminal. Two quadruples result from the attachment of the edge to the adjacent branches on the left side, and two other quadruples result from attachment to the right side, and there is, of course, the original situation.

In fact we have shown now that the methods of exchanging internal connections and the regrouping of quadruples are equivalent procedures. Both of them can be considered to be the smallest type of adjustment that can be made on a tree graph in which internal nodes always have three connections. Moreover, if splitting an internal node does not have all four possible outcomes, then one of its neighbours is terminal and there is no corresponding quadruple.

The efficient implementation of the exchange of attachments in the computer program is still far from obvious. It is related to the single line tree representation mentioned at the end of section 3.2, and, therefore, it will also be treated in **Appendix A**.

In the original pilot tree, intermediate nodes which connect close branches have a low numerical label, while intermediates connecting branches which are further apart have higher numbers. It can be convenient to have an indication of the distances of the connected branches (larger distances usually yield less reliability). In exchanging attachments this information is at least partially lost, but in the algorithms of **Appendix A** we try to diminish this loss as much as possible. (So the numerical label of an intermediate node still gives an indication of the distances of the connected branches.)

4.2. *Dichotomy support*

The pilot tree introduced in section 3.2 is constructed from a the *similarity score* which is a kind of a distance function. This means that we restrict the information to *pairs* of manuscripts. It is generally agreed that genealogically the most revealing type of information results from the opposition of some manuscripts against some others. From this type of information it is difficult to construct any type of tree when there are many contradictions. It is much easier to improve an existing tree by means of small steps. For that reason we start from the pilot tree (call it T) which is made from the similarity score and which usually is not far from the pursued tree which optimises the full information of the version formulas. In that way we never have to construct a tree from the oppositions themselves, but during the improvement steps we can balance the influences of the inevitable contradictions.

Suppose that we have decided upon some tree T , and that we want to compare the results to other trees which emerge from T by the small adjustments as were made in the previous section. Then we need some objective criterion to do so. This criterion should explicitly be based on the version formulas and should define an order on the trees from which we can distinguish "better" from "worse". The easy way to do so is to assign a real valued numerical value to the combination of a tree and a version formula, which tells something about the "goodness of fit". The sum of these values over all version formulas for a fixed tree T will be called the *support* for T .

In this section we first define such a support function and then we supply a fast and easy algorithm to compute the support simultaneously for a tree T and all other trees which can be obtained from T in a single exchange of attachments.

Several of the provisions to be made to speed up the program have no consequences for the final result of the adjustments. Those details will be supplied in **Appendix B**.

Some paragraphs are supplied under the *design choice* heading. Those details can be changed, if the material or the intuition requires it, but usually these changes will have (probably minor) effects on the final result. The other paragraphs will have the *general* heading.

General

A tree consists of *edges* and the omission of each edge will divide the tree in exactly two parts. An edge which connects a single available manuscript to the rest of the tree does not carry quadruple information, and so such an edge is not useful in determining which quadruples should be exchanged. The other edges connect two internal manuscripts and the deletion of such an edge results in two non-trivial subtrees of the original tree. The deletion of such an edge will be called a *dichotomy*. The two internal nodes are each incident to two other edges, thus yielding four branches which are adjacent to the *dichotomy*. The exchange of attachments admits the regrouping of those branches into two pairs. We restrict ourselves in the definition of a support function to those *internal connections* only, and for each of those connections we have two alternatives from the exchange of attachments.

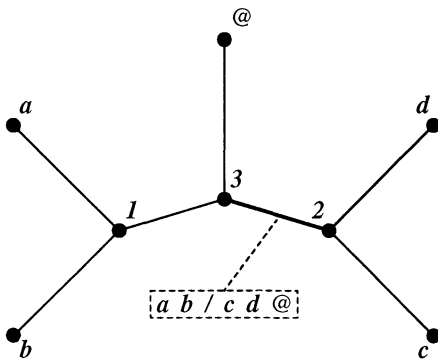


Fig. 5. Dichotomy consistent with a formula

Furthermore, every internal connection of the tree should be compared to every formula in the database. If all available manuscripts are mentioned in the formula, then we can be sure that all four neighbouring branches contain at least one available manuscript, and thus exchanging that connection favours some groupings of the manuscripts above others. If some manuscripts are not mentioned in that formula, then it can happen

that some branches do not contain available manuscripts at that point, and those formulas have only a limited value in deciding in favour of one grouping over the others.

In order to use the information content of a formula we subdivide it into those available manuscripts which share the same version. Those parts will be called the *readings* of the formula. If we have precisely two readings and a dichotomy puts all manuscripts with the same reading in the same subtree, then we have maximal agreement between the formula and the dichotomy. In those cases we say that the formula *supports* the dichotomy. In Fig. 5 the formula $ab@/cd$ *supports* the dichotomy 3–2⁴. The opposite situation is that more than one reading has available manuscripts in both subtrees of a dichotomy (for Fig. 5, e.g., ad/bc). In that case the formula is said to *contradict* the dichotomy. There exist situations in uncontaminated pedigrees in which precisely one reading has manuscripts in both subtrees. In Fig. 5 this is the case for formula $ab/cd@$ and dichotomy 3–2. This could occur because this formula supports dichotomy 3–1 and the pedigree also uses formula $ab@/cd$ which is somewhere else in the database. If each of the five manuscripts in the text tradition has single readings somewhere, then we have a completely uncontaminated pedigree. We will say that in such a case the formula and the dichotomy are *consistent*.

Design choice

In many text traditions there are several formulas which express a difference with more than two alternatives (a complex variant, cf. Greg 1927-1928). Before we continue with the general argument we will discuss the possibilities of using that information properly. We could leave the formula complete, but if there are many readings belonging to one side of a dichotomy and only two which have manuscripts in both sides, then the entire formula will be considered a contradiction. We are then obliged to dismiss the formula in view of the maximal support of the tree, but by doing so we would disregard the fact that only a part of the formula is in contradiction to the dichotomy, and, furthermore, discarding the formula would mean a loss of much supporting information. Moreover, if in some exchange of attachments more than one reading ends up at different sides of the dichotomy, this information gets lost if the formula is not split up.

⁴ The same thing can be said if there are more than two readings, but every reading has all its manuscripts in the same subtree. For example (see Fig 6.1), a formula $ab/cd/ef/gh$ would support the dichotomy 5–6.

We, therefore, tried to maximise the use of a complex variant by splitting up the information. We split up the formula in single oppositions and evaluate all pairs of readings: a formula $ab/cde/fg$ is split up into three partial formulas ab/cde , ab/fg and cde/fg , and each part is processed separately. In our program we have chosen for this solution, so we would become aware of its consequences.

First of all, since we only consider pairs of readings emerging from the partial formulas with a single split, many manuscripts will be missing in the formulas which we really evaluate⁵. For every single opposition formula all members of other alternative readings will have the effect of being lacunary. So any bias for lacunary manuscripts will have nasty consequences on the results. We will have to keep this in mind when we later fix the values of the design parameters.

If several alternatives exist, then many partial formulas will be evaluated, and the formula will have too much influence if there is no compensation for it in some way. If there are N alternatives, then there exist $\frac{1}{2} \cdot N \cdot (N-1)$ pairs, and every reading will occur in $N-1$ oppositions. As an example we compare the formulas $ab/cdefgh$ (a type-2 formula in Greg's terminology) and $ab/cd/ef/gh$ (a complex variant). The second formula will split up in 6 partial formulas and count six times, but it would be ridiculous to consider it as so much more important if the weights of the original formulas are equal.

The loophole out of the problem is an extra weight factor. If we emphasise the grouping of, e.g., ab , then we see that those manuscripts go together in both cases against the others. So we could also say that for the connection between a and b the information contents of the original formulas are equal. Our choice is that the weight factor for each of the partial formulas should be $\frac{1}{N-1}$ or $1/3$ in our example. This balances the effect that ab occurs three times in an opposition, i.e., in ab/cd , in ab/ef and in ab/gh . In its totality the formula with the large number of splits will be more important than formulas with a single split, but this extra weight factor is nevertheless reasonable, for it is taken into account later on. If we look at the information contents of those formulas and on the level of readings, the effect of this weight factor is consistent.

⁵ For the first part of the formula $ab/cde/fg$ which is evaluated, we have ab/cde and f and g are missing. In the second part we evaluate ab/fg and miss c , d and e . In the third part we evaluate cde/fg and a and b are missing.

General

We will be evaluating the branches and dichotomies of the pilot tree with the help of the initial formulas which have been decomposed if they result from complex variations. We compute the support for each dichotomy in the pilot tree.

To be prepared for all kinds of lacunary formulas, we must decide what to do with dichotomies in which some branches do not contain manuscripts of the formula. In order to handle these partial formulas, we described the different possible situations and studied the behaviour of support and contradiction in those cases. Then we tried to make reasonable design choices to cope with them. We have the advantage that the number of possibilities is small since we only consider internal nodes with three edges, and thus have a classification in only four branch types.

In Fig. 6 we illustrate the possible support with regard to the dichotomy 5–6. If the (partial) formula contains no information for certain manuscripts, then the branches in the pilot tree are represented by dashes. Keep in mind that a dichotomy always has four branches (cf. section 4.2).

The first possibility is that all four branches of some dichotomy contain manuscripts which exist in a given formula. This case will be called a *quadruple*. The name will not clash with the earlier definition of the *quadruple*: essentially the same object is considered in a different way.

In Figs. 6.1 and 6.5 we show two supporting formulas for the dichotomy between 5 and 6 which are of the *quadruple* branch type. It is possible to have formulas on the same branch type as in Fig. 6.1 which are only consistent, e.g., $abcde/fgh$, or contradicting, e.g., $abce/dfg$. In fact all formulas which contain the eight manuscripts $abcdefgh$ have the branch type of Fig. 6.1. Similarly, all formulas which contain the manuscripts $abceg$ have a *quadruple* branch type of Fig. 6.5. Among others we have, e.g., the contradicting formula acg/be and the consistent formula ab/cge . Also the formula bc/age is consistent with the dichotomy between 5 and 6, but certainly not with the dichotomy between 1 and 5, for which it is a contradiction.

The second branch type occurs when one of the branches of the dichotomy contains no manuscripts of the formula. Figs. 6.2 and 6.6 show two supporting cases for the dichotomy between 5 and 6, i.e., $abcd/ef$ and ac/f . This branch type is called a *triple*. For the type of Fig. 6.2 we can also have consistent formulas, e.g., $ac/bdef$, and contradicting formulas, e.g., abf/cde .

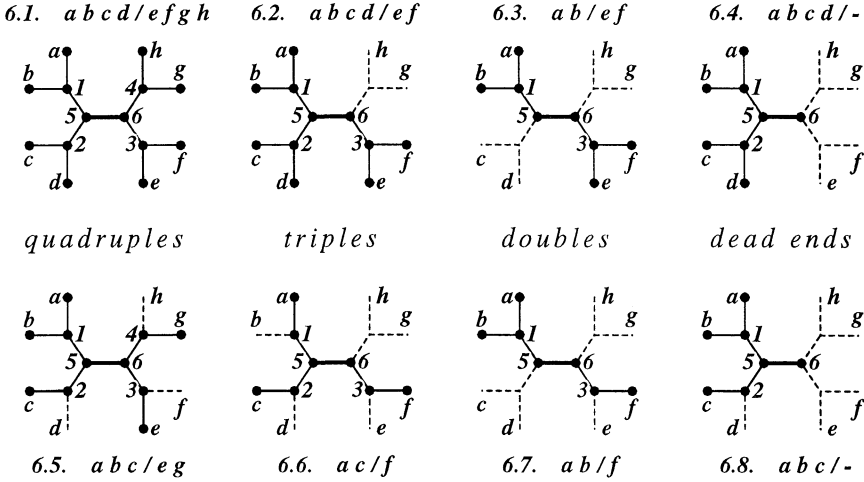


Fig.6. Supporting formulas for the four branch types

For the type of Fig. 6.6 there are only supporting and consistent formulas, since there are only three manuscripts involved while for a full contradicting formula we need at least four manuscripts.⁶

In the third branch type there exist only two branches for the dichotomy, but each branch is connected to a different node. We will call a member of this type a *double*. Figs. 6.3 and 6.7 show two supporting formulas for *doubles*. For the branch type of Fig. 6.3 there are also contradicting and consistent formulas, e.g., $a e / b f$ is contradicting and $a b e / f$ is consistent. For Fig. 6.7 there is no contradicting formula like in the case of a three manuscript triple and it is possible to have consistent formulas, e.g., $a f / b$ is consistent.

In Fig. 6.4 the formula $a b c d / -$ is an abstraction from, e.g., $a b c d / x y z w$. Since the group $x y z w$ does not take part in the group of manuscripts implicated in the 5–6 dichotomy, this formula contains no information about the reliability of 5–6. We call this branch type a *dead end*: a branch with node in the dichotomy which is only connected to branches without any manuscript in the formula. By definition such a formula always supports the dichotomy, but it carries no information. This type can be seen in Fig. 6.8 as well as in Fig. 6.4.

⁶ Several authors cannot use this branch type because they restrict themselves to formula which describe "type-2 oppositions" (cf. Salemans 1990).

Apart from the branch types we want to direct our attention to formulas which separate a single reading from some others. We see them in Fig. 6.6 and 6.7 Those formulas can never be contradicting, and even when they support a dichotomy their information content is disputed. We will speak about "type-1 variants" (following Greg 1927-1928). Their importance is much less than "type-2 variations". If the original formula before it was split up in pairwise oppositions consisted of many single readings and a small group with the same reading, then it is probable that there is some connection within this group; therefore these type-1 variants should not be omitted entirely, but it is reasonable to reduce their weight in some sense. Oppositions of one manuscript versus another single reading should be dropped because they do not carry any sensible information.

Design choice

After the classification of the previous *general* paragraphs, the design choices are just a matter of determining a sensible factor for each of the discussed cases. Let $w(f)$ be the *weight factor* of a formula f . For a given *dichotomy* d in a tree and a fixed formula f in the database we can give a support value for each of the possible situations.

Table 1. *Weight factors*

	<i>quadruple</i>	<i>triple</i>	<i>double</i>	<i>dead end</i>
<i>supporting</i>	$1 \cdot w(f)$	$0.4 \cdot w(f)$	$0.16 \cdot w(f)$	0
<i>consistent</i>	$0.125 \cdot w(f)$	0	0	<i>non-existent</i>
<i>contradicting</i>	$2 \cdot w(f)$	$1 \cdot w(f)$	$0.5 \cdot w(f)$	<i>non-existent</i>
<i>supporting type-1 variant</i>	<i>non-existent</i>	$\frac{w(f)}{64}$	$\frac{w(f)}{160}$	0
<i>consistent type-1 variant</i>	<i>non-existent</i>	0	0	<i>non-existent</i>

With this table we have a possibility of evaluating a single opposition formula on a dichotomy. The $w(f)$ which is used here does not need to be the original weight of the formula since we have already adapted it when it contained more than two readings in the previous design paragraph.

Moreover, because we observed that it is of little merit for a triple to be consistent with a formula, we do not give credit for it. Since the type-1 variant cannot be a quadruple and is at least consistent, we divided its credit for support by a factor 8.

We keep in mind the effects of exchanging attachments. Triples will always go to triples, and the same is true for quadruples; dead ends usually change to doubles and doubles often change to dead ends. For example, if we exchange the branches of $5-6$ in Fig. 6 in such a way that 1 and 3 are going together versus 2 and 4 , then the dead end in Fig. 6.4 changes into a double and the double in Fig. 6.3 changes into a dead end.

This means that the doubles and dead ends are much more important than a superficial observation would suggest. We sometimes lose and sometimes gain support by resolving a contradiction, if we change from a double to a dead end. If a double is changed to another double we never have a different support result. In triples it is also impossible for full support to change into full contradiction by exchanging attachments, and so it is sensible to emphasise quadruples.

In order to find the entire support $W(d)$ of a dichotomy d , we can add all contributions for it over all formulas f . To find the support of a tree we just take the sum of the $W(d)$ for all dichotomies d . Moreover, it is also easy to see how well a formula fits a tree by summing the contributions of that formula over all dichotomies of that tree.

We should constantly keep in mind that many of our formulas are lacunary. We tried to obviate this problem by making a design choice. If we look at the situation of Fig. 6.3, we would have the situation of an ordinary quadruple if the manuscripts $cdgh$ did not exist at all. The support for the quadruple case ab/ef would be $1 \cdot w(f)$, the support in the case of Fig. 6.3 would be $0.16 \cdot w(f)$ for the edge $5-6$, $0.4 \cdot w(f)$ for the edge $1-5$ and $0.4 \cdot w(f)$ for the edge $6-3$. That means that the total amount of support we get in this case is almost equal for those two cases. Our factor is 0.96 . In the case that we have the contradiction ae/bf the two triple dichotomies $1-5$ and $6-3$ would each have a factor $1 \cdot w(f)$, and the double $5-6$ would contribute $0.5 \cdot w(f)$. This would together give a factor 2.5 instead of the factor 2 which we would have in case of a simple quadruple.

The existence of many lacunary formulas has the disadvantage that many essentially quadruple formulas are split up into triples and doubles. The evaluation factors should be chosen in such a way that the split of a quadruple into two triples and a small (possibly zero) number of doubles does not have much influence on the sum of the support values.

The next choice to be made involves the steps along which minimality should be pursued. It seems reasonable to look at an initial tree and compare it to all other trees which could be derived from it by exchanging one attachment.

If the number of available manuscripts is M , then the number of fictive intermediate manuscripts is $M-2$ and the number of dichotomies is $M-3$. These numbers are a consequence of our initial assumption that every internal node should have exactly three incident edges. Since each internal edge admits two alternative dichotomies, we compare $2 \cdot M - 5$ trees, the original plus twice the number of internal edges. We choose to maximise the support function by starting from the initial tree and changing that internal connection which gives the largest improvement by the comparison as long as the exchanging of an attachments yields more support.

An additional advantage of this design is that if d' and d'' are the two alternatives for d , then it is sufficient to compare for these exchanges only $W(d)$, $W(d')$ and $W(d'')$ and not the support over the entire tree.

General

In the previous sections we showed that we construct a pedigree on the basis of a pilot tree based on general tendencies in the database. Some traditionalists will object that the only way to obtain the maximal security for having picked the right tree is to compare all possible trees.

We agree, then, that it is not obvious that these methods supply a tree which maximises the chosen support function. The general theory of complexity shows that in the related subject of the *Steiner trees* only the brute force attack of comparing all possible trees surely leads to the pursued optimal tree. If, however, the initial guess made by the similarity scores is not separated from the optimum by some low support valley, then the previous algorithm will surely lead to the summit.

In order to avoid the most probable type of suboptimality, it is advisable to make one additional try after the previous algorithm has stopped, i.e., by considering every other single exchange of attachments over a larger distance in the tree just once. If the maximum is real, this procedure will not give any improvement of results, but it could quiet our conscience. If we have a substantial gain, we advise to restart the procedures with the better tree as initial guess.

5. Edge Contraction

Up to now we have considered trees with internal nodes with precisely three edges. Other trees could be obtained by contraction of some of these edges, as shown in Fig. 1 of the introduction. The process of this contraction itself is easy: the real question is when it is favourable to do it.

From the previous chapter we can obtain a support value for every dichotomy, and it seems reasonable to contract internal edges which have negative support in the optimal tree. If the support for a dichotomy is negative, this renders no reason to choose this dichotomy in favour of its alternatives, and we could better do without it. So we *contract* the edge, which means that we discard it and identify its two incident nodes (cf. van Mulken 1992: 64-65). In addition, in the case that the support is positive but one or both of the alternatives have nearly the same amount of support, then a contraction should be considered.

To know when to contract a terminal connection is more difficult. No manuscript is perfect: even a true intermediate will have some versions which do not occur anywhere else. The "zéro caractéristique" of Quentin (1926) will practically never occur in reality. The criterion should be that we only have single readings which the medieval scribe should directly spot as being inconsistent, erroneous or "not done", while the alternative is unique and obvious.

Since this last criterion does not fit very well in a computer algorithm, we should try to manage with a numerical substitute, e.g., almost all of the encountered variants of the supposed intermediate should coincide with at least one of the alternatives of its environment. The next paragraph is meant to define precise criteria for handling those situations.

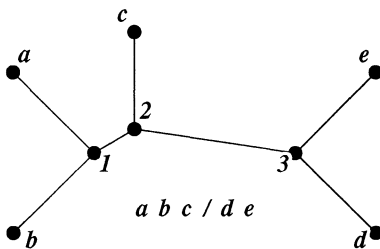


Fig. 7. Contracting 1 - 2

Design choice

Since we designed this procedure precisely for handling non-straight-forward manuscript traditions, we should always expect a certain amount of contradicting formulas due to coincidence. This is usually called "noise", but up to now we have not yet defined a measure for the degree of noise.

First, we can decide in the support evaluation program which part of the formulas is really contradicting. This *contradicting percentage* P_c will be used to estimate which contractions should occur. It is determined as the weighted average of the contradicting formulas divided by all formulas. The weight is the value the formula would obtain in case of support. This weight is used after the branch type and type-1 variant factors are accounted for, and in order to avoid zero weight, dead ends will be treated as doubles. The P_c is multiplied by 100 % in order to get a percentage.

This percentage is compared to the *supporting percentages* which are defined for every dichotomy and terminal edge in order to decide upon contraction.

As weight for the *supporting percentage* for the combination of a dichotomy with a simple opposition formula, we take the support value if support is possible, but only the consistency value if no support is possible even after an exchange of attachments. With this weight factor we can now evaluate the *dichotomy support* P_{d0} for every dichotomy as a percentage. We simply divide the sum of the obtained support values over all oppositions by the sum of the weights, and multiply it again with an extra factor of 100 %.

For every dichotomy we have evaluated the difference of support for the exchange of attachments. With the same weight factor we can make those differences also into a percentage if we divide the sum of the difference scores by the sum of the weights and multiply by 100 %. This yields the *dichotomy difference percentages* P_{d1} and P_{d2} . The contraction of the internal edges is now decided upon by comparing those three percentages with the contradicting percentage P_c .

As a general indication of the goodness of fit of the entire procedure up to now, we can use the weighted average of the P_{d0} with the same weight factors which were used to establish the *supporting percentages*. In this way we obtain the *total supporting percentage* P_t .

It is certainly possible that some dichotomy has a reasonably high score P_{d0} but that P_{d1} or P_{d2} are small. This happens in particular when we have three or more manuscripts which almost always go together (cf. Fig. 7). The reason could be that they are all close to a common source. Because we required that up to this stage all internal nodes have degree 3, we forced a dichotomy in the group (in Fig. 7 the edge 1-2).

Exchanging attachments of this dichotomy always forces one member of the group outwards, and so we often have consistency, and seldom support or contradiction. The effect is that the score P_{d0} is highly positive, and that the scores P_{d1} and P_{d2} are both small. It is obvious that edges of this type should be contracted. The threshold for contraction is set by our program at :

$$P_{d1} + P_{d2} \leq \min(2 \cdot P_c, 8.0 \%).$$

It can be the case that some of the manuscripts have many single readings, but that still the group has a common source. In that case only one of the P_{d1} and P_{d2} will be small. Also in this case we want to contract, but with a much smaller threshold.

We put as an indication for contraction :

$$\min(P_{d1}, P_{d2}) \leq \min(0.5 \cdot P_c, 2.0 \%).$$

The third reason for internal contraction is that the overall support for a dichotomy is poor. The obvious value for the threshold would be 0 , but for heavily contaminated traditions this would hide all internal structure that still exists. It is sensible to keep at least half of the dichotomies at this stage and choose those that are the best supported in this miserable collection. As a reasonable compromise we put here as a threshold :

$$P_{d0} \leq P_t - 10 \%,$$

and hope for the best. In this case we need not restrict ourselves to negatively supported dichotomies, because a dichotomy which scores 10% or more under the average is undesirable anyway.

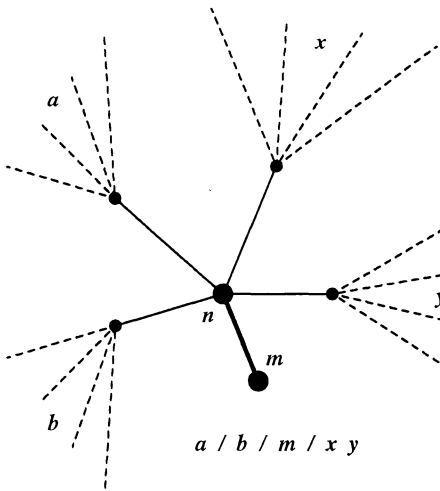


Fig. 8. Terminal contraction

In a similar way we should try to establish when a terminal connection should be contracted. Since this is the last stage in the contracting process we should by now be aware of internal nodes with many links. Let m be an available manuscript which is liable to be an intermediate one. Then there exists a link from m to some internal node n which should be contracted (cf. Fig. 8). We define that a formula is a *witness against contraction* for m if there are at least two available manuscripts x and y such that: (i) x and y have the same reading in this formula

and (ii) x and y are in different branches emerging from n and (iii) the reading of x and y is different from the reading of m .

This definition is in agreement with the notion of intermediarity of Quentin (1926), see also Dekker (1986). Let us put W_m for the accumulated weight of the *witnesses against contraction* of m and G_m for the sum of all weights of formulas containing m , then we can again obtain a percentage against contraction P_m by dividing those numbers and multiplying the result with 100% .

As a threshold for contraction we use :

$$P_m \leq \min(2/3 \cdot P_c, 5\%) .$$

Since sometimes more than one manuscript satisfies this criterion for the same internal node n , it is sensible to contract terminal edges in increasing order with respect to the values of P_m .

6. Final Remarks

Although the principles of the sections **4.2 Dichotomy support** and **5 Edge contraction** are quite general, it is impossible to use all of the discussed techniques without putting values for the evaluation factors and the thresholds. The values we have chosen are arbitrary, but we tried to choose these sensibly. Fine calibration of them has been done with the help of two well-known French text traditions, i.e., the *Charroi de Nîmes* and the *Perceval* of *Chrétien de Troyes*. The first of these traditions has a highly supported pedigree, but a fragment which is only on a small fraction of the text, and thus it could be used to calibrate the variables for lacunary formulas. The second text tradition is so highly contaminated that the support is quite negative, and it could be used to calibrate contraction to save some of the underlying structure.

In this report we have not discussed the algorithms which should be used to reduce the number of manuscripts to a manageable amount. This problem is studied in the article of Wattel in this volume. If the relations between the manuscripts change over the different parts of the text, then we may have a case where different quires have had a different history. This problem of change of relationship is discussed in the other article of Wattel and van Mulken in this volume.

This paper can be seen as only a small part of a big machinery. The ideas of how stemmatological information should be coded and processed evolved over several years. In Salemans (1990) some remarks are made with respect to the classical results of Lachmann. The research in this area received an extra stimulance from the ideas of Dees (cf. Dees, Dekker and Hogehout Mulder 1988-1989; and Dees 1992). Several papers in the annuals of "*Corpusgebaseerde Woordanalyse*" show part of this evolution (cf. Dekker 1985; Dekker 1986; van Mulken 1993), and also the articles of van Mulken and Wattel in this volume. The ideas of this particular paper evolved from discussions with Marcel Dekker, Pieter van Reenen, Lene Schøsler and several other people of the stemmatology colloquia.

References

- Dees, A., M. Dekker and M. Hogeboom Mulder. 1988-1989. "Een voorbeeld van stamboomreconstructie: Karel en de Elegast". *Spektator* 18-2. 96-118.
- Dees, A. 1992. "Les chartes dans la recherche linguistique et philologique". *Le médiéviste et l'ordinateur* 25. 23-27.
- Dekker, M. 1985. "Stambomen". *Corpusgebaseerde Woordanalyse*. 34-39.
- Dekker, M. 1986. "Intermediariteit". *Corpusgebaseerde Woordanalyse*. 61-69.
- Greg, W.W. 1927-1928. "The calculus of variants. An essay on textual criticism". Oxford.
- Maas, P. 1957. "Textkritik". Leipzig: Teubner. 27.
- Mulken, M. van. 1992. "Gerd Simon: Ein neues Kraut?". *Corpusgebaseerde Woordanalyse*. 113-127.
- Mulken, M. van. 1993. *The manuscript tradition of Chrétien de Troyes, a Stemmatalogical and Dialectological Approach*, Dissertation, Faculty of Language. Amsterdam: Vrije Universiteit.
- Quentin, Dom. H. 1926. "Essais de critique textuelle (Ecdotique)". Paris: Picard.
- Salemans, B.J.P. 1990. "Text genealogical remarks on Lachmann, Bédier, Greg and Dearing". *Leuvense Bijdragen* 79. 27-468.
- Schøsler, L. 1993. "New Methods in Textual Criticism: The Case of *Charroi de Nîmes*". *Medieval Dialectology*, ed. by J. Fisiak, 225-276. Berlin: Mouton de Gruyters.

Appendix A : *Tree representations and exchanging attachments*

We recall the pilot tree of Fig. 2. This tree had a two-line representation of the form :

<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>1</i>	<i>2</i>	<i>3</i>
<i>3</i>	<i>1</i>	<i>3</i>	<i>4</i>	<i>2</i>	<i>1</i>	<i>2</i>	<i>4</i>	<i>4</i>

Originally this double line table will end with the last two nodes on the rightmost places, but those entries can be deleted without any loss of information. Since we have a list of the original manuscripts in the database heading line, we can know for sure that the first line of the table should now be *a b c d e f 1 2*. At this moment this line does not supply any information. Why not omit it? The entire tree is now represented as a single line, in which each virtual node occurs exactly twice. In this way we have as the final one line representation for the pilot tree :

<i>3</i>	<i>1</i>	<i>3</i>	<i>4</i>	<i>2</i>	<i>1</i>	<i>2</i>	<i>4</i>
----------	----------	----------	----------	----------	----------	----------	----------

It is clear that this tree representation is highly suitable to use in an computer algorithm.

In the tree structure, the labels of the terminal nodes correspond to available manuscripts, but the internal nodes are at best hypothetical manuscripts and probably only auxiliary mathematical constructs for the algorithms, which makes the labelling of those manuscripts unimportant. In the pilot tree the internal nodes with low numerical labels connect rather closely related external nodes, while the high numerical labels connect branches which are far apart. If during the process of readjustment of the tree some quadruples are regrouped, this property of the tree labelling can not be guaranteed. Among other reasons, this is important because the one line tree representation is no longer valid if the two links with the highest numerical labels are no longer connected. Therefore we prefer that in any labelling of the trees, all branches which start out from the largest label and end in a terminal node pass through a decreasing sequence of internal labels. In order to guarantee this for the tree after the regrouping of a quadruple, we make the following two provisions :

- 1) The two internal labels of a regrouped quadruple are reassigned in such a way that the largest internal label is adjacent to the largest external label. Since the branches which are connected to the external labels are not touched, in most of the cases this requirement will be sufficient to keep the paths from the largest node to the terminals decreasing.
- 2) As long as some path from the largest node to some terminal is not decreasing we exchange the two labels involved. If there are several possibilities we first exchange labels which are close to the largest node.

If the node with the smaller number has two different immediate successors on two different paths, we exchange it with the largest of the two, and thus take care of the other clash simultaneously.

The effect of these label exchange rules can be seen in Fig. 9 in which the quadruple regrouping causes the difficulties, although we numbered it according to rule 1). Fig. 9.a displays the original situation; Fig. 9.b is obtained by a regrouping of the quadruple.

Since 8 is the maximal label, it should be adjacent to 7 after regrouping, and thus the label 3 is adjacent to 1 and 6. If the original tree had no labelling problems, then the maximal label can be found in the branch from 7 to 8 and may be beyond that. The labels 3 and 6 are in the wrong order and can be exchanged. This is done in Fig. 9.c. After this first exchange of labels we still have a problem, because 3 now clashes with both 4 and 5. According to the rule 3 should be exchanged with 5, thus also removing the clash with 4, as can be seen in Fig. 9.d.

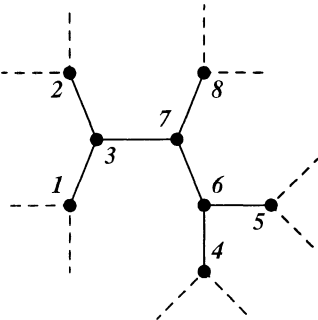


Fig. 9.a. Original

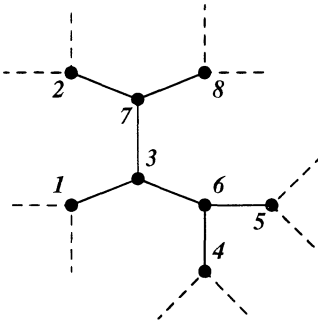


Fig. 9.b. Quadruple regrouped

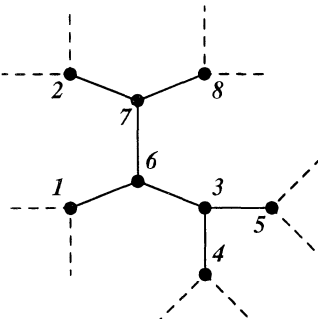


Fig. 9.c. Once relabelled

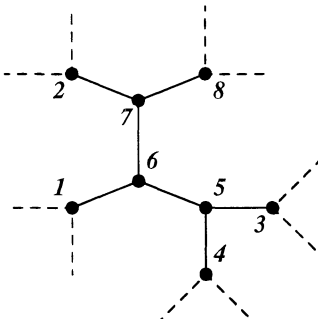


Fig. 9.d. Twice relabelled

This process of relabelling is bound to terminate, because the problems are always localised in the neighbourhood of the low label involved in the regrouping of the quadruple. Because this label follows a single branch and encounters ever smaller labels in that branch in the process, it is bound to find its place soon.

Appendix B : *Implementation of the support function*

The number of version formulas is usually much larger than the number of manuscripts that can be handled in a tree construction. It therefore makes sense to make a list of all branches of dichotomies. This is done first for the initial tree. For every internal edge of this tree we have two possibilities to change attachments.

So for every internal edge we only have to consider the difference in support if the branches are rearranged in the two possible alternative ways. If we, e.g., consider in Fig. 10 the edge $1-5$, we have the branches a , b , $2cd$ and $7@6\cdots$. Now a and b go together and 2 and 7 go together. The two alternatives for $1-5$ are: a and 2 together versus b and 7 together, resp. a and 7 together versus b and 2 together. So we must register the support function for the initial tree and the difference in support for the $2 \cdot (M-3)$ attachment exchanges for the $M-3$ dichotomies. We can compare those branches with the formulas in a single pass through the database.

We expect that it can cost several changes of attachments before we have obtained a final tree which maximises the support function. It thus makes sense to put considerable effort into the efficiency of the program. One of the ways in which this can be done is in the coding of the information. First an edge is coded by the minimum of its two incident nodes in which manuscript code letters are supposed to precede numbers. Thus the edge $a-1$ in Fig. 10 has code a , $@-7$ has code $@$, $3-6$ has code 3 and $5-7$ has code 5 . Since dichotomies are always internal edges, they obtain small numerical codes.

Even more important are the code numbers of the branches. We obtain them by assigning to every manuscript a bit position in an integer. The most significant bit will not be used in order to have a *null*-pointer possibility, and since the available manuscripts should also have a coding possibility, it is advised to use $2^n - 1$ as the maximum number of manuscripts if the long machine-word has n bits. We use 31 as a limit because our machine has 4 -byte longs, and so we have up to 28 dichotomies and we list at most 112 branches. To assign a bit to a manuscript, we simply number the manuscripts.

For Fig. 10 we obtain: $@=0$, $a=1$, $b=2$, ..., $h=8$. In the integers we use the *low order byte last* convention, so $@ \rightarrow 1$, $a \rightarrow 2$, $b \rightarrow 4$, ..., $h \rightarrow 256=0X100$ in hexadecimals. The branches can now easily be coded by the manuscripts they contain and for every branch we have a simple long integer. The four branches of dichotomy 5 (i.e., the dichotomy on the edge 5–7) contain ab , coded as 6, cd , coded as $24=0X18$ (hexadecimal), $@$, coded as 1, and $efgh$, coded as $480=0X1e0$ (hexadecimal). The initial dichotomy can now be coded as $30=0X1e$ versus $481=0X1e1$. The two alternatives are 7 versus $504=0X1f8$, resp. $25=0X19$ versus $496=0X1e6$. We will call alternative pairs of complementary integers the *dichotomy separators*.

In the start of the program we construct a branch table for each dichotomy. After constructing three dichotomy separators from the four branches, we can pass through the entire set of separation formulas. Formulas are subdivided into readings and every reading is assigned an integer in the same way from the manuscripts which share it. Next we can see whether a reading has manuscripts at both sides of a dichotomy by merely taking the *bitwise and* function of the code of the reading with both numbers in the dichotomy separator. If both *bitwise and*s are non-zero, then the reading is not entirely at one side of the dichotomy. If this happens for two readings simultaneously, then the opposition defined by those two readings is contradicting the dichotomy. If two different readings both have a zero in a dichotomy separator, then the dichotomy is supported by the opposition of those two readings. If we find one zero in a dichotomy separator compared with two readings, then the dichotomy is consistent.

Let us consider, e.g., the formula $abc/d@/e/fh$ in Fig. 10. The four readings get the codes $14=0Xe$, $17=0X11$, $32=0X20$ and $320=0X140$. If we take the *bitwise and* of $14=0Xe$ with the two integers $30=0X1e$ and $481=0X1e1$ of the original dichotomy, we get $14=0Xe$ and 0. So this reading is completely at one side of the dichotomy. The reading $d@$, which is coded by $17=0X11$, has as the *bitwise and* the outcomes $16=0X10$, resp. 1, and thus it exists at both sides of the dichotomy. In this manner $abc/d@$ is consistent with 5–7, which we could have seen immediately in the picture. If we compare the readings abc with e , we have an obvious case of a type-1 variant. e is coded by $32=0X20$; this number has a 0 *bitwise and* with $30=0X1e$ in the dichotomy separator. Both readings are thus at one side of the dichotomy and we have a supporting formula abc/e . Note also that e/fh is supporting in this sense, but we have an extreme case of a dead end here.

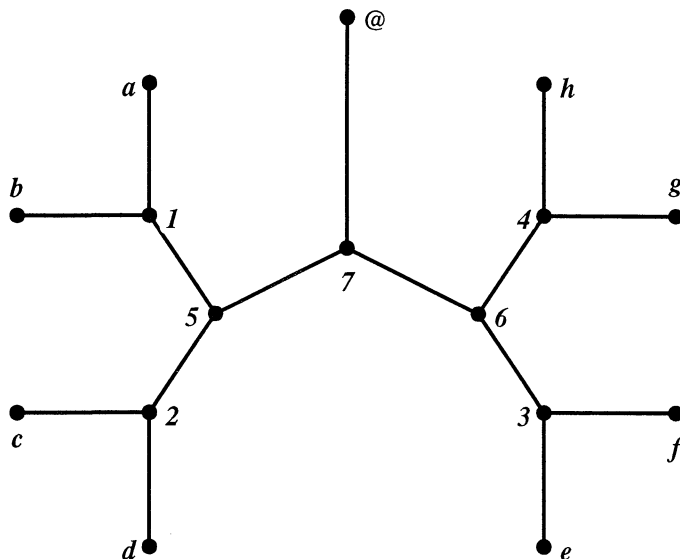


Fig. 10. A simple pedigree

Next the *bitwise or* of two readings is *anded* with the four branches in order to see whether we have a quadruple, a triple, a double or a dead end. We also check whether one or both readings consist of a single manuscript in order to recognise type-1 variants. With a single consultation of a table for the encountered combination, we can establish the evaluation factor for the opposition involved. Each of those operations can take place in a fast machine instruction, so that the program will be fast, although we have not improved the theoretical computational complexity.

It goes without saying that all of those steps can be taken for every formula in the database. By means of the assigned supporting factors we can compute the support for the tree and decide whether a simple exchange of attachments gives an improvement, and when it does, choose the best one. Along these lines it should not be difficult to write the necessary code to compare and exchange trees until an optimal value is found.

The Analysis of Early Islamic Traditions and Chains of Transmission

A.D. van Reenen

Department of Arabic and Islamic Studies

Vrije Universiteit, Amsterdam, NL

This paper presents a survey of the study of early Islamic traditions. After a general introduction to the texts and their use by Muslim and western scholars as legal and historical sources, respectively, methodological and practical problems regarding the analysis of traditions will be discussed. It will be argued that, given the prevailing lack of criteria for the dating of traditions, more traditions have to be described and a greater number of different characteristics must be taken into account than has been the case thus far. From this data, general profiles of traditions about a specific theme or from a particular place or collection of traditions can be deduced. These general profiles will enable the single tradition to be understood better and hence to be used as a historical source in a more reliable way.

In the second half of this paper, those characteristics of early Islamic traditions which most closely resemble the objects of stemmatological research will be described. In the field of early Islam, most of these phenomena have been discovered and described by G.H.A. Juynboll. Early Islamic traditions are particularly interesting because they are all preceded by a chain of transmitters. There is a distinct possibility, however, for each chain that it was forged and that the persons mentioned were invented. Even with a reliable chain of transmitters, clues have to be found to determine who originated the tradition and why. The most serious problem in this respect is the circular nature of much of the reasoning, since most of the discussions in the East and the West take the traditions themselves as both proof and argument.

1. A brief introduction to early Islamic traditions

One of the most difficult fields concerning literary sources and historical authenticity is the study of early Islamic history, i.e. the period between ca. 600 and 950 A.D. Though there is a myriad of available texts, few texts provide data that are reliable from the historian's point of view. Of early Islamic sources, the genre that presents the most difficulties with regard to historical authenticity is Tradition or 'Hadith' literature.

Islamic traditions claim to be authentic reports of deeds or sayings of the Prophet Muhammad, who reportedly lived between 570 and 632. The reports are intended as examples for believers to follow in their daily life and belief. Besides the Qur'an itself, traditions are the most important legal or normative source for Muslims. Subjects range from usury, purity, stories about the Prophets, rituals, clothing and heritage to the use of garlic in the mosque, prostitution, the Resurrection and dyeing one's hair.¹

On the average, each text is five lines long, including the chain of transmitters, usually six to eight in number, which precedes the actual text or story. The transmitters allegedly passed this text on, or "transmitted" it, from one to the other, starting with the first transmitter, supposedly an eye witness to what is related in the text, down to the last transmitter, who recorded the story in a tradition collection. The unbroken "chain of transmitters" between Muhammad and the compiler of the collection was intended to supply the text with the only authority that could make it binding, i.e. the authority of the Prophet himself. A typical tradition could be the following:

The tradition collector Bukhariyy heard from the transmitter Hisham who passed the story on from 'Abd-Allah, who transmitted it from Husayn, from Abu Sa'id, from Yahya, from Yusuf, from 'A'isha, the wife of the Prophet, that the Prophet, upon returning home, saw a curtain with images in the house and said: "Angels do not enter a house in which there are images". 'A'isha then removed the curtain.²

There are many tradition collections, the most important of which had been completed between the end of the second and the beginning of the fourth

¹ Cf. Robson 1971, vol.III:23-28; Juynboll 1982-1989, vol.6:45-49; vol.11:510-513.

² In this document "he" and "his" will be used for reference to the word "transmitter", whether male or female; in Ibn Hajar's *Tahdhiib al-Tahdhiib*, Hyderabad, 1325-1327 A.H., 12 vols., a biographical dictionary about transmitters, ca. 10.000 male and ca. 300 female transmitters are listed. The limited number of female transmitters occur almost exclusively as first in the chain, i.e. as eyewitnesses of the events described in the transmitted text part.

Islamic century (ca. 775 - 925 A.D.). These collections probably contain more than a hundred thousand texts.³

Within this large number of texts there are, however, many groups, that consist of, for example, three or thirty or more variants of what is basically the same text. These variants are found scattered throughout several collections, although variant texts can also be found on the same page. Completeness was considered to be more important than simply recording one version of "what really happened".

A good example of a cluster of variants is the group of texts containing the saying "Angels do not enter a house in which there are images, dogs, an impure person or bells". Distributed over twelve collections, the saying occurs in:

- 31 variants without any context
- 7 variants that start with the saying followed by a story about a sick man in whose house images were found
- 3 variants, also with a story about a sick man, in which it is discussed what word one of the transmitters used for "images"
- 30 variants, in 6 subgroups, about dogs in a house
- many other variants in 4 other groups of texts, where the variants appear as an addition to another text or in combination with other sayings.

It is difficult to reconstruct the modes of transmission of traditions in the first few centuries of Islam, since later Muslim writers superimposed their subsequent interpretations on the sources of this period, one main reason being to authenticate texts which had by then gained canonical status.

A number of the variants have characteristics that seem to indicate oral transmission. Lists of items that are permitted or forbidden are often set out in pairs of opposites, in rhyme or with parallel grammatical constructions. Some texts as a whole have a numeric structure, such as three parts or two times two related elements. A further example is that many stories have a similar narrative structure (Introduction including time and place, Action by believers, Action or saying by Prophet, Reaction by believers, Specification by Prophet). Although this structure seems to stem from the function of the story as a source for legal rules, its presence will also have been reinforced by the conditions of oral transmission.

In spite of the difficulties in reconstructing this period, it seems clear that

³ The Islamic calendar is lunar and begins at the year of the 'hijra', which is 622 A.D.; therefore, the abbreviation A.H. (Anno Higrensensis) is used; 175 A.H. - 325 A.H. = 791 A.D. - 936 A.D.

the transmission of traditions evolved gradually from an originally fully oral transmission to a predominantly written transmission at the end of the early Islamic period. However clear and “logical” this development may seem, the relationship between the oral and written stages of Islamic tradition and their respective characteristics still raises many questions. It is not yet clear precisely when orally transmitted texts began to be written down. Studying the texts themselves for clues does not yield clear criteria for determining the stage of development of specific texts in terms of the mode of transmission. Furthermore, writing the text down may not always have been the final stage of a text in transmission. It is probable that “new” oral versions continued to be made after the text was recorded in one of the collections that later came to be regarded as the canonical collections.

2. Authenticity

As already stated above, traditions claim to be authentic reports of an event in the life of the Prophet Muhammad. The authenticity of most of these “historical stories”, however, is questionable for a variety of reasons.

As early as the eighth century, Muslims themselves had realised that many traditions were spurious, that frequently the contacts between transmitters simply could not have taken place and that often texts contradicted each other. Many Muslim scholars devoted themselves to establishing the reliability of the texts. Extensive biographical dictionaries about the transmitters, still preserved today, contain information on the place and date of birth and death of more than ten thousand transmitters (a number of whom are completely fictitious), on those to whom and from whom they transmitted, accompanied by statements of tradition scientists about their reliability and details from their lives.

An intricate system of technical terms was developed, denoting a large variety of transmission situations. These terms reflect the later Islamic dogmas about the development of the Islamic community, its sacred writings and their authority through a chain of originally oral transmitters. In spite of this fact, it is still worthwhile studying these technical terms, since they do contain information about the transmission situation in the first few centuries and about the practices of transmitters who forged traditions and caused chains to comply with gradually evolving standards.

Apart from general practices and evaluations, it appears that the greatest attention was devoted to the study of the chain of transmitters, the manner of transmission and the reliability of the transmitters. Also, but much less

frequently, terms were used to discuss the textual part of a tradition and the position of a tradition in a comparison of texts. Examples of each of the categories will be listed below.⁴

General terms

- *da'if* 'weak'; not fulfilling the required conditions
- *maqbul* 'received' generally; fulfilling all conditions
- *mawdu'* invented, false
- *mu'allal* a chain or text with a hidden fault
- *sahih* 'genuine'; fulfilling all conditions; also: collection of traditions of such type

The chain of transmitters

- *'aziz* 'precious'; an authentic tradition coming from two Companions (the first generation of transmitters after the Prophet are called the Companions; depending on the definition, they knew him, had only seen him or were merely contemporaries)
- *gharib* authentic, but resting on the authority of only one Companion
- *hasan* 'beautiful'; not absolutely faultless
- *maqtu'* a chain which is interrupted or 'cut'
- *marfu'* record of a word or deed of the Prophet recorded by the Companion who saw it
- *marshhur* a tradition vouched for by more than two Companions
- *mu'allaq* 'suspended'; without the name of the Companion
- *mu'dal* a chain from which two or more transmitters are lacking
- *mudallas* a tradition falsely ascribed to an early authority
- *munqati'* 'interrupted'; a chain from which a name has disappeared
- *mursal* a text without a chain, or one with an incomplete chain, or without the name of the Follower (the generation of the Followers is the second generation of transmitters after the Prophet)
- *musalsal* having a chain of authorities reaching back to the Prophet
- *musnad* a tradition whose chain goes back to the Prophet; also: a collection of such 'supported' traditions
- *mutawatir* reported by numerous authorities or transmitters
- *muttasil* a tradition with an uninterrupted chain

The reliability of the transmitters

- *mubham* coming from one of whom one knows nothing but the name
- *mutqin* accurate reporter
- *thiqā* trustworthy reporter

⁴ This list is a compilation of the most important terms found in Guillaume 1924:181ff. and Robson 1971. The meaning of a technical term can vary in different periods and with different Hadith critics.

The textual part of the tradition (the *matn*)

- *'ijmaliyy* referring to many things
- *mudraj* a tradition in which a gloss or observation, interpolated by a transmitter, is impossible to separate from the words of the Prophet
- *mudtarab* a tradition in which a word has become 'misplaced', added or suppressed, or suffered has any kind of derangement
- *nasikh* abrogating; in other words: rendering a previous tradition or verse from the Qur'an invalid

A comparison of texts (apparently with the same subject)

- *ma'ruf* weak tradition, yet known because it is confirmed by another
- *mahfuz* one of two suspicious traditions which has a slight advantage over its rival
- *mardud* a tradition from a doubtful source which contradicts a tradition of good standing
- *mukhtalif* two traditions which are in apparent contradiction, but which can be reconciled
- *munkar* a tradition of weak authority contradicted by a weaker one
- *shadhdh* 'exceptional'; a tradition of good authority yet in conflict with another similarly attested tradition
- *shahid* a tradition from a Companion 'bearing witness to' or confirming one from another Companion

According to the sources, the circa hundred thousand texts in the now accepted or canonical collections are only a fraction of the original number. Critical as all this may seem, however, the Muslims in that era used these texts for completely different purposes than Western islamologists now do, given their entirely different cultural and scientific tradition. From the viewpoint of Western islamology, the Muslims were, and are, far less critical than the sources seem to demand. It has, furthermore, become clear that much of the information used by Muslim scholars for assessing the value of the traditions and writing the biographical dictionaries was derived from the traditions themselves. Sometimes a transmitter who never existed appears in a biographical dictionary with an entry describing his entire life. Probably only minimum research would prove the interesting law that "the longer since the death of a transmitter, the more information becomes available about his life".⁵

⁵ Cf. Creel 1963:48: The Traditional Chinese History records a long line of early emperors who are supposed to have reigned before the establishment of the first dynasty, during the third millennium B.C. There are no scientifically established historical data about these model emperors, among whom Yü, Yao and Shun are most famous, and they are not mentioned in documents written before Confucius. Starting with Tsui Shu, as early as

Though relying heavily on the works composed by their colleagues of a millennium previously, Western islamology has different objectives, and, although material and findings do not differ greatly from what Muslim scholars had already found, the approach and the conclusions drawn from the findings vary considerably.

Western islamology has always attempted to use traditions as a historical source. In the last century it became clear through the work of Goldziher, in particular, that most of the traditions probably did not come into being before 700 A.D. and that many in fact originated later. A tradition dating from, let us say, 750 A.D. would be projected back to the Prophet. To complete the chain from the actual date of origin to the Prophet, chains have been completed with real or fictitious transmitters, preferably with names that make a reliable impression or names of people who had already been declared trustworthy; on occasion even whole new chains were fabricated.⁶

Many scholars after Goldziher considered the chain to be hardly more than a rhetorical device whose single function was to convince the audience of the normative value of a text. The only historical information that could be extracted from these texts would be about those who fabricated them and not about the life of the Prophet Muhammad, the actual subject of the text.

Schacht and Juynboll in particular (see references) devoted more attention to the chains of traditions, showing that much historical information can be derived from them. Juynboll also showed that, exercising extreme caution, historical information can be unearthed from the biographical dictionaries.

As far as the textual part is concerned, it has become clear that texts were altered both by transmitters and by the final collector, according to different contexts and purposes, e.g., as biographical, political, legal and other sorts of traditions. Sayings were embedded in a story, texts were combined, shortened or, on the contrary, embellished with legal and narrative details, literary *topoi* were inserted and enumerations of forbidden practices were

already the 18th century, Chinese scholars have discovered a relation between literature and these mythical emperors: the later an emperor appears in literature, the earlier is the date assigned to him and the more detailed the information about events in his life. The reason for this is that, "as this mythology expanded, new figures had to be assigned to vacant periods, and only the earlier periods remained vacant"; also cf. Gates 1936.

⁶ Goldziher 1890. An interesting phenomenon is described by Juynboll (1991:156): A number of transmitters, mainly from Kufa, allegedly have "reached ages at death between one hundred and one hundred and twenty years or more". They seem to have been the answer to "the demand for simple strands bridging some one hundred years via as few people as possible" in the first century (p. 171).

made, with varying length.⁷

Considering the fact that many traditions probably did not originate at the actual time of the Prophet, it is important to establish a date of origin for the texts, so that they can provide information for a reconstruction of the early Islamic period. This task involves the following questions:⁸

- 1) First, where, when and on whose authority did the tradition originate, and what were the contents and the form of this original version?
- 2) Second, who transmitted it subsequently, and how was it changed between the point of origin and the point of fixation in the tradition collections?
- 3) Finally, what are the date, contents and form of the tradition in its present form?

In practice, traditions are dated mainly by gathering all the texts on one subject and then comparing and contrasting the contents of the text. Ideally, the transmission chains should be studied just as intensively, but few scholars have devoted themselves to the study of this aspect of the traditions.

Until now, no criteria have been found to distinguish between reliable and unreliable information. Each debate at an international Hadith colloquium is devoted, in part, to the question of authenticity. Opinions vary from "much material can be accepted as authentic and credible" to "we do not know who Muhammad was, who wrote the Qur'an and when", and "most of the traditions are fabrications from at least a century after Muhammad".

My own position is that it is too early to discuss the dating of traditions and the determination of their authenticity. Western islamology is hindered by its major objective, i.e. the use of traditions as a source for historical information. Eagerness to date a specific tradition has prevented us from establishing characteristics for traditions in general, their variants, their text structure and general or specific features of the transmission chains. As to methodology, arguments are voiced like "but of course, there is no sense in denying the authenticity of all these texts, since you would have nothing left to work with and you would have to start all over again"; or: "we do not need to go through the methodological debates which other disciplines have had to go through". Many arguments are statements rather than evidence and are hardly convincing. The methodological debate will probably continue as it has done for the last century, i.e. without noticeable progress.

⁷ Stetter 1965; Reenen 1990:56-60.

⁸ Cf. Juynboll 1983:73.

Therefore, instead of contributing my own, rather sceptical, arguments to the discussion, I refrain from dating and I will, in fact, start all over again. In the project I am involved in at the Vrije Universiteit in Amsterdam, I restrict myself to a description and analysis of a large number of texts and their similarities and differences. This will enable us to base discussions on the description of a corpus which is sufficiently large to allow us to draw some general conclusions.

3. Description and Analysis: Theory

In order to be able to describe and analyse a large number of texts, I considered a number of articles written by colleagues and myself and systematised the questions that arose from them. This I used to derive what kind of questions are asked and what data are required to answer these questions.

It appeared that a distinction can and must be made between questions about the narrative and the normative characteristics of the texts. I developed the hypothesis that the narrative elements and structure used were selected to convey the normative message of the tradition. For example, if someone in the second or third century after Muhammad would like to propagate his opinion about the conditions of purity, he would usually choose or compose a story in which the action takes place in the house of the Prophet and the main actor, apart from the Prophet, would be a woman.

Added to the initial historical objective of this field of study, we find three aspects or approaches that can, in my opinion, be discerned in the study of early Islamic traditions: a textual, a normative and a historical approach. The texts contain information relevant and applicable to each of these approaches: for Muslims they function as narrative texts and as normative sources for legal debates, and for the present-day researcher as historical sources. For a clear understanding of early Islamic traditions, all three aspects have to be studied.

Another result of the analysis of questions raised in the study of early Islamic traditions was that it appeared that historical unreliability was not the only problem involved in the dating of early Islamic traditions. The most important of these problems will now be discussed.

The first problem is a practical one and involves the sheer number of traditions and large biographical works about the transmitters. In order to make representative statements, a great number of texts is required. A considerable number of these works is, however, not easily accessible: they are sometimes hardly legible, often unvocalised and sometimes without indexes.

Consequently, usually only a small group of texts, taken from the more accessible sources, is studied.

The second problem is that most of the other early Islamic sources are equally historically unreliable. In fact, there are hardly any uncontested data, which could be used as points of reference. For a proper evaluation of traditions, we are forced to rely almost entirely on these selfsame texts: our reconstruction of that period is based on sources for a proper evaluation of which we need this same reconstruction. The danger of circular reasoning is obvious.

The third problem is that traditions, as we have seen, are short: five or six lines, including the transmission chain. As a consequence, such texts contain only little information and the usual methods for textual analysis can not be used as there is not enough material to discern a particular style, frequencies are too low, etcetera. A further, related, complication is that texts in a group of variants of one story or saying usually closely resemble each other, having few distinctive features.

In order to overcome these problems, one has to describe large quantities of traditions, using every possible characteristic. A tradition can only be properly identified as unique within a group of variants once many different characteristics are used. Only large quantities of data will reveal the relations between individual traditions that would otherwise go unnoticed. Finally, there is a need to establish which possible characteristic or combination of characteristics have been determinative in the origin and development of traditions.

After describing the texts exhaustively, one can compare and contrast all of this information and study texts as clusters of characteristics in relation to other texts or clusters. In this way, texts could be better understood, and thus be more precisely dated, both as unique examples and as members of groups of texts sharing certain characteristics. It would then be possible to make profiles of individual traditions, or of all traditions with a particular subject, or of the texts and transmitters of a particular city. Hopefully, in the future, profiles of specific periods can also be produced. Traditions that present problems for historians can be matched against these profiles.

This approach may appear to be obvious and simplistic, but a comparison of all of the texts is not a straightforward task. Existing methods of analysis for early Islamic texts have failed to date traditions, because groups of tradition variants are often only discussed as if they formed a self-contained unit with little or nothing in common with other groups of tradition variants.

One can have serious doubts about these limitations from the viewpoint of oral tradition theory (Vansina 1985); in addition, a reading the *Concordance*

et Indices de la Tradition Musulmane (Wensinck *et al.*, eds. 1939-1988) provides proof of the opposite. There are more similarities and less differences between individual traditions than would initially appear; this is equally valid for traditions on a wide range of subjects. Traditions can not be studied in isolation from the genre as whole.

The many similarities between these texts involve not only the wordings, but also the narrative and legal structures, stereotypical roles of characters in narratives, patterns in the transmission chains and similarities between texts that circulated in a particular city during a specific period. The originators, the transmitters and the collectors / editors were acquainted with many traditions and were probably influenced by this knowledge when they transmitted or created new traditions. According to Noth (1973:24), in dealing with historical texts, and Raven (1988:209), in dealing with traditions, one should have as much knowledge as possible of other texts of the same genre in order to be able to judge one text properly.

All of this is in itself not sufficient, since traditions often are, as we have already seen, members of a group of variants. How can we best compare these groups within the genre of Hadith as a whole? In principle, of course, each characteristic provides a good basis for comparison. As was stated earlier, there is a need to establish which possible characteristic or combination of characteristics have been determinative in the origin and development of traditions. An obvious basis for comparison is the subject of a tradition, which must have determined the degree to which texts influenced each other in terms of their textual development. Transmitters will have been especially versed in certain areas of belief, ritual or history, and many professional transmitters probably knew how to create a typical theological tradition about the Day of Judgement or a typical entertaining story about the Prophet.

Therefore, adapting the statements expressed by Noth and Raven, we could say that in order to be able to judge one text properly, one should have as much knowledge as possible, to begin with, of other variant texts, then of other traditions with the same subject, of traditions with related subjects and, finally, of the whole genre of Islamic traditions in general.

When applied to the analysis of traditions and their dating, this means that, first, traditions should be described using as many characteristics as possible. Thereafter a comparison of all variants of a certain story or saying should be made in order to combine the findings about various traditions covering roughly the same subject, thus allowing a clearer picture of the differences and similarities between related traditions to emerge. The result could further be combined with the findings of the analysis of related sub-

jects and these again with Tradition literature in its entirety and early Islamic texts in general. Ultimately, one can collect and work with a variety of general data based on large quantities of details and many analyses of these data on the different levels mentioned.

To summarise, the levels are as follows:

- 1) a variant tradition; for example, a saying about the medical practice of bleeding
- 2) a group of variants with roughly the same text
- 3) a group of variants with the same subject; for example, about medical practices
- 4) a group of variants of related subjects; for example, about medicine in general
- 5) early Islamic traditions in general
- 6) early Islamic texts in general

This process of examining and comparing texts on each of these levels involves viewing each text as determined by the fact that it is a unique variant or text, but also by the fact that it is a member of a group of variant traditions, which in turn is a member of a group of variants with the same subject, which is a member of a group of related subjects, which is a member of the group "Tradition literature", which is a member of early Islamic literature. A single text has certain characteristics of each of these groups. At each level we can establish characteristics, compare the texts and try to determine what elements were developed by whom, where, when, why and how: in other words, we attempt to date traditions on all of these levels.

Returning to the problems raised by the dating of early Islamic traditions, we find that the fourth problem is found within the dating process itself. Traditions have multiple aspects and there are numerous uncertain elements of information, apart from the existing lack of general knowledge about early Islamic society and transmission circumstances. If extensive descriptions are needed, as stated in the discussion of the previous problem, data should be readily available for use in the necessary complex analyses. This, however, is not the case.

A practical and simple example of misleading information involves the dates of birth and death of the transmitters in the biographical dictionaries. When we look, for example, at the entry for 'Ibn Shihab Al-Zuhriyy' in the biographical dictionary the *Tahdhib al-Tahdhib* of Ibn Hajar, we find that he was born in Madina in the year 50 or 51 or 56 or 58 and died in the year 123 or 124 or 125 or 126. For the analysis of a chain of transmitters it is

important to know when the alleged transmitters are supposed to have lived and what their age was. Combining all possibilities, we find that the age of this transmitter might have been 65, 66, 67, 68, 69, 70, 72, 73, 74, 75 or 76.⁹

A more complex example is found with historically unreliable data, for example persons named in the text and in the biographical dictionaries. The mention of these persons could lead us to assume that the name appearing in a text belongs to a particular transmitter and to conclude that he transmitted this story. Most of the information, however, is historically unreliable and there is as yet no criterion to help us decide whether a certain transmitter did or did not transmit a text, or even that he actually ever existed.

To solve this problem, a system has been designed that allows the researcher to give a label to all the data in the analyses and in the database used in this project indicating both the origin of the piece of data, or information entity, and its status with regard to its historical reliability.

When this is applied to the occurrence of a transmitter in a tradition, the above-mentioned 'Ibn Shihab Al-Zuhriyy', when found in a tradition, will be treated as a name, not a person, with the label 'from a text'. When found in a biographical dictionary the same name will get a label "from biographical dictionary 'x'". During the analysis the question will have to be answered whether the names in both sources are supposed to refer to the same 'entity' and only then, whether these names refer to a person that actually existed, and to what degree of certainty this can be presumed. These identifications of names from different sources with specific historical persons will all be made explicitly.

Of course, this does not solve the problem of historical unreliability, but it enables us to use data irrespective of their claim to historicity, and even to experiment with different probabilities of this claim. An advantage is that each researcher's decision regarding the historicity of a text element is documented in an explicitly made change of status of the same information entity. Thus, the same information entity can be retained in the research irrespective of its previous status or its label of historical reliability. The researcher can choose, for example, to work only with data of which he is more than 75% certain or with all data which he declared to be factual after a specified date.

⁹ Ibn Hajar, *Tahdhib al-Tahdhib*, Hyderabad, 1325-1327 A.H., 12 vols.: vol. 9, p. 445, no. 732; his full name is: muhammad bn muslim bn 'ubayd allah bn 'abd allah bn shihab bn 'abd allah bn al-harith bn zuhra bn kilab bn marra al-qurashiiyy al-zuhriyy al-faqih abuu bakr al-hafiz al-madaniyy.

To summarise, statements and implicit assumptions about information can be transformed into links explicitly laid between information entities. Therefore, with increased subtlety and more potential for checking and cross-checking than ever before, all data can now be used, regardless of possible historical unreliability.

4. Description and Analysis: Practice

After this discussion of problems and possible solutions in the analysis of traditions, we now investigate the practical aspects of the suggested approach. Translating the proposed solutions into requirements for a practical analysis, computer programs and a database structure, the following list of directives emerges:

- tradition texts must be recorded alongside their characteristics
- the biographical information about the transmitters must be recorded
- the information about the collections must be recorded
- it must be possible to link all of this information
- it must be possible to attach the above-mentioned labels of origin and reliability to all information entities (whether words or characteristics)
- it must be possible to make profiles of the characteristics of sets of texts, biographical entries and collections.

With these demands in mind, the following files have been made and prepared for entry into the database.

- 1) Approximately 2750 Arabic texts have been typed into the computer in transliteration. In order to be able to encode the historical, literary and normative information, a flexible system of markers and codes was developed, usable not only for words, but also for phrases, including discontinuous phrases, structures and whole texts. With this system it is possible to find the position of a word or phrase as well as its relative position, i.e. before, after or between certain other positions in the text.
The encoding records such characteristics as key words, persons, objects, practices, literary *topoi*, narrative elements and structures, historical and legal elements, text genre, Qur'an citations, transmitters and their relations in the different transmission chains.
- 2) For the transmission chain in the text, a separate file has been made in which the transmission relations in the chain are formalised into a succession of pairs of transmitters, to which position numbers and codes for chain composition and chain length are added. A full discussion of

these chain relations will follow below.

- 3) A third file contains biographical information on all of the transmitters, involving circa 1600 different transmitters: it presents the full name, date and place of birth and death. This file is structured and encoded in the same way as the text file.
- 4) In a fourth file, the persons in the texts and in the biographical dictionaries are linked to each other.
- 5) Finally, the information about the collections is entered into the database with the place, period, genre and legal school, along with the titles of chapters and paragraphs (which may contain useful information about the use of a text).

After entering these data into the relational database program Ingres, it is possible to ask complex questions involving a variety of characteristics of traditions. An example of such a question is:

What is the average age of the transmitters in chains of texts

- where the subject is 'wine'
- where the text type is 'legal discussion'
- where the transmitters lived in 'Mecca'
- where the collections were made after 'the year 250 after the Prophet'.

Naturally, the results of such queries have to be analysed; here the above-mentioned levels will play a crucial role. In this final stage, help from statisticians is needed, for with dependent texts, which variants in groups are, one can not simply add the data from any tradition whatsoever.

The profiles of traditions, subjects, places and transmitters, which we consider a valuable contribution to the debate about early Islamic traditions, can now be made. An important question concerns what correlations can be found between items of information, or combinations of these, taken from the textual part and items of information taken from the chain, and vice versa.

5. Stemmatological Problems

The comparison of variants and transmission chains has thus far only been treated in a superficial way within the context of the whole project. We proceed to describe more thoroughly those phenomena which are most interest-

ing to those interested in stemmas.¹⁰

In order to be able to analyse a large number of chains and make use of their information, we must enter the chains and their elements into the database in such a way that they can be retrieved by a search program. Since the text was composed for reasons other than entry into a database, it has to be transformed into a more formalised representation. A chain is a normal part of the text and it contains a diversity of names, verbs referring to the transmission of texts, other comments and variants. Furthermore, the actual order of the names of the transmitters in the text chain is not always as chronological as the words intend to depict. This problem of storing chains in a database must be solved, however, since we are dealing with numerous chains. In this project, at least 2700 chains have to be studied, since more than 2700 texts are analysed and each text has at least one chain of transmitters.

As we recall from the introduction, a typical tradition looks like this:

The tradition collector Bukhariyy heard from the transmitter Hisham who passed the story on from 'Abd-Allah, who transmitted it from Husayn, from Abu Sa'id, from Yahya, from Yusuf, from 'A'isha, the wife of the Prophet, that the Prophet, upon returning home, saw a curtain with images in the house and said: "Angels do not enter a house in which there are images". 'A'isha then removed the curtain.

For the purposes of chain analysis, we distinguish the following elements: a tradition collector, several transmitters, transmission between them and the story told by the last transmitter in the chain before the Prophet, in this case his wife 'A'isha.

A 'transmitter' can be defined as a person occurring in a chain, who is supposed to have received, then passed on a story to another person, i.e. another 'transmitter'.

'Transmission terms' are the terms that refer to the transmission relation between two or more transmitters, for example, 'he reported to' (Ar.: *axbara*). Apparently, there is also a beginning, the Prophet or the first Transmitter, and an end, the Collector who records the textual part and the chain in his collection.

Using only these elements, the chain looks like this:

¹⁰ I would like to thank E. Wattel of the Faculty of Mathematics and M. Dekker of the Faculty of Arts for introducing me to a new way of thinking about transmission chains.

Collector - transmission term - Transmitter - transmission term - Transmitter - transmission term - Transmitter - transmission term - Transmitter - transmission term - Transmitter - transmission term (saw) Prophet

Using abbreviated symbols and chronological order, a chain looks like this (Col = collector, trm = transmission term, [A-Z] = transmitter):

Prophet - trm - B - trm - C - trm - D - trm - E - trm - F - trm - G - trm - Col¹¹

When we consider the practical meaning of the words in a chain of transmitters, they seem to refer to the following sequence of events: Person A saw the Prophet do something; Person A then passed this on to Person B, who thus heard it from A; Person B then passed the tradition on to person C, who heard it from B.

Each transmitter appears to have two functions: as a witness he hears something from a spokesman; he, in turn, becomes a spokesman himself and passes it on to the next witness. Each person occurs in two relations, one backwards, and one forwards; he is receiver and sender. A period of time, however small, separates the two functions or relations, for when any transmitter B hears something from A, he only reports this to C at a later moment. If B would report it to C at the exact same time, C would have to be present at the transmission moment from A to B and he would probably claim to have it heard from A himself.

A whole chain could, consequently, be described as a series of relations

¹¹ One could also number a chain in reverse: Col - trm - B - trm - C - trm - D - trm - E - trm - F - trm - G - saw Prophet, which is the order of transmitters as we find it in the text. The starting point differs: the Prophet or the Collector. For questions concerning the position number of a transmitter in a chain, it is probably safest to start counting at the Prophet, so that a query using these numbers will arrive at people in roughly the same layer (for example: what is the third and fourth transmitter in all 'images' traditions and compare these to the third and fourth transmitter in all 'usury' traditions).

Defective chains, in which the Prophet and sometimes the first transmitter are not present, pose a problem. A person in such a chain will have a position number that is one or two positions lower than his position in a complete chain. With a query as mentioned above, this person could disappear from a selection of persons occurring in the third to fifth position.

We encounter the same problem when a chain is numbered in reverse. The collections date from periods between ca. 775 and 925 A.D. and some even later. Counting from these collections back to the Prophet would give a much broader variety of transmitters found in a specific position number than is the case with the first way of numbering. In this project, therefore, transmitters will be numbered from the Prophet upwards to the collections. A higher position number is later in time.

between at least two transmitters, from a beginning point, usually the Prophet, to an end point, usually the Collector (see Figs. 1 and 2).

In chains (and in 'chain bundles', see below), the transmitters are represented by a dot or point and their relations by a line. In combination with the dates of death of the transmitters, the line suggests the lapsing of time and it seems that with these dates of death the length of this line in historical time can be calculated. The length of a line between for example, A, who died in 80 A.H., and B, who died in 120 A.H., would be 40 years: the difference between both dates of death.

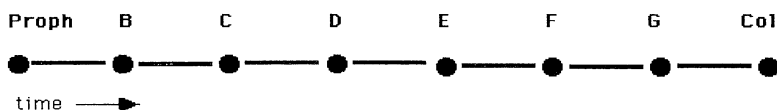


Fig. 1. A text chain as a series of relations between people, as worded in a text

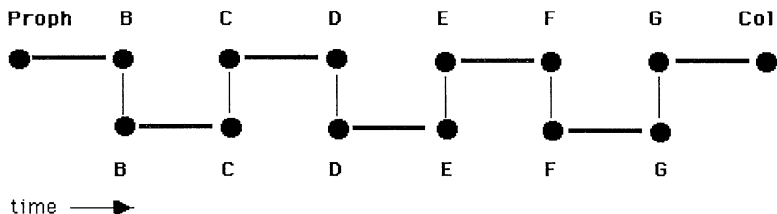


Fig. 2. A chain as a series of relations between two people, differentiating between the transmitter as receiver and as sender

The date of death of a transmitter is an important piece of information in the analysis of a chain (and of a bundle, see below), especially when we attempt to assign a text to a specific time and place. Already in the early centuries of Islam, these dates were important for Muslims for determining the value and authenticity of a tradition, since an important criterion was that all transmitters had to have had the opportunity, in terms of both place and time, to transmit a text received from the previous transmitter. Analysing a chain with these aims, we would say that when the difference between the dates of death of A and B is too large, it is unlikely that there was an actual transmission relation between A and B.

Looking back at the text, however, we see that in actual fact these 40

years between A and B are the length of the period that the text remained in the memory or scriptures of one transmitter A. The moment of transmission between A and B was possibly only five minutes long. Hence, it would be more precise to represent the moment of transmission, the meeting, with a point, rather than with a line, and the period that the text remains with the transmitter, as a line, rather than with a point.

From the viewpoint of graph theory, one could say that to both persons and meetings the characteristic of 'historical time' can be attached. The points indicate persons who have characteristics like dates of birth and death. The lines indicate what the two persons have in common, in this case the meeting. For the line the characteristic 'historical time' is unknown, and it therefore indicates the course of time during which the meeting has taken place. In the practice of studying chains of early Islamic traditions, the line refers to both the transmission relation, from the viewpoint of the history of the transmission of the text, as well as to the historical course of time during which the actual meeting has taken place.

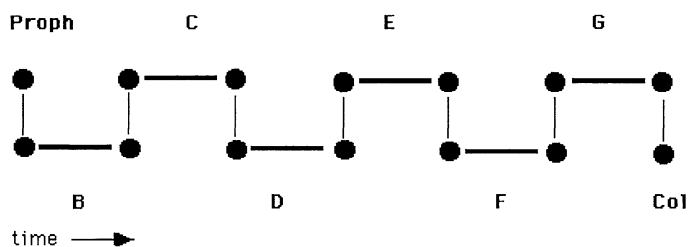
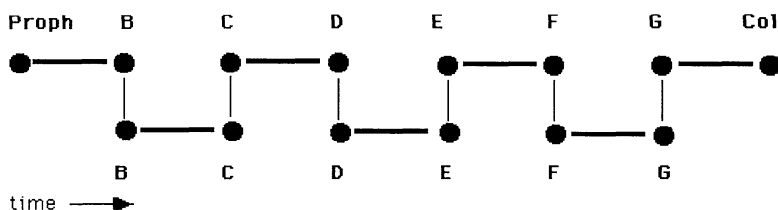


Fig. 3. A chain as a series of relations between two people, differentiating between the transmitter as receiver and as sender and representing the moment of transmission as a point and the transmitter as a line

When we return to the first definition of the chain using pairs of transmitters, represented by points joined by lines, we can redefine the chain as an abbreviation of the transmission process as it actually took place. By describing a chain in terms of its relations, the transmission information in the text becomes explicit, including the two different functions of each transmitter.

For a research project using a database, the problem seems to be solved and the relations between the people can now be entered in a simple way; when the beginning and end points are also indicated, all information is readily available (see Fig. 4).

I



II

Proph	→	B
B	→	C
C	→	D
D	→	E
E	→	F
F	→	G
G	→	Col

III

Proph, beginning	→	B,	middle
B, middle	→	C,	middle
C, middle	→	D,	middle
D, middle	→	E,	middle
E, middle	→	F,	middle
F, middle	→	G,	middle
G, middle	→	Col,	end

Fig. 4. A chain as a series of relations between two people, differentiating between the transmitter as receiver and as sender (I), represented by a table of pairs of transmitters (II) and by the same with supplementary information (III)

As we can see, there are as many relations as there are transmission terms, and there is usually one more transmitter than there are transmission terms, i.e., the number of relations is usually the number of transmitters minus one. The beginning, middle and end of the chain are recorded. Using these basic pairs of transmitters we can describe all chains and search for successions of pairs within a chain.

There is, however, a slight problem. Until now we have discussed straight lines, but there are also complex chains, in which a transmitter reports a text to more than one other transmitter [$A \rightarrow B_1$, $A \rightarrow B_2$]. This is also, by definition, valid for bundles of chains.

A 'chain bundle' is a set of two or more chains which share the subject of the textual part in a particular type of text; the texts of such a bundle are considered to be variants.¹² A bundle is examined to gain insight into the origin and development of the texts. The study of bundles of chains and the

¹² Compare a critical discussion of this definition, below.

stemmatological phenomena of bundles were virtually all first recognised by Juynboll, the one specialist in this field to devote many articles to this aspect of research into early Islamic traditions.¹³ His method, based on an earlier finding of Schacht and known as the “common link theory” will here be explained and critically discussed.

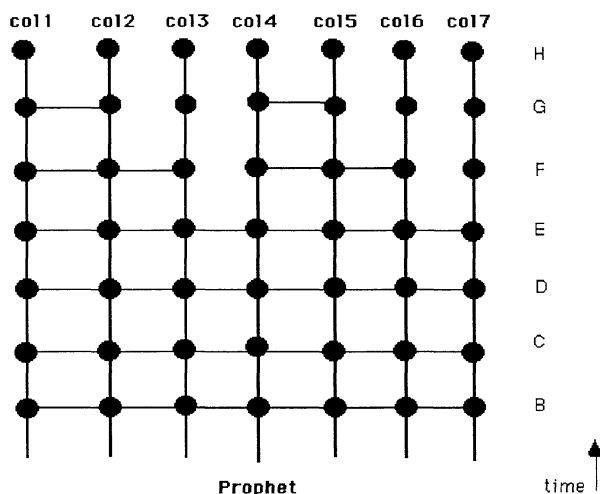


Fig. 5. *Making a bundle, phase one: a number of chains are combined into a bundle; names in common are recognised (here connected with horizontal lines)*

The procedure for converting a selection of chains into a bundle is as follows: the separate chains of a selection are compared and if they have a certain transmitter in common, i.e., when the names of transmitters seem to refer to the same person, the identical names are assimilated, leaving only one occurrence of the name (Figs. 5 and 6). By representing several occurrences of a name by one, to which all initially separate transmission lines are attached, these separate lines can form a tree structure, giving a good representation of the relations between the transmitters of a specific text, at least as presented to us in the sources (Fig. 7).

¹³ For a selected bibliography of Juynboll, see References below; Schacht 1950:171-175.

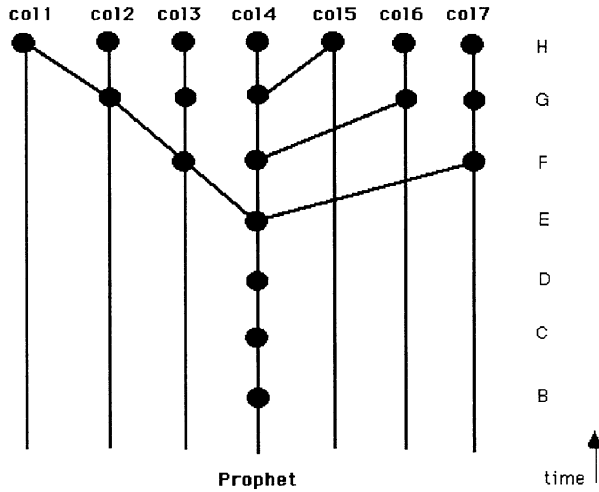


Fig. 6. Making a bundle, phase two: names in common are assimilated

When making these bundles, Juynboll rediscovered an interesting phenomenon:

Contrary to what we might have expected, for a great number of traditions this strand [from the Prophet upwards in the direction of the Collector] is a single strand for the first three, four or five names.¹⁴

In other words, often chains have the first few names after the Prophet in common and only thereafter do transmitters start having transmission relations with more than one transmitter. Sources about this period report a lively transmission activity. Therefore, from the Prophet onwards each relation upwards in time should have been reported as a 1:n relation and not a 1:1 relation as they have been in the majority of cases. It is more probable that the last transmitter of these 1:1 relations (E in Fig. 7) had been active in originating and distributing this tradition. Furthermore, it is most likely that these nodes, called common links, were also responsible for the one-to-one chain from the last transmitter downwards to the Prophet. The main motivation for this process would be that at the beginning of the second Islamic century transmitters were increasingly required to have their

¹⁴ Juynboll 1991:172; see Schacht above.

traditions authenticated by a chain reaching all the way down to the Prophet. Put differently: this common link probably composed the text, related it to his pupils, also reporting that he had heard it from one man, who had heard it from one man, who had heard it from the Prophet.¹⁵

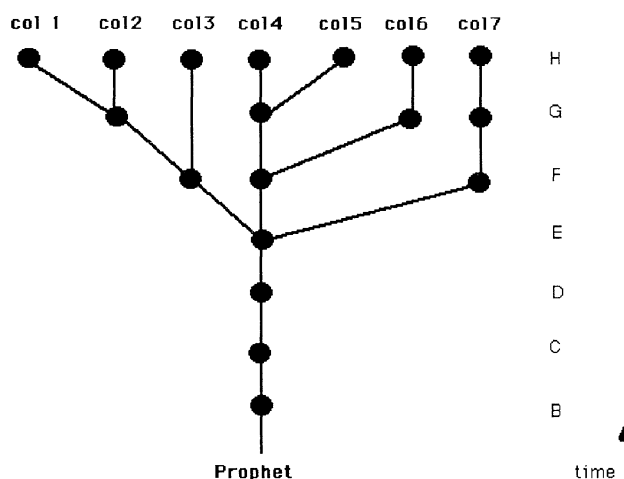


Fig. 7. *Making a bundle, phase three: an example of a simple bundle of chains; a summary of seven separate lines; E is the 'common link'*

The designation of a single node in a bundle or tree as likely to have been active in the transmission of the text is not straightforward. At this point we move beyond the stage of the graphical representation of the chains in a bundle to the stage of interpreting a bundle as a tool for historical analysis. The activities of the nodes can vary from inventing the whole text to changing a word, and from lying about all earlier colleagues to only changing one name in the chain. Juynboll states,

identifying the cl [common link] in a bundle is of crucial importance, in that it points to the originator of the tradition supported by that bundle as well as to the method that originator used to get the saying authenticated.¹⁶

Attributing one of these roles to a transmitter involves a careful use of such

¹⁵ Cf. Juynboll 1992; The name 'common link' is confusing, since it is not the link that is common to several chains, but the transmitter. The 'node theory' would have been more appropriate.

¹⁶ Juynboll 1991:172; the text between brackets is mine.

data as the biographical information about the transmitters, the characteristics of the text, the information from Islamic tradition critical works and the role of the transmitter in other bundles with regard to the contents of the tradition and the mode of its transmission.

Without this knowledge, the layman is confronted with considerable difficulties in applying the common link theory, some of which will be discussed below. Since making a chain bundle and using the common link theory can be important tools for the study of early Islamic traditions, we will describe its most important stemmatological phenomena in terms of relations between two transmitters. This will enable us to find and analyse bundles on a large scale within a database.

It appears that most of the stemmatological phenomena can be described as specific kinds of activities by nodes: transmitters with more than one relation in a certain direction in historical time. In other words, when a transmitter has more than one relation with transmitters from whom or to whom he transmits a text, he is a node. According to the common link theory, he is, consequently, also a candidate for a historically accepted active role in transmitting the text.¹⁷

The different types of nodes (Fig. 8) and two other phenomena will here be presented with a definition using pairs of transmitters and their transmission relations:

- 1) a 'common link' (c.l.): the last transmitter shared by the majority of the chains of a bundle, having two or more transmitters, pupils, transmitting from him. [$D = \text{c.l. in: } A \rightarrow B, B \rightarrow C, C \rightarrow D, D \rightarrow E_1, D \rightarrow E_2, D \rightarrow E_3 \dots$]
- 2) a 'partial common link' (p.c.l.): a common link which occurs later in a bundle than the "first" common link. Usually the partial common link has fewer pupils and is often the pupil of the common link or of another partial common link. [$E_3 \rightarrow F_1, E_3 \rightarrow F_2, E_3 \rightarrow F_3$]
- 3) an 'inverted common link' (i.c.l.): a transmitter having two or more transmitters, or teachers, from which he transmits. This is the opposite of 1) as the direction of the multiple relations is reversed: in this case the multiple transmitters are earlier in the chain than the one common link. [$C_1 \rightarrow D, C_2 \rightarrow D, C_3 \rightarrow D$]

¹⁷ Cf. especially Juynboll 1994:1.

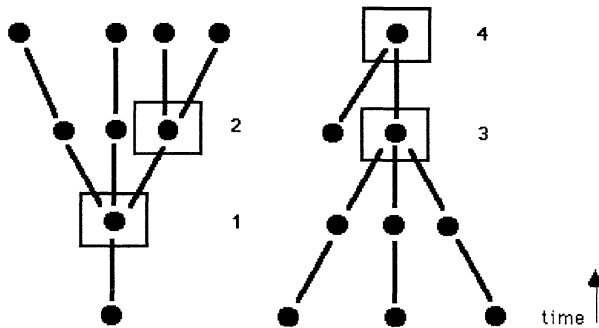


Fig. 8. A schematic representation of four types of nodes in a bundle (the nodes are found in the squares)

- 4) an 'inverted partial common link' (i.p.c.l.): a transmitter having two or more transmitters, or teachers, from whom he transmits. This is the opposite of 2) as the direction of the multiple relations is reversed: in this case the multiple transmitters are earlier in the chain than the one partial common link. [$F1 \rightarrow G$, $F2 \rightarrow G$, $F3 \rightarrow G$]
- 5) a 'single strand': a chain in a bundle in which most transmitters only have a transmission relation with one other transmitter, i.e., when making the bundle, the 'single strand' had almost no names in common with other chains in the bundle.
- 6) 'spider': a bundle that "consists exclusively of single strands [cf. 5], even those showing up what looks like a cl [common link]"; [$A1 \rightarrow B1$, $B1 \rightarrow C1$, $C1 \rightarrow D1$, $D1 \rightarrow E1$; $A1 \rightarrow B2$, $B2 \rightarrow C2$, $C2 \rightarrow D2$, $D2 \rightarrow E2$; $A1 \rightarrow B3$, $B3 \rightarrow C3$, $C3 \rightarrow D3$, $D3 \rightarrow E3$].¹⁸

It appears from these definitions that all of these phenomena can be described using pairs of transmitters and their relations (in a graph: two points connected by a line).

Apart from these phenomena from the common link theory, more factors must be taken into consideration when chains and bundles of chains are

¹⁸ For the phenomena 5) and 6), see Juynboll 1994:2:184 and note 32, p.192: "single strands forming a spidery bundle do not constitute a basis for positing the historicity of the transmission point of its cl [common link] if there is one ; he is in fact no more than a 'seeming' cl. As stated earlier, a cl's position becomes historically feasible only when a number of his alleged pupils are themselves pcls [partial common links] characterized by two or more alleged pupils."

entered into the computer.

A can transmit to two people and these two people (B1, B2) could both transmit to C. However, sometimes in both chains and bundles, it can occur that A transmits to B and B to C, but that, at the same time, A also transmits directly to C [$A \rightarrow B$, $B \rightarrow C$ and $A \rightarrow C$].

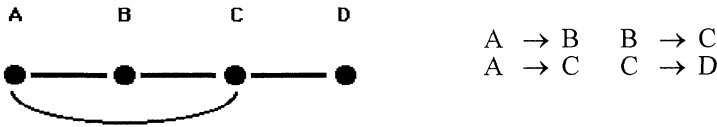


Fig. 9. The phenomenon [$A \rightarrow B$, $B \rightarrow C$ and $A \rightarrow C$] in a diagram and as a table of pairs of transmitters

This phenomenon involves an interesting aspect of the description and recording of the transmission of traditions, i.e., the length of the transmission line between A and B. We saw above that the characteristic of historical time can be attached to both persons and meetings. Persons have dates of birth and death and we could regard the line as referring to both their transmission relation from the viewpoint of the history of the transmission of the text, as well as to the historical course of time during which the actual meeting has taken place. The date of the actual meeting must, of course, be between the date of birth and death of both persons.

In the case of this phenomenon [$A \rightarrow B$, $B \rightarrow C$ and $A \rightarrow C$], A has two transmission relations: one which covers, for example, 30 years and a second one which covers, for example, 60 years. As we can see in Fig. 9, it is again possible to record and retrieve this in terms of pairs of transmitters. We can ask for all transmitters to whom A has had a relation (yielding B and C), or all transmitters from whom C has transmitted texts (yielding A and B). Counting the number of transmitters or layers in the transmission of a text presents difficulties, however. [$A \rightarrow B$, $B \rightarrow C$, $C \rightarrow D$] involves three steps while [$A \rightarrow C$, $C \rightarrow D$] involves two, in which the layers of A and D remain the same. Of course, the moments at which [$B \rightarrow C$] and [$A \rightarrow C$] occur can also differ in time and place. The only unambiguous fact from such transmission lines is that the following relation [$C \rightarrow D$] is a single one and that it must have occurred before the death of C.

Until now, we have focussed on the actual chains and the use Western scholars make of them to construct chain bundles. It is possible to present

an abstraction of these in diagrams which take into account the following nine phenomena as represented in Fig. 10.

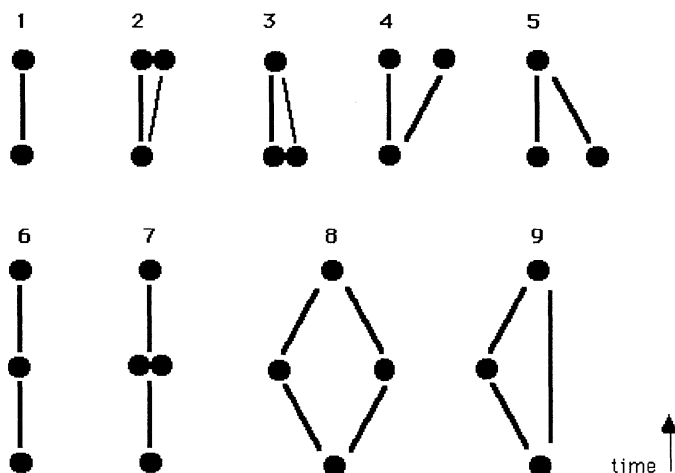


Fig. 10. Phenomena in chains and bundles of chains

- 1) one transmitter to one transmitter [$A \rightarrow B$]. The transmission can be direct when transmitters are at the same place; transmission can be indirect via written permission, in which case transmitters are not at the same place.
- 2) one transmitter to two transmitters [$A \rightarrow B_1, A \rightarrow B_2$]. The act of transmission occurs for each relation at the same time and place.
- 3) two transmitters to one transmitter [$A_1 \rightarrow B, A_2 \rightarrow B$]. The act of transmission occurs for each relation at the same time and place.
- 4) one transmitter to two transmitters [$A \rightarrow B_1, A \rightarrow B_2$]. The act of transmission occurs for each relation at a different time and place.
- 5) two transmitters to one transmitter [$A_1 \rightarrow B, A_2 \rightarrow B$]. The act of transmission occurs for each relation at a different time and place.
- 6) a succession of single transmission relations [$A \rightarrow B, B \rightarrow C$]. The transmitter at the beginning point is a sender only, the transmitter at the end point is a receiver only; the intermediary transmitters are both receivers and senders.
- 7) a succession of transmission relations [$A \rightarrow B_1, A \rightarrow B_2, B_1 \rightarrow C, B_2 \rightarrow C$]. A combination of 2) and 3), where the receivers in 2) are the senders in 3). The acts of transmission for both intermediate transmitters occur at the same time and place.

- 8) a succession of transmission relations [$A \rightarrow B1$, $A \rightarrow B2$, $B1 \rightarrow C$, $B2 \rightarrow C$]. A combination of 4) and 5), where the receivers in 4) are the senders in 5). The acts of transmission for both intermediate transmitters occur at a different time and place.
- 9) a succession of transmission relations [$A \rightarrow B1$, $B2 \rightarrow C$, $A \rightarrow C$]. A combination of 1) and 6), where the sender of 1 is identical to the first sender in 6) and the receiver in 1) is identical to the last receiver in 6).

The recurring elements and aspects in the description of these phenomena are the following:

- transmitter; with the functions of sender and receiver and, in historical time, beginning and end of a relation
- relation between transmitters, with two possible directions in time: from earlier to later or vice versa
- succession of relations between transmitters
- time of transmission
- place of transmission
- length of a transmission relation in historical time or in terms of its relative length compared to another or several other transmission relations (see Fig. 9).

We have been able to describe all phenomena in the transmission of early Islamic traditions using pairs of transmitters and their relations. Through this process, the basic elements and aspects of these phenomena have come to light. It appeared possible to record these pairs and put them into a database and retrieve them. As stated above, it is not difficult to link to the people in the chains the information about the transmitters found in the biographical dictionaries. Using this and the fact that the people in the text are linked with the information about the text, we can now ask, for example, for all chains for which is valid for a pair of transmitters that transmitter C is 'Abd Allah and transmitter D is Husayn. A further condition can be to restrict the search to texts of a particular type, or about a particular subject and to texts copied from a collection with particular characteristics.

Useful as this may be, it must be remembered that the answer to such a query would only show that these transmitters have a particular transmission relation in a certain text and that this relation is part of a succession of relations. It is not clear from these pairs and successions of pairs when B actually spoke to C (time of transmission) or where (place of transmission). We only know that it must have been after the birth of B and before the death of C. Conclusions about the length of the transmission lines in terms of historical time can only result from an analysis of the chain.

Other problems arise from the interpretation of the stemmatological phenomena. For example, opinions about the workability of the common link theory differ greatly and many scholars refrain from using it, partly for the following reasons.

The first and most general difficulty involves the question “what is the text?”. The text we have in our collections has been transmitted for more than a millennium by transmitters, collectors and copyists down to, finally, the editors of the modern text editions. Which phase of the text are we dating? Which aspect? The wordings? The basic message? In terms of dating, should we differentiate between the originator of the text in all of these phases?

A related problem deals with the selection of texts. Earlier, we defined a bundle as a set of two or more chains which share the subject of the textual part in a particular type of text. A subject, for example, ‘drinking wine’, however, often has a number of aspects which are expressed in sayings, stories, legal maxims, etcetera. One aspect is sometimes found only in one text type, but often there exists a narrative version of, for example, a legal maxim, or the legal maxim could be uttered by one of the actors of a narrative. Often completely different subjects are found in the same enumeration, as, for example, in the following text: the Prophet forbade bleeding someone and paying for bleeding, tattooing and getting tattooed, paying for dogs and paying for images.

When texts are selected for a bundle, should we include both a legal maxim and a narrative in the same bundle? Should an enumeration be accepted in a bundle of texts dealing exclusively with tattooing? If the same enumeration of subjects was also part of a larger narrative which only deals with paying for dogs, would that make any difference? Can we include it in both bundles?

It is clear that a more precise criterion has to be found for the selection of texts than Schacht’s “the chains of a given tradition”.¹⁹

For both smaller and larger selections, arguments can be presented. In otherwise different texts, a bundle of chains selected purely on the basis of a small text detail can reveal a definite common link for this detail (cf. Reenen 1990:65ff).

We could argue, on the other hand, that it is only logical to include, for example, all 350 chains from the texts about the “prohibition of making

¹⁹ Schacht (1950:171ff) speaks of “the isnads [i.e. chains] of a given tradition” but he also says “...when the isnads of different, but closely connected traditions showed a common link”.

images" into one bundle. Many bundles based on this subject share common links and partial common links. It would become visually difficult to analyse such a bundle, but it is a good methodological exercise to find reasons for not doing so. At any rate, the selection criteria for chains of a bundle should not depend on the number of chains that can be found for a bundle or on whether a common link and originator can be found (cf. Reenen 1990:68ff).

A second point which concerns the absence or presence of chains in a bundle is the possibility that not all variant texts of a saying or story are available. Many collections are still difficult to use, the indexes contain mistakes and the chapter numbers differ in various editions. An added factor is that at the end of the early Islamic period, stricter criteria for the acceptance of chains had been formulated and many texts were sifted out by the collectors. These texts might have constituted interesting variants and branches to a tree. Some of the phenomena which are visible in a tree might only have been reinforced, but others will have disappeared forever since its chains have been deleted. In more general terms we must ask the following question: how can we draw conclusions from a bundle that is almost by definition incomplete?

When we, furthermore, recognise that whole chains were fabricated, that people were invented and that the transmission of a certain text was sometimes forced upon a transmitter by his later colleagues, we must realise that drawing a bundle is only a small step towards its analysis.

Another problem arising from the interpretation of the information in a text chain discussed above will now be rephrased. The problem of "nominal record linkage" involves the identification of the name of a person (transmitters) in a source with a name in another source, and the identification of these names with a historical person (cf. Morton 1994).

The procedure of making a chain bundle was described earlier: all but one of the names of transmitters which seem to refer to the same person are omitted. By representing several occurrences of a name by one point in a bundle, the separate lines can form a tree structure. This identification, however, may be obscure, especially since many names are ambiguous. First of all, most of the Arabic names of transmitters consist of a forename, name of the father or mother (and often many more generations) and several other names, indicating his place of birth, the legal school he belonged to, his profession or some physical characteristic. A person can often be referred to by a number of these or a combination thereof. Finding the man or woman in question, i.e., linking the text name with a biographical name, is a task that requires meticulous care.

There is a further danger in assuming that names of transmitters consistently refer to the same person who frequently appears as a transmitter in a chain. If "Ibn Shihab al-Zuhriyy" and "Al-Zuhriyy" are mentioned in the same position in a bundle, many readers of the bundle will conclude that the names refer to the same person; however, there are many other people called "Al-Zuhriyy" who were contemporary to the famous one at the beginning of the second Islamic century. It is a distinct possibility that a later transmitter deliberately referred to his less famous teacher by a shortened name in order to encourage an association with the famous "Al-Zuhriyy", even if this were not the case.

The lack of reliability of the data also affects such seemingly simple data as the ages and dates of birth and death of the transmitters in the biographical dictionaries. We have already seen that these are subject to a large degree of variableness; the same applies to the places where a certain transmitter is reported to have heard or passed on traditions. The geographical information is essential since locating the texts is one of the tasks involved in dating them. Reports from the early Islamic period tell us that transmitters would travel from one place to the other in search of traditions (a practice called *talab al-'ilm*). Usually, therefore, each transmitter has two or three toponyms associated with him. When we combine the places of transmission for all transmitters in a bundle to determine the place of origin of a text, we find a multitude of possible successions of places. Perhaps we should only consider places connected to the common links or partial common links, but then the procedure would be circular since precisely this information is needed to determine the common link and his role in the transmission of a text. Therefore, also with data concerning the place of origin of a transmitter, extreme caution must be used, even more so when several of these pieces of information are combined in a complex analysis.

Leaving the problems with the data aside, we proceed to the interpretation of bundles. The decision to take the graphical node as the historical originator, be it as common link or only as partial common link, is not a simple one. Examples of this can be seen in the attempts to detect the measures taken by transmitters to claim the transmission of a text for themselves by creating a private chain or part of a chain. The result of this practice can be a perfectly regular tree, in which all traces of the cover-up have been erased.

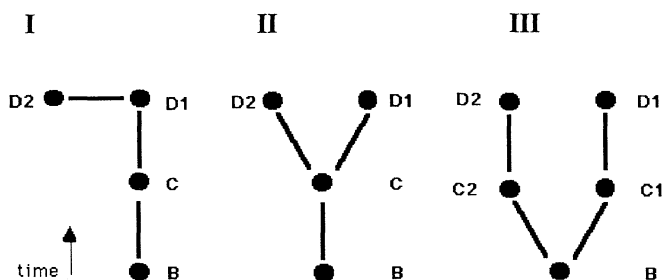


Fig. 11. The "Cook-principle": D2 sees D1 has a line with C (I) and wants his own line with C (II) or a contemporary of C (III)

A small scale example of this is a phenomenon that has come to be called the 'Cook principle', after its originator, M. Cook. The situation is as follows (see Fig. 11): Transmitter D2 hears that C has transmitted a text to his contemporary D1. D2 would have liked to get the credit for the transmission himself. Instead of referring to D1, he also claims to have heard the text from C, or even, from another teacher than C. In the first option C, originally in a 1:1 relation to D1, suddenly becomes the common link; the result would be [C→D1, C→D2] (Fig.11, II). In the second option, B, originally in a 1:1 relation to C (C1), suddenly becomes the common link; the result would be [C1→D1, C2→D2] (Fig.11, III). This would particularly be the case if a number of transmitters in D's generation would have done the same, in spite of the fact that D might have invented the text without ever having known C. If practised on a large scale, the C, common link due to this practice, is not distinguishable from the common link that really was the originator (here C or D). Juynboll recognises this and he states that some of the lines following a common link, or node, could be explained as a result of this mechanism. With large numbers of chains, this would require so much coincidence that one would be forced to start considering a conspiracy of some sort. Since such a theory lacks support, one of Juynboll's major adages remains, in spite of the Cook principle:

The more transmission lines there are, coming together in a certain transmitter, either reaching him or branching out from him, the more that moment of transmission, represented in what may be described as a "knot", has a claim to historicity.²⁰

²⁰ Cook 1981:Ch. 11; cf. also Juynboll 1989:352-354.

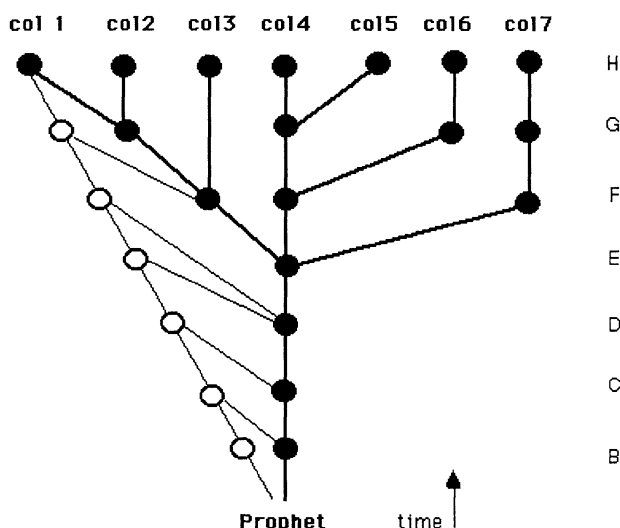


Fig. 12. 'Divers under the common link' (with dotted lines), who avoid having to rely on the common link (E), as so many others already did; they can join the main line at any point under the common link

Greater aspirations can be ascribed to those who “dive under the common link” (Fig. 12), a sort of expanded Cook principle, which is described by Juynboll as follows:

As from the early third/ninth century, tradition scholars had apparently begun to feel uneasy about the well-nigh miraculous, and for a lot of people perhaps no longer believable, key position of many cls [common links]. ... In an attempt to take away some of the “pressure” bearing down on a cl, we see how tradition scholars start interfering in isnad bundles [chain bundles] in a fundamental way: ... they begin to practise “diving under the cl” by creating new strands which bypass the cl and end up in transmitters given in the single strand [1:1] under the cl, at first a Successor [the second layer of transmitters after the Prophet], later also a Companion [the first layer of transmitters after the Prophet], and very rarely (?) even the Prophet himself.²¹

A theoretical danger, for which I have no examples, could emerge in the following situation: suppose the actual transmission took place via three quite ordinary chains [1, 2, 3] from the common link onwards and that

²¹ Juynboll 1994:158; text in square brackets is mine.

later two or more collectors took one of these chains [3] and attached a few of their own branches to it, one or two positions after the common link, let us say at position E. In an analysis of the ensuing bundle, we would conclude that 3E is the common link and that the lines 1 and 2 are loose strands and can therefore be ignored. The question is raised, again, of whether this practice and the resulting bundle could be distinguished from a bundle in which 3E really is the common link.

The basic assumption behind the common link theory is one that still awaits corroboration from research into transmission practices in the first centuries of Islam. The lack of insight into this period, already mentioned in connection with the question of oral versus written transmission, makes it impossible to prove that it is unlikely that one transmitter only told his text to one single other transmitter.

Earlier we saw that the main rule of analysis in the common link theory is that

The more transmission lines there are, coming together in a certain transmitter, either reaching him or branching out from him, the more that moment of transmission, represented in what may be described as a "knot", has a claim to historicity.

Conversely, when a branch in a bundle consists of 1:1 relations only, during analysis it will tend to be cut off like a wild shoot of a fruit tree. The first few names from the Prophet to the node are not considered to be historical, which is also the case for the 'single strand' from the common link upwards to the collections and the 'divers', avoiding the common link, from a low position in the tree to the collection. What remains is a solid little bush, in which each branch has side branches.

Probable as this may be, the possibility can not be excluded that some people did have 1:1 transmission relations, i.e., that they only told a text to one other transmitter. Furthermore, we will never be sure whether all chains have been found that have ever been attached to texts about the subject of a bundle. Concerning the 1:n relations, the opposite is true. As a result of practices like the 'Cook-principle' and 'diving', a common link or partial common link might have gained this status by completely invented chains or parts of chains. Furthermore, there was a general desire to authenticate a certain text by as many, possibly homemade, chains as possible.

Summarising, one could say that each chain that is present can be historical or invented but also that each chain that is absent can be historical or invented. With regard to the '1:n = historical rule', the possibility remains that each and every single 1:1 transmission relation is a historical one. The opposite remains to be proven.

A completely different point of interest is the assumption behind some of the phenomena described and used in the common link theory. Some phenomena lead us to assume that the transmitters and collectors were aware of the fact that a tradition had already been transmitted via other chains and transmitters. With the specialised knowledge that many transmitters and collectors must have acquired, this need not surprise us. In the traditions themselves, comments are interpolated that ascribe the words or a particular variant to a certain transmitter. Many of the critical terms, listed above, suggest this awareness, whereas some of the collectors also wrote books about tradition criticism.

If, however, a collector like Ahmad bn Hanbal uses one of his frequent 'loose strands' to establish a tradition with one of his own chains, which has nothing or little in common with those in the rest of the bundle, including the common link, then there is greater evidence of awareness of bundles than expected (cf. Reenen 1990:67). The question, therefore, is to what extent were the transmitters aware of the interrelations between the chains which we represent in a bundle? Our analytical assumptions must not be imposed on the transmitters in that period; on the other hand, we should not underestimate the transmitters when we see the complex ways in which they impose their needs for authentic texts and chains on the material.²²

To summarise, we can say that implementing this common link theory requires large amounts of data and many careful decisions. Juynboll has already shown that much information can be gained from chain bundle analysis and that many mechanisms in the transmission of early Islamic texts can be recognised. The previous discussion has also made clear, however, that when enough data are retrieved the basic rules of the common link theory must again be tested and many implicit assumptions must be made explicit.

For the near future, much work will continue to be done by hand. Theoretically, if we can make decisions on the basis of explicit criteria, we should be able to express them in rules and enter these into a computer program. With clear criteria and decision procedures, it should be possible to select, with a high degree of accuracy, common links who are originators.

Within this project and the database developed in it, however, more modest objectives are set. Only some of the tools for analysis are supplied that might help us to make decisions and formulate criteria, e.g., a list of all

²² An interesting clue to this awareness is discussed by Juynboll (1994:158); a "centre", "pivot" or "common link" seems to have been recognised by fellow researchers in former times.

the nodes, common links, partial common links, etcetera, of all 2700 texts of the corpus of this project, with their subjects. The type of texts that are ascribed to a certain transmitter and the type of text of certain collections are also being charted. Patterns of information in the biographical dictionaries will also be a result of the analysis, for example, dates of birth and death and average age in a city or period.

These findings can also be established without the help of a database, and sometimes even better; however, it is the speed with which this can be carried out, the number of texts that can be analysed, the number of conditions that can be added to a question and the connections that can be made between information from texts, biographical dictionaries and collections, that make such analyses so useful. Additionally, more freedom in analysis is offered by the possibility of slightly altering the queries and re-submitting them.

There is one thing, however, that can be done by anyone working with early Islamic traditions, with or without computers, a matter which has not become normal practice. In a more strict procedure than is often followed, four different stemmas can be made during the analysis of early Islamic traditions:

- 1) a stemma of the chains as they are recorded in the texts; this stemma gives a picture of the origin and development of a text as it is presented to us in the chains;
- 2) this stemma must be analysed, which will result in a stemma that represents a more critical view of this development;
- 3) a third stemma can be made which reflects the interdependencies of the text parts and their variant wordings;
- 4) a combination of 2) and 3) will lead to a stemma which is based on both parts of an early Islamic tradition.

Ideally 2) and 3) result in the same picture; I doubt whether this will often be the case in practice.

References

- Cook, M.A. 1981. *Early Muslim Dogma - A Source-Critical Study*. Cambridge.
 Creel, H.G. 1963. *Chinese Thought, From Confucius to Mao Tse-Tung*. Chicago.
 Gates, J. 1936. "Model Emperors of the Golden Age in Chinese Lore". *JAOS* 56. 51-76.
 Goldziher, I. 1890. *Muhammedanische Studien*, Vol. 2. Halle.
 Guillaume, A. 1924. *The Traditions of Islam - An Introduction to the Study of*

- Hadith Literature*. London.
- Ibn Hajar, n.d., *Tahdhib al-Tahdhib*, ed. 1325-1327 A.H., 12 vols. Hyderabad.
- Juynboll, G.H.A. 1982-1989a. "Hadiith". *Dictionary of the Middle Ages*, vol. 6, 45-49. New York.
- Juynboll, G.H.A. 1982-1989b. "Sunna". *Dictionary of the Middle Ages*, vol. 11, 510-513. New York.
- Juynboll, G.H.A. 1983. *Muslim Tradition, Studies in chronology, provenance and authorship of early Hadith*. Cambridge.
- Juynboll, G.H.A. 1989. "Some isnād-analytical methods illustrated on the basis of several woman-demeaning sayings from Hadith literature". *al-Qantara*, X, 343-384.
- Juynboll, G.H.A. 1991. "The Role of the Mu'ammārūn in the Early Development of the Isnād". *Wiener Zeitschrift für die Kunde des Morgenlandes*, 81. Band, 155-175. Wien.
- Juynboll, G.H.A. 1992. "Some Notes on Islam's first Fuqahā' distilled from early Hadith literature". *Arabica*, XXXIX, 287-314.
- Juynboll, G.H.A. 1994a. "Early Islamic Society as reflected in its use of Isnads". *Le Museon*, Tome 107, Fasc.1-2, 151-194.
- Juynboll, G.H.A. 1994b. "On the Origins of the Poetry in Muslim Tradition Literature". *Festschrift Ewald Wagner zum 65. Geburtstag*. (= *Beiruter Texte und Studien*, Bd. 54), 192-207. Beirut.
- Morton, G. 1994. "Presenting the Self: Record Linkage and Referring to Ordinary Historical Persons". *History and Computing*, vol. 6, no. 1, 12-20.
- Noth, A. 1973. *Quellenkritische Studien zu Themen, Formen und Tendenzen frühislamischer Geschichtsüberlieferung*. Teil I: *Themen und Formen*. Bonn.
- Raven, W. 1988. "Some Early Islamic Texts on the Negus of Abessynia". *Journal of Semitic Studies*, XXXIII, 2(Autumn), 209.
- Reenen, A.D. van 1990. "The Bilderverbot, a new survey". *Der Islam*, Band 67, Heft 1, 27-77.
- Robson, J. 1971. "Hadith". *The Encyclopaedia of Islam*, new edition (1960 - ...), Vol. III, 23-28. Leiden: E.J. Brill.
- Schacht, J. 1949. "A Reevaluation of Islamic Traditions". *JRAS*, 143-154.
- Schacht, J. 1950. *The Origins of Muhammadan Jurisprudence*. Oxford.
- Stetter, E. 1965. *Topoi und Schemata im Hadith*. Tübingen.
- Vansina, J. 1985. *Oral Tradition as History*. London.
- Wensinck, A.J. et al., ed. 1939-1988. *Concordance et Indices de la Tradition Musulmane*. Leiden.

PART II

Subject-Related Observations

Medieval Dossiers and Modern Stemmas

An Exploration of Manuscripts of the *Epistola Alexandri ad Aristotelem*

J.B. Voorbij

Department of Computer and Humanities
Utrecht University, NL

0. Introduction

The introduction to a text edition that contains a *stemma codicum* usually offers a pedigree of only that particular text.¹ The work in question is presented as an isolated phenomenon. This is, in fact, curious, since many medieval manuscripts contain at least two texts. The implication is that several texts may not have been transmitted individually, but rather in association with other works. There is a chance that a specific set of texts appears repeatedly, and that it was copied as a group from manuscript into manuscript. Only rarely does one find a stemma that displays the tradition of a combination of texts. Most editors take neither the transmission contexts of their texts nor the possible consequences of these contexts for the stemmas into consideration. It is disputable whether an editor always can take the transmission context into account, but he should be aware of the advantages this method offers.

¹ This contribution is the result of a long-term research project on medieval Latin Alexander texts, executed in collaboration with Martin de Ruiter (Department of Medieval Studies, University of Groningen); Aafje Lem, Mieke Lens, Bennie Werink and Marinus Woesthuis (all of the same Department) joined this cooperation later. I appreciate the many vivid discussions I have had with them and gratefully acknowledge the material they have always shared with me without reserve. In the process of writing this article I enjoyed the assistance and advice of Martin de Ruiter (University of Groningen), Jos Biemans (University of Amsterdam), Frans van Liere (University of Groningen) and especially Randy Tollefsen (Utrecht University).

In this contribution I will discuss the use of fixed combinations of texts to construct more reliable stemmas. My point of departure is a corpus of ca. 190 manuscripts, each of which contains at least one Latin text dealing with Alexander the Great. One of these texts, the *Epistola Alexandri ad Aristotelem* (hereafter: *Epistola*), assumes a prominent position within the corpus. It has been edited several times; its latest critical edition, made by Walther Boer (1953), shows serious defects. I will bring into question a crucial part of Walther Boer's *stemma codicum* and propose an alternative to it.

Next, the question will be approached of whether an editor, when constructing a stemma, may make deductions from information about repeatedly occurring combinations of his text with others. If used, the information must be consistent; the types of data on which this consistency depends will be described. I call a combination of texts dealing with the same subject matter a 'dossier'. Terminology related to this idea will be presented; this terminology discriminates between the manuscript as a physical copy of a text and the manuscript as a textual witness.

Further, based on the corpus of Alexander manuscripts, it will be shown that the identification of dossiers may provide insight into text transmission. Finally, it will be demonstrated that part of my alternative stemma of the *Epistola* is confirmed by the traditions of texts that constitute a particular dossier, as far as those traditions have been studied.

1. The Corpus of Alexander Manuscripts

For some years I have been interested in the main sources of the histories of Alexander the Great found in Helinand of Froidmont's *Chronicon* and in Vincent of Beauvais' *Speculum Historiale*.² These sources are the so-called Zacher-*Epitome* and the *Epistola*; Vincent also added substantial fragments of the *Collatio Alexandri cum Dindimo* (hereafter: *Collatio*) to his account. The adaptation of these texts in Helinand's and Vincent's histories does not always correspond to the modern editions of the three Alexander texts, and they differ notably from the edition of the *Epistola* by Walther Boer (1953). To find an explanation for these differences, I searched for the manuscript sources actually used by Helinand and Vincent. A valuable tool for this

² An edition of the accounts of Alexander the Great in the *Chronicon* and the *Speculum Historiale* is in preparation; see Voorbij (1991:1-2) and Smits (1991:269) with further references. I discuss neither Helinand's nor Vincent's account of Alexander the Great and their manuscript sources here, as they will be the subjects of other publications.

quest was David Ross's check list of the three Alexander texts, enumerating about 150 manuscripts written between ca. 800 and ca. 1500 A.D.³ The check-list showed that the three texts frequently occurred together within the same manuscripts.

The three Alexander texts themselves date from before 800 A.D.⁴ The *Zacher-Epitome* goes back to the Greek Alexander Romance, the author of which is commonly called Pseudo-Callisthenes. Dating from the third century A.D. in Egypt, this work describes Alexander's descent from the Egyptian magician-king Nectanebus, his youth, his accession to the throne of Macedonia, his campaigns in Persia and India where he triumphed over kings Darius and Porus, his miraculous adventures in the East, and, finally, the conspiracy leading to Alexander's death.

There are several versions of the Greek romance, two of which were translated into Latin: the α - and the δ -versions of Pseudo-Callisthenes. Only the α -version is important for our corpus. It became known through its translation by Julius Valerius (ca. 310)⁵ that had its widest circulation in an abridged form written before the ninth century. This abridged version is known nowadays as the *Zacher-Epitome*, after the first editor (Zacher 1867), and survives in 80 manuscripts.

For the sake of completeness, the Latin translation of the δ -version of Pseudo-Callisthenes must be mentioned here. In the tenth century Leo Archipresbyter of Naples translated a copy of this δ -version. The translation itself, the *Nativitas et Victoria Alexandri Magni*, has not survived in its original form, but can be reconstructed from later adaptations. One of these, produced in the eleventh century, improves the style of Leo's translation and expands it with material taken from other texts dealing with Alexander the Great, such as the *Epistola* and the *Collatio*. The result, nowadays known as the first interpolated version (*I*¹) of the *Historia de Preliis*, was adapted twice: one adaptation, executed in the late eleventh or early twelfth century, is commonly called the second interpolated version (*I*²) of the *Historia de Preliis*; the other, an adaptation independent of *I*², is known as *Historia de Preliis I*³.⁶ Although the derivatives of Leo's translation survive in many

³ Ross (1956). Due to a few misprints and improper or uncertain identifications, it is still impossible to deduce the exact number of manuscripts or total occurrences of each text in this list. For additions to Ross's list, see Hahn (1980) and Voorbij (1984), with corrections to Hahn; the check-list lists ca. 190 manuscripts now.

⁴ For general information on medieval Alexander literature see Cary (1956) and Ross (1988).

⁵ Edited by Kübler (1888); for the dating 'ca. 310' see Herzog (1989:212-214).

⁶ The most recent survey of Leo's translation and its Latin derivatives, with a critical dis-

manuscripts that are found all over Europe, they appear to be almost totally absent from our corpus of Alexander manuscripts. Only one copy has been found so far in combination with two of the three Alexander texts under consideration.⁷ I, therefore, choose to ignore the *Historia de Preliis* tradition.

The *Epistola* and *Collatio* also draw us back to the Greek literature of late Antiquity. In those days fictional letters from Alexander to contemporaries were in circulation; their actual authors remain unknown. One of these pieces was a letter to Aristotle describing Alexander's miraculous adventures in India during and after his successful campaign against king Porus. An abbreviated version of this letter was incorporated into the Greek Alexander Romance. Several translations of the letter survive. Julius Valerius probably was the first to put it into Latin: he incorporated an amplified rendering of the letter in his translation of the Greek Alexander Romance mentioned above (Kübler 1888:123-135). A second, even more elaborate translation came into existence between the fourth and seventh centuries. This translation, known as the *Epistola*, has been found in about 135 manuscripts.⁸

The Latin *Collatio*, a fictional correspondence between Alexander and Dindimus, king of the Brahmans, was composed before the end of the fourth century; it also may descend from a Greek original. The correspondence consists of five letters in which Alexander enquires about the life of the Brahmans; Dindimus describes their ascetic life and attacks the Macedonian way of living and Alexander, in the final letter, ridicules the Brahmanian asceticism.⁹ The *Collatio* survives in more than 80 manuscripts.

cussion of editions and translations, is given by Smits (1993). Kratz (1991) must be added to his list of translations.

⁷ MS. Madrid, Biblioteca Nacional, 9783 (early 13th century) contains, among other texts, the *Epistola*, the *Zacher-Epitome* and the *Historia de Preliis I'*.

⁸ Modern critical editions by Kübler (1888:190-221) and Walther Boer (1953), the latter mentions earlier editions based on one manuscript each; for English translations see Gundersen (1980) and Kratz (1991:107-125). Before the year 1000 a revision of the *Epistola* was made, known as the *Epistola II*; see Voorbij (1984:117) with further references. Since the *Epistola II* is found exclusively in connection with Leo's translation, it has been left out of our corpus.

⁹ The *Collatio* was edited by Kübler (1888:169-189). A revision of the correspondence, referred to as *Collatio II*, was made around the tenth century. This revision in its turn was adapted later, probably in the eleventh century; the text is known as *Collatio III*. Since *Collatio II* and *Collatio III* are exclusively connected to the textual traditions of Leo's translation and the *Historia de Preliis*, they have been left out of our corpus. For these revisions see Voorbij (1984:118) with further references.

As stated above, Ross's check-list shows that two or all three of our Alexander texts are frequently transmitted together in the manuscripts. The *Zacher-Epitome* and the *Epistola* survive together in more than 60 manuscripts, whereas the combination of the *Epistola* and the *Collatio* has been found in over 30 copies. All three Alexander texts occur together in more than 30 manuscripts. Only the combination of the *Zacher-Epitome* with the *Collatio* has not been found so far. The dominant presence of the *Epistola* in our manuscript corpus justifies that I concentrate my attention on this text.

Other works dealing with Alexander the Great, such as the *Historia de Preliis* (mentioned above) and the *Iter ad Paradisum*, can be found in our corpus; however, since they occur only occasionally in combination with the *Zacher-Epitome*, the *Epistola*, the *Collatio* or other Alexander texts, they do not play a role in the composition of the corpus.

2. Walther Boer's Edition of the *Epistola*

The great number of surviving manuscripts is not the only reason for my focus on the *Epistola*: the editorial state of affairs also guided my choice. Zacher (1867) used seven manuscripts for his critical edition of the *Epitome*, but he also relied on three copies of the unabbreviated Julius Valerius text as well as on secondary sources such as Vincent of Beauvais' *Speculum Historiale*. Probably he was unaware that more than 70 other copies of his text existed. With a few exceptions, most manuscripts of the *Zacher-Epitome* have not been studied in detail.¹⁰ A similar situation exists concerning the *Collatio*: for his edition, Kübler (1888) used three manuscripts and two earlier editions. The actual contents of the remaining more than 80 copies is still unknown.

For the *Epistola* the state of affairs is much better. Whereas Kübler (1888) used eight copies, Walther Boer (1953) mentioned 67 manuscripts. A closer look into the *apparatus criticus*, however, reveals that Walther Boer, for different reasons, selected 28 out of these 67 copies, the variant readings of which he used for the construction of his critical text. Ross (1956) showed that almost half of the extant manuscripts did not come to Walther Boer's attention.

¹⁰ One exception is the so-called Oxford-Montpellier *Epitome*, which offers a different redaction of the *Zacher-Epitome*; see Hilka (1911:31-69).

The *Epistola* has a complicated textual tradition. The existence of several distinct redactions as well as contaminations of these redactions and other sorts of adaptation suggests that the *Epistola* was a *texte vivant*. Editing this Alexander text is a laborious if not Herculean task, as Kübler (1888:XXXI) remarked, though this may not function as an excuse for negligence or inconsistency. In the past years I studied Walther Boer's *apparatus criticus* and compared it with several *Epistola* manuscripts, copies which Walther Boer had seen, as well as copies that had been found since 1953. I was able to make several interesting conclusions, the most important of which will be discussed here briefly.¹¹

Among the manuscripts studied, Walther Boer distinguished four families which he indicated with the Roman numerals I-IV. These numerals also express a value judgement. Family I represents a closed recension bearing the best text whereas, on the other end of the scale, family IV is an open recension with many inferior readings.¹² Walther Boer based his edition primarily on the common readings of family I. The *apparatus criticus* shows only part of the textual variants found in the other families. To prevent the apparatus from being overloaded, the editor omitted any variants found in just one copy (Walther Boer 1953:XXXVI). Although understandable from the editor's point of view, this method is regrettable in view of the great number of manuscripts that have since been recorded. Because individual readings, which can be valuable in reconstructing relations between manuscripts, have been omitted, it is often impossible to determine the position of additional copies properly within Walther Boer's *stemma codicum*.

An objection must be made to the omission process itself as applied by the editor. In the past years I have compared the readings of several *Epistola* manuscripts used by Walther Boer with the text and critical apparatus of his edition. It appears that some unrecorded single variants do occur in several copies. Consequently they are not unique and should have been mentioned in the *apparatus criticus*. Walther Boer either overlooked these variants or applied the omission principle on his own manuscript collation during the collation process itself. Once having skipped a variant as unique to one copy, upon meeting this variant in a next one, he may not have verified whether this reading had already been found in earlier colla-

¹¹ Within the space allotted to me here, it is impossible to discuss these and other conclusions in detail. I will return to them at another time.

¹² For closed and open recensions see, e.g., West (1973:29-47). Schnell (1989:66-67) disputes Walther Boer's stemma of the *Epistola*'s family IV.

tion stages. To what degree the omission process affected his stemma remains to be studied.

Walther Boer's leading textual witness of family I is MS. Leiden, University Library, B.P.L. 20, from the Benedictine monastery Le Bec in Normandy, indicated with the symbol *H*. He emended its text from other copies and used the edition by Kübler (1888) as well, as is evident from a series of references. Walther Boer studied Kübler's critical text, but apparently neglected its preface wherein Kübler not only quoted but also endorsed Zacher's observation that the scribe of the Leiden manuscript repeatedly changed and amplified the poor text of the *Epitome*.¹³ Nowhere does Walther Boer show awareness of this quotation by Kübler. Kübler (1888: XXIX) had the same criticism of the Leiden text of the *Epistola*.

Codicological evidence undermines Walther Boer's *stemma codicum*.¹⁴ The two oldest copies of family I appear both to originate from the same cultural area, namely, twelfth-century Benedictine monasteries in Normandy. The first copy is the Leiden manuscript, mentioned above, that comes from Le Bec. It is made up of two parts; the sections were bound together by 1164. Apart from the *Epistola*, the first part, dated ca. 1139, contains the *Gesta Normannorum Ducum* by William of Jumièges, Einhard's *Vita Karoli Magni*, the Zacher-*Epitome*, the *Abbreviatio Gestorum Regum Franciae* and the *Genealogia Comitum Flandriae*; the second part, dated ca. 1160, contains Geoffrey of Monmouth's *Historia Regum Britanniae*, the *Historia Brittonum* and a fragment of the *Historia Ecclesiastica* by Orderic Vitalis.¹⁵

The second copy, MS. Paris, Bibliothèque nationale, lat. 5062 (early 12th century), belonged to the monastery of Fécamp. In the Paris manuscript the last part of the *Epistola* text is missing. Fortunately, this part was located by Avril (1965) during his quest for Norman manuscripts; it is kept at the Vatican library as an organic item of MS. Ottob. lat. 909. Apart from the *Epis-*

¹³ Zacher (1867:X): "Der schreiber von B [= the Leiden manuscript] stand, der in Bec herrschenden gelehrsamkeit entsprechend, an kkenntnis und geschmack weit über der niedrigen stufe seiner vorlage, und hat demzufolge den text ziemlich oft, und meist auch ganz geschickt verändert und wirklich verbessert; doch darf der herausgeber ihm nur sehr vorsichtig folgen, da er den text meist correcter und eleganter gestaltet hat, als der verfasser selbst vermocht und beabsichtigt hatte."

¹⁴ See Figure 1. Apart from the Leiden (= H) and Paris/Vatican (= Pari) manuscripts discussed below, the following copies appear in the stemma: Cambridge, Gonville and Caius College, 177/210 (late 12th century; = Gc); London, British Library, Cotton Nero D VIII (late 12th, early 13th century; = Cn); and Oxford, Lincoln College, 96 (early 13th century; = Ol).

¹⁵ For a description of Leiden B.P.L. 20, see Van Houts (1982:229-231; 1992:CIX-CX).

tola, the Paris/Vatican copy contains Hegesippus' *Historiae*, the Zacher-*Epitome*, Einhard's *Vita Karoli Magni* and Paulus Diaconus' *Historia Langobardorum*. According to Avril, the Paris/Vatican manuscript is about three decades older than the Leiden copy. He suggested that the texts dealing with Alexander the Great and Charlemagne in the Leiden manuscript were copied from the Paris/Vatican manuscript.¹⁶

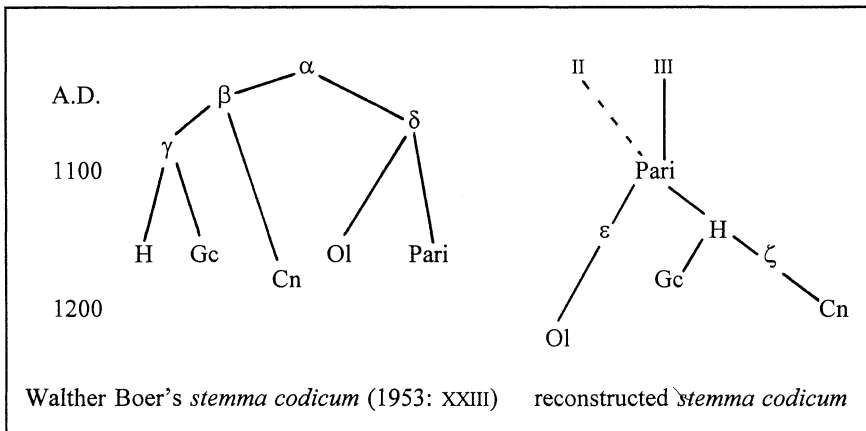


Figure 1. *Stemma codicum* of the *Epistola*'s family I

At least for the *Epistola*, Avril's suggestion is confirmed by my own comparison of both copies. Originally the text of the *Epistola* in the Paris / Vatican manuscript represented Walther Boer's family III. After its completion, however, a second hand collated the text with a copy of family II, replacing in several places the original text with readings from this family II, as is evident from erasures and interlinear and marginal additions. Thus, this second hand produced a contamination. Afterwards, the scribe of the first part of the Leiden manuscript copied the contaminated text in full from the Paris/Vatican exemplar, incorporating all changes in the running text and correcting a few minor errors produced in the earlier contamination process itself. The Leiden manuscript, in turn, underwent adaptations by several later hands (Van Houts 1982:229-231; 1992:CIX-CX). Apart from some minor elements, one of these scribes wrote down, as a marginal addition, an apocryphal letter from king Porus of India to Alexander (cf. Walther Boer

¹⁶ Avril (1965:211-216). See also Pellegrin (1975:452) for the Vatican manuscript.

1953:26). All descendants from the Leiden manuscript incorporated this letter into the body of the text.

This evidence leads to two conclusions. In the first place, the position of all copies within family 1 of the *Epistola* needs to be reconsidered. Priority must be given to the Paris/Vatican exemplar, to which most probably all other members of family 1 go back. In the second place, and of more importance, the position of family 1 itself within the *stemma codicum* must be reconsidered. Whereas in Walther Boer's opinion this family represents the best text of the *Epistola*, it should be treated as a Norman contamination from the early twelfth century. Any edition of the *Epistola* should use as point of departure one of the other families distinguished so far.

3. Selection and Classification: Support from the Transmission Context

The editor of the *Epistola* has to face similar problems as the editors of many other texts. Confronted with a large quantity of manuscripts, part of which show up contaminations or adaptations, he has to collect and study all copies, construct a *stemma codicum* and, finally, decide which textual witnesses meet his editorial standards.

When the tradition of a certain text has not been studied before, the editor, at first in random order, selects and collates manuscripts and proceeds exploratorily while grouping and classifying the witnesses. In the latter processes he runs the risk of becoming involved in a long sequence of retrials. If the editor lacks a proper overview of all the material, he might lose much time.

When earlier editions exist, this risk may be reduced, since already available information can be used. In this situation, the order in which manuscripts are studied as well as the arrangement of tentative classifications may be based on the groupings made by preceding editors as a point of departure. Continual carefulness is needed. The discovery of additional manuscripts or a reconsideration of copies familiar to the editor's predecessors may lead to a divergent opinion about the importance of witnesses and their correlations. Too great a dependence on classifications by earlier editors may form a pitfall in the end.

One may ask whether the editor can use any other – external – information to help in the editing process. What sort of information can assist in the selection and classification of manuscripts? In many instances the answer to these questions can be found in the study of the transmission context of a text.

The size of a text has implications for its manuscript transmission. Vincent of Beauvais' *Speculum Historiale*, for instance, is so bulky that it cannot be contained in just one manuscript volume, but requires at least two volumes (cf. Voorbij 1991:46-55). As opposed to this, many texts, due to their modest size, are grouped with other texts in one manuscript. A systematic comparison of the contents of many manuscripts may reveal that the same group of two or more texts occurs repeatedly. An example of this phenomenon is the group of the *Zacher-Epitome*, *Epistola* and *Collatio* that survives in more than 30 manuscripts of our corpus (see section 1).

This repetition of a group suggests that several combinations as such were fixed. It also suggests that, apart from traditions in which texts have been copied one by one from different sources, textual traditions existed in which groups were transmitted from one manuscript to a second, and so on. This may have consequences for the editing process. If the manuscript tradition of a certain text shows a repeated grouping with one or more other works, the editor, on constructing a *stemma codicum* of just this individual text, possibly draws up a classification of that particular combination of writings.

If this hypothesis is correct, the question arises whether the editor can make a classification indirectly by using other texts. If he is confronted with problems in the *stemma codicum* of the text he is editing, can the editor use the pedigrees of works repeatedly grouped with his and derive from them additional arguments to establish relationships between copies of his text? I believe so.

I decided to employ this technique to tackle the problems caused by Walther Boer's stemma of the *Epistola*. At this point, the combination of the *Epistola* with the *Zacher-Epitome* and the *Collatio* is of no help. Because of the editorial state of affairs, no stemmas or other information on relationships between copies of those two texts and the *Epistola* is available (see section 2). A cursory survey of our corpus, however, shows that the Alexander texts themselves occur with some regularity in combination with one or more of the following works: Pseudo-Seneca's *Epistolae ad Paulum et Pauli ad Senecam*, Dares Phrygius' *De Excidio Troiae*, Paulus Diaconus' *Historia Langobardorum*, Einhard's *Vita Karoli Magni* and Geoffrey of Monmouth's *Historia Regum Britanniae* (cf. Voorbij 1984:119). These texts have been edited previously and some editions offer information on textual traditions. A further study of catalogues of manuscripts showed that, apart from the writings mentioned above, several others survive in repeating groups with one or more of the Alexander texts.

A few years ago systematic research into the entire corpus of Alexander manuscripts was begun at the University of Groningen.¹⁷ Its aims are to establish whether a particular combination of texts is medieval or post-medieval; whether each text of a group has been copied individually from a different exemplar or whether all were copied from one and the same manuscript; whether or not the existence of that particular group of texts was coincidental; why a particular combination came into existence, if created on purpose; and how popular particular groups of texts were.

As much information as possible regarding all texts occurring in the corpus (ca. 700) as well as codicological information on each of the ca. 190 manuscripts has been collected in a database. Part of the data, however, appears to be inconsistent and needs revision and normalisation. The degree of consistency obtained depends upon evidence on the following topics:

- *Definitions.* A terminology concerning combinations of texts is needed to describe properly the phenomena encountered in manuscripts. I will return to this point in section 4.
- *Codicological descriptions.* For some manuscripts, no descriptions exist. In other instances, especially for composite manuscripts, the available descriptions are incomplete, uninformative or unsatisfactory from a codicological point of view. They frustrate any clear understanding of the relationship between the structure and the content of the manuscript. Fortunately, good manuscript descriptions do exist; in the other cases, reliable descriptions or additional information must be obtained by scrutiny of the source itself.
- *Characterisation of subject matter.* Our axiom is that a combination of texts should constitute a meaningful unity in terms of subject matter. A combination of a treatise on tanning with an exposition on the Lord's Prayer and Dares Phrygius' history of Troy does not seem to make sense; it is possibly an accidental collocation of texts. On the other hand, a combination of the *Epistola*, which describes phenomena of the Indian animal world, with travel stories of the East or encyclopedic works seems sensible. These texts may have been collected because they deal with the same subject matter: the East or animal life. Consequently, and where possible, one or more keywords have been attached to each text in the database to characterise its subject matter.

¹⁷ This research is being carried out in collaboration with Martin de Ruiter, Aafje Lem, Mieke Lens, Bennie Werink and Marinus Woesthuis (all of the Department of Medieval Studies, University of Groningen). We are now in the process of refining and normalising the data.

- *Information offered by editors and other scholars.* Some editions or studies of texts abundantly document the manuscript tradition of the work they are dealing with, whereas others limit themselves to just a few remarks. The latter publications are usually of little value for the recognition of combinations of texts. The former must be used with care: upon comparison, their stemmas and other data – codicological observations or quotations of rubrics or initial and final words – should produce a consistent picture. In case of inconsistency, unambiguous data need to be obtained.

4. Dossiers: Texts and Transmission

An insight into the relationship between structure and content of each manuscript as well as an understanding of the subject matter of each text are crucial to the correct interpretation of groups of texts. Similarly, an understanding of the manuscript as a physical copy of a text – the copy is complete or incomplete – as opposed to the manuscript as a textual witness – the text is complete, fragmentary, adapted or represents a particular family – is also important. The sum of these types of information will help to clarify how texts have been copied and transmitted: one by one from different exemplars, combination by combination, or by a mixture of both.

The following terminology is used to describe the relations between manuscript and text:

- *Manuscript:* a sequence of quires and leaves written by one or more scribes that is intended to be bound as a unity in a binding. The manuscript may survive completely or fragmentarily (fragment, *membrum disiectum*).
- *Composite manuscript:* a bundle of two or more (fragments of) distinct manuscripts not originally intended as a whole but assembled by recipients: medieval as well as post-medieval book collectors and librarians. Of special interest in this context is the ‘booklet’: a group of single leaves forming at least one quire that usually has been produced separately, to be bundled with one or more other booklets into a larger, composite manuscript. Features which may characterise a booklet are described by Robinson (1980) and, more precisely, by Hanna (1986).
- *Volume:* the physical unity of a binding and the manuscript(s) contained in it. A volume may contain a manuscript or a composite manuscript.

- *Text*: any of the various forms in which the original words of a chronicle, romance, sermon, charter, treatise, letter, etc., survive. The text may be complete or incomplete. An incomplete text may represent a part, a fragment or an excerpt. Interpolations and continuations may have been added to any text. Another type of change is identified in relation to the following item.
- *Dossier*: a group in one manuscript of two or more texts, the subject matter of which constitutes a unity; especially such a group occurring with some regularity within a corpus of manuscripts. A dossier reflects the interest and taste of medieval recipients. The dossier is fixed as regards the texts that belong to it; the order in which the texts are presented, however, may vary. Moreover, in the course of time the dossier may show changes: additional texts may have been added, resulting in a new fixed combination; conversely, a text may have been omitted from the combination, resulting in a smaller dossier.

Texts belonging to a dossier may have been contextualised, i.e., conflicts between their contents may have been eliminated, the transition from one text to the next may have been eased by adaptation of the concluding or introductory phrases or by an interpolation of the text.¹⁸ Individual texts showing these kinds of changes possibly originate from a dossier.

In several instances the dossier coincides with a booklet or with an entire manuscript.

- A *dossier manuscript* contains one or more dossiers. Apart from this, such a manuscript may have one or more other, individual texts.

The relations between the terms may be represented as follows:

volume	—	no composite manuscript	—	1 manuscript
volume	—	composite manuscript	—	2 or more manuscripts
manuscript	—	1 text	—	no dossier
manuscript	—	2 or more texts	—	<i>possibly</i> 1 dossier
manuscript	—	4 or more texts	—	<i>possibly</i> 1 or more dossiers

The transmission process may have obscured part of the distinctions made so far. For instance, a composite manuscript that was assembled in the Middle Ages may in turn have functioned as a dossier manuscript. The

¹⁸ A striking example of contextualisation within a dossier survives in MS. Paris, Bibliothèque nationale. Nouv. acq. lat. 310 (second half 12th century); see Schnell (1989).

selection of manuscripts or booklets made by the medieval collector possibly served, in part or in full, as a dossier that has been copied later. The original manuscripts themselves may have contained dossiers as well. Apart from this, the copying of manuscripts may have obscured the presence of distinct physical units in their exemplars, a process that may be called codicising (cf. Hanna 1986:103). The copy may have integrated or imbedded the original sections of the composite manuscript, which had been produced separately, into one sequence of quires or other units. Similarly, booklets may have been linked into greater units and the original breaks between them may have disappeared.

Some implications of the transmission process are displayed in Figure 2.

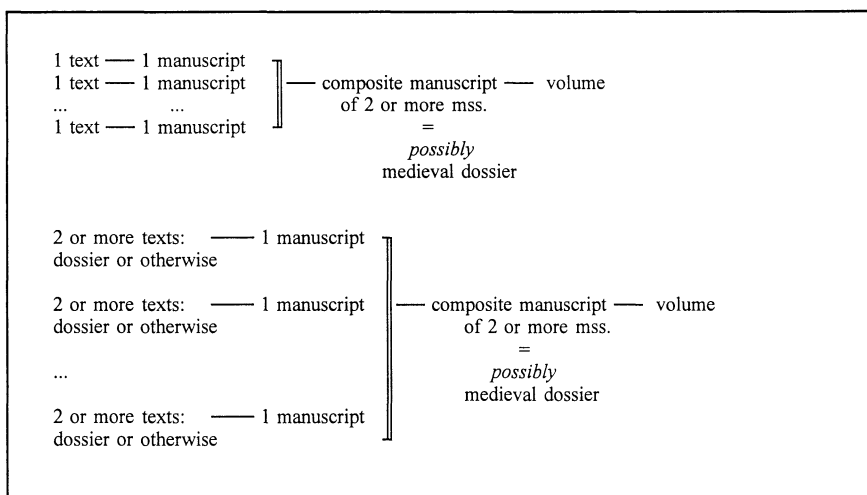


Figure 2. Text and dossier transmission (a selection of paths of transmission).

Using the terminology above, the following types of dossiers have been identified within the corpus of Alexander texts so far:

- *Collatio* + writings by Church Fathers: dossiers on Christian rules (the Brahmins as Christians *avant la lettre*);
- *Collatio* + letters by (pseudo-)Seneca: moralistic dossiers;
- *Collatio* + a collection of letters by several authors: dossiers of specimen letters;
- *Epistola* + *Collatio* + descriptions of the East: geographical dossiers on the East;
- *Epistola* + book of penances + travel books: dossiers for pilgrimages;

- *Epistola* + works on natural history, especially Solinus: encyclopedic dossiers of flora and fauna, more specific of the marvels of India;
- *Zacher-Epitome* + *Epistola* + historical works (literary accounts included, like Dares Phrygius' on the fall of Troy): dossiers of historiography.

I return now to the question whether the editor can use the pedigrees of works grouped with the text he is editing and derive from these additional arguments to support his own *stemma codicum*.

First, the editor has to establish which works occur frequently in his corpus of manuscripts and whether they constitute one or more dossiers with his text. Using the corpus of Alexander manuscripts, I present the texts most frequently found in the list below, the Alexander texts excluded (I limit myself here to a frequency of at least 5). They belong mostly to historiographic dossiers, sometimes we find them in moralistic or encyclopedic dossiers:

	number of occurrences within the Alexander corpus
Dares Phrygius, <i>De Excidio Troiae</i>	21
Einhard, <i>Vita Karoli Magni</i>	8
<i>Epistola Presbyteris Johannis</i>	9
Geoffrey of Monmouth, <i>Historia Regum Britanniae</i>	17
<i>Historia Appollonii Regis Tyrii</i>	11
Pseudo-Seneca, <i>Epistolae ad Paulum et Pauli ad Senecam</i>	13
Seneca, <i>Epistolae Morales ad Lucilium</i>	11
Solinus, <i>Collectanea Rerum Memorabilium</i>	10
Pseudo-Turpinus, <i>Historia Karoli Magni et Rotholandi</i>	8
William of Jumièges, <i>Gesta Normannorum Ducum</i>	5

The second step is to consult editions of and studies on those texts that constitute a dossier, and to determine which of these publications offer unambiguous and consistent information. At this point, one meets a variety of situations, from silence on textual traditions to a display of solid documentation. In the corpus of Alexander manuscripts the following situations occur:

- The available editions and studies do not contain any conclusive evidence concerning the *stemma codicum* or the tradition of the text under consideration. Sometimes there is neither a systematic list of manuscripts nor a stemma available (see, for instance, Meister's 1873 edition of Dares Phrygius' *De Excidio Troiae*). These publications offer too few possibilities to study a text in relation to its transmission context. In other instances the textual tradition has been studied abundantly and groups of

manuscripts have been isolated. No more detailed classification, however, let alone a stemma, has as yet been established. This applies, for example, to Geoffrey of Monmouth's *Historia Regum Britanniae*: the impressive studies by Crick (1989; 1991) identify groups of manuscripts and isolated copies, but only few transmission lines. The available information offers limited possibilities for their identification.

- The available stemma outlines the transmission of individual text traditions (families, versions, redactions); sometimes it also indicates in which specific regions of Europe they have circulated. The stemma, however, does not display the relations between individual manuscripts, or does this only partially or vaguely. This type of classification is available in editions of, e.g., Solinus' *Collectanea Rerum Memorabilium* by Mommsen (1895) and the *Epistola Presbyteris Johannis* by Zarncke (1879). These data do have a limited value: they may help to determine where and when a particular dossier has circulated.
- The editions and studies contain no satisfactory *stemmata codicum*, as they present only a subset of the known manuscripts. This holds especially true for texts edited following the principle *codices recentiores codices deteriores*. Their editors often omit information on copies later than the twelfth century, even if the texts in question enjoyed a great popularity during the later Middle Ages. An example of this method is the edition of Pseudo-Seneca's *Epistolae ad Paulum et Pauli ad Senecam* by Barlow (1938). Since they leave too many gaps in the transmission lines, these publications, too, have only limited value for the study of dossiers.
- The available publications offer a well-balanced amount of information consisting of stemmas, quotations of rubrics and incipits, and unambiguous codicological data that express the relationship between structures and contents of manuscripts. This category of studies, for instance the studies by Van Houts (1982; 1992) on William of Jumièges' *Gesta Normannorum Ducum*, offer ample opportunities for comparing stemmas of different texts.

Finally, the editor has to evaluate and weigh the available information. He has to look for data that support or contradict his own stemma. At this stage he will primarily look for parallels between his stemma and those of other texts within germane dossiers. If the information obtained results in a consistent image of the transmission of both the text and the dossier(s) to which it belongs, he can safely rely on it to ascertain his *stemma codicum*.

5. Stemmas of a Historiographical Dossier

In the light of the preceding discussion, I will reconsider now the reconstructed stemma of the *Epistola*'s family 1. Apart from establishing the contamination which underlies this family, my reconstruction leads to the following differences with Walther Boer's stemma: ms. *Pari* is placed at the top of the stemma, only two intermediary stages (ϵ and ζ) are assumed, and *H* is designated as the manuscript from which at least two surviving copies are derived (*Gc* and *Cn*).¹⁹ Is my stemma supported by the pedigrees of other texts, using dossiers instead of individual textual witnesses to construct it?

To identify dossiers in the five manuscripts of family 1, the codicological structure of each copy must be studied first. It appears that four out of these five manuscripts are composites (see Figure 3). A fifteenth-century section was added to the Vatican part (*Ottob*) that constitutes a unity with *Pari*. It is not known if these sections were assembled before or after the section containing the *Epistola* had been split up (cf. Pellegrin 1975:452). The volume indicated with *Ol* is composed of three manuscripts. The first two, on folios 1-87 and 88-121, respectively, contain theological and canonical writings. The third manuscript, on folios 122-138, includes among other works the *Epistola*.²⁰ The Leiden volume (*H*) is made up of two manuscripts, the first of which contains the *Epistola*. The two manuscripts were brought together by 1164 (Van Houts 1992:CIX). In the last composite manuscript, *Cn*, three manuscripts are assembled; the first contains the *Epistola*. It is not known when these sections were bound together (Van Houts 1982:245; 1992:XCVI).

The codicological history of the fifth manuscript, dated in the late twelfth century, is different. At some point before 1560 it was divided into three parts, two of which have been recovered: *Gc*, in which the *Epistola* is contained, and *Lv*, which is now a part of a composite manuscript in which Sir Robert Cotton assembled fragments of twelfth-, thirteenth- and fourteenth-century manuscripts. The third part, containing the pseudo-Gildasian *Historia Brittonum*, is still missing (Van Houts 1982: 237-239; 1992: CXII-CXIII).

¹⁹ See Figure 1. To the list of symbols used so far (cf. footnote 14) I add the following two: *Lv* (= MS. London, British Library, Cotton Vitellius A VIII, late 12th century) and *Ottob* (= MS. Vaticano, Ottob. lat. 909, early 12th century).

²⁰ These data have been supplied by Frans van Liere (University of Groningen) who examined *Ol*.

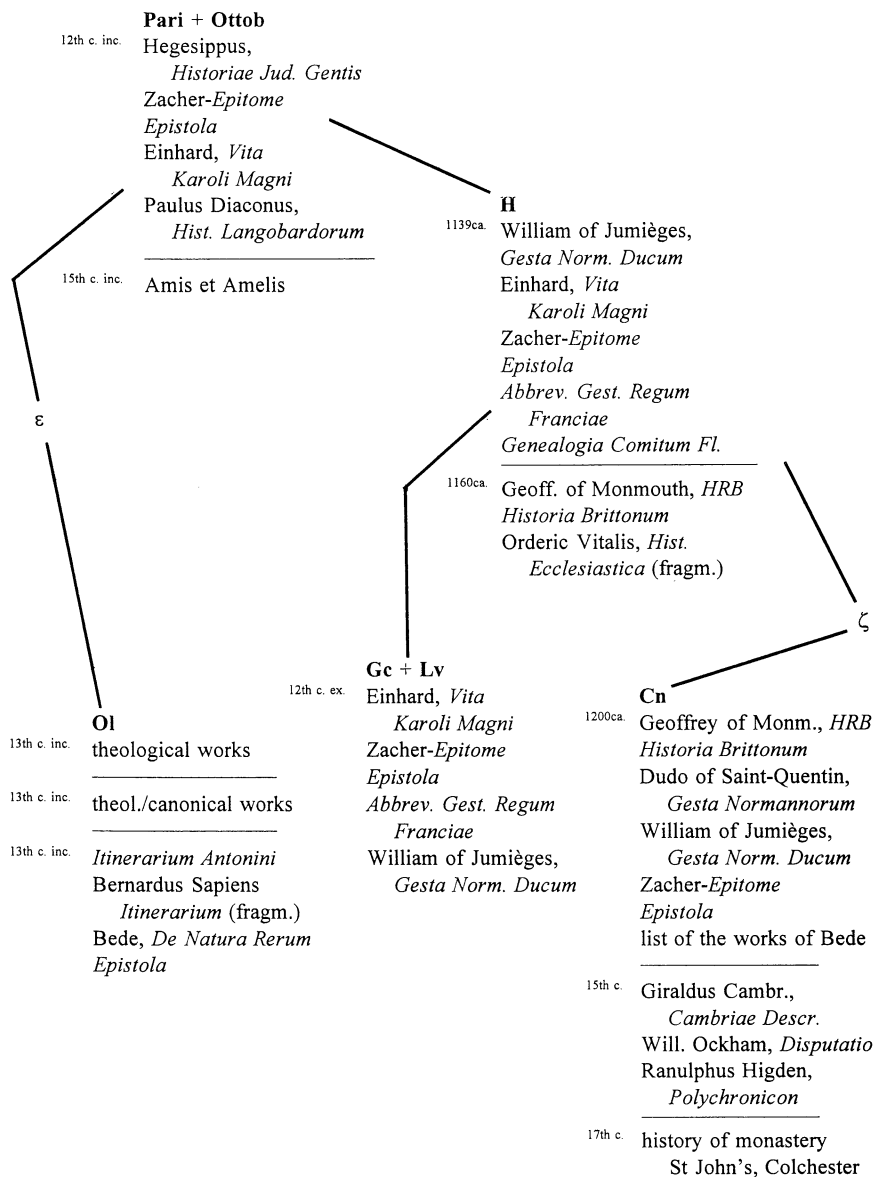


Figure 3. Structures and contents of manuscripts belonging to the *Epistola*'s family 1 (reconstructed stemma)

Let us concentrate now on the subject matter of the manuscripts or sections in which the *Epistola* is contained, discussed according to my stemma. In four out of the five copies the presence of a gradually changing historiographical dossier can be discerned. *Pari/Ottob*, at the top of the stemma, presents, more or less in a chronological order, the history of the Jewish people, of Alexander the Great as depicted in the *Zacher-Epitome* and *Epistola*, of Charlemagne and, finally, of the Lombards.

In the left branch of the stemma the subject matter changes drastically. The section of *Ol* in which the *Epistola* is incorporated also contains two travel books and Bede's *De Natura Rerum*. This group of texts is best characterised as a geographical dossier on the East, or possibly a dossier for pilgrimages. The reasons for this change of subject matter remain to be studied.

As opposed to this, the right branch of the stemma shows a shift of historiographical interest. In the first section of *H*, the history of the Anglo-Norman dukes and kings is placed prominently in front. Next, attention is paid to the history of the Frankish and French leaders, in the middle of which the history of Alexander the Great has been placed. The line Anglo-Normans – Charlemagne – Alexander the Great suggests that the history of the two latter kings served as material for comparison, in order to show the greatness of the Anglo-Normans. This use of histories has been observed several times, especially with regard to Alexander the Great (cf. Cary 1956). By 1164 the first section was bound with a second that was almost exclusively devoted to the history of British leaders. It contained, apart from a fragment of Orderic Vitalis' *Historia Ecclesiastica*, Geoffrey of Monmouth's *Historia Regum Britanniae* and the pseudo-Gildasian *Historia Brittonum*. Because of this twelfth-century assemblage, the scope of *H* was extended to include a group of historiographical texts on Western-European kings and realms in which a prominent place was allowed to the Anglo-Normans and in which the history of Alexander the Great was used as reference material. For medieval users, both sections of *H* together could have served as a dossier, conform the last transmission scenario displayed in Figure 2.

This dossier could be used in different ways, as is shown by *Gc+Lv* and *Cn*. Most of the texts of *H* were copied into *Gc+Lv*, although they were reorganised in a geographical sense: first, texts dealing with the Frankish and French kings, then William of Jumièges' history of the Anglo-Norman leaders and, finally, in the still missing part of the manuscript, the *Historia Brittonum*. Here the history of Alexander the Great serves as reference material for the history of Charlemagne. On the other hand, the section of *Cn* that contains the *Epistola* concentrates on the history of the British kings

first and is followed by the history of their successors, the Anglo-Normans. The presence of Alexander the Great right after William of Jumièges' account once again emphasises the greatness of the Anglo-Normans.

According to my stemma, the central works of this historiographical dossier were originally the *Zacher-Epitome*, the *Epistola* and Einhard's *Vita Karoli Magni*. In *H*, William of Jumièges' account and the *Abbreviatio Gestorum Regum Franciae* joined them as fellow travellers; all five pieces were copied into *Gc+Lv*. Furthermore, *H* also offered the opportunity to compose a new group of texts: the *Zacher-Epitome* and the *Epistola* were now associated with Geoffrey of Monmouth's *Historia Regum Britanniae*, the *Historia Brittonum* and William of Jumièges' work, as is apparent from *Cn*.

These paths of transmission also become visible, although only partially, from publications dealing with texts that, in my stemma, are associated at least twice with the *Epistola*. Unfortunately, neither Zacher's edition (1867) of the *Epitome* nor Waitz's edition (1851) of the *Abbreviatio Gestorum Regum Franciae* offer any information on stemmas. Publications dealing with three of the four remaining texts, however, do provide some evidence on the grouping of the manuscripts. In his edition of Einhard's *Vita Karoli Magni*, Holder-Egger (1911) assigned both *H* and *Gc* to the A-family of this text, as derivatives from *A2* (= MS. Vaticano, Reg. lat. 339). Unfortunately, neither Holder-Egger nor any of the other modern editors of Einhard's work was familiar with the manuscript *Pari/Ottob*. Consequently, they give no evidence in support of the link I established between *Pari/Ottob* and *H*. The two other texts confirm another grouping in my stemma, that of *H* and *Cn*, and in this order. These copies belong to the so-called Leiden group of Geoffrey of Monmouth's *Historia Regum Britanniae* (Crick 1991:187-188) as well as to the so-called group 3 of the *Historia Brittonum* (cf. Crick 1991:51).

The *stemma codicum* of the so-called F-redaction of William of Jumièges' *Gesta Normannorum Ducum*, however, supplies the best evidence in favour of the right branch of my own stemma (cf. Van Houts 1982:77): it shows similar relations between *H*, *Gc+Lv* and *Cn* (see Figure 4).

This stemma, however, also raises an interesting question. Could my intermediary stage ζ be identical with Van Houts' *F3*? This manuscript *F3* (Rouen, Bibliothèque Municipale, 1173 – Y.11) is a composite of three sections. The first, written in the second half of the eleventh century, contains Dudo of Saint-Quentin's *Gesta Normannorum*. The second section, dated at the end of the twelfth century, contains – apart from the F-redaction of William's text – Einhard's *Vita Karoli Magni*, the *Abbreviatio Gestorum Regum Franciae* and the *Iter Ierosolimitanum Caroli Magni*. Soon

after its completion this part was bound with the first. The third section, probably written before 1189, contains the A-redaction of William's text, a list of the dukes of Normandy, and a fragment of an unidentified chronicle. It is unknown when all three sections were assembled (Van Houts 1982:232-234; 1992:xcv, cx).

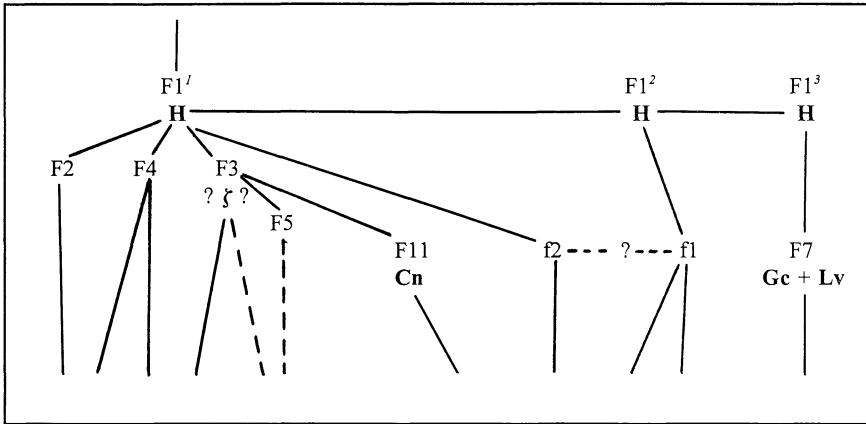


Figure 4. Reconstructed stemma of *Epistola*, family 1 integrated within the stemma of William of Jumièges' *Gesta Normannorum Ducum*, F-redaction (both stemmas partly displayed)

It is tantalising to speculate that the first two sections of this manuscript constituted a dossier from which *Cn* was derived. Such a speculation would imply that the second section also contained Geoffrey of Monmouth's *Historia Regum Britanniae*, the *Historia Brittonum*, the *Zacher-Epitome* and the *Epistola*, and that these texts are now missing. Thus far, however, no codicological evidence has been found in support of this speculation. Moreover, no confirmation can be deduced from available publications about one of the texts in question, except from the F-redaction of William's text. The relations between the copies of Dudo's *Gesta Normannorum* in *F3* and *Cn* have not yet been established (cf. Huisman 1984). Furthermore, neither Einhard's text of *F3* nor the *Abbreviatio* has been studied by an editor; no information on their relations with other copies is available. Similarly, there is no evidence for a missing copy of either the Leiden group of Geoffrey of Monmouth's work or group 3 of the *Historia Brittonum*. Until further evidence is obtained, this question will remain unanswered.

From all this material the following may be concluded. The information regarding the dossiers identified in copies of the *Epistola*'s family 1 does not

conflict with my stemma. However, no external evidence has been obtained in favour of the link between *Pari/Ottob* and *H*, and the question about the identity of $\zeta/F3$ is unanswered; at these points, only arguments deduced from the *Epistola* itself can be used. Whereas no support for the left branch of my stemma could be obtained, the right branch is corroborated by the traditions of several other texts. *H* is designated as the manuscript from which *Gc+Lv* (directly) and *Cn* (indirectly) were copied. The available information, if unambiguous, offers a consistent image.

6. Conclusion

The preceding sections show that, under certain circumstances and under fulfillment of specific conditions, editors can use medieval dossiers of texts to assist in the construction of more reliable *stemmata codicum*. The main prerequisite for this approach is an understanding of the context(s) in which a text was transmitted, in codicological terms as well as in terms of subject matter and content. The success of this approach depends on the availability of lucid, codicological information, i.e., unambiguous descriptions of the structures and contents of manuscript volumes. Thus, on the one hand, codicologists should be aware of the importance of their work for others; on the other hand, editors should use codicological information and study texts in relation to their transmission context(s). In the introductions to their editions they should display and document transmission context(s) of their texts. Such documentation is invaluable for editors studying works that are found together with some regularity in medieval dossiers. Any editor who presents such documentation indirectly contributes to the works published by fellow editors and in this way increases the understanding of the transmission of medieval texts in general.

July 1993

References

- Avril, F. 1965. "Notes sur quelques manuscrits bénédictins normands du XI^e et du XII^e siècle". *Mélanges d'Archéologie et d'Histoire* 77.209-248.
- Barlow, C.W., ed. 1938. *Epistolae Senecae ad Paulum et Pauli ad Senecam* <quae vocantur>. Rome: American Academy.
- Cary, G. 1956. *The Medieval Alexander*. Reprinted 1967, ed. by D.J.A. Ross. Cambridge: Cambridge University Press.

- Crick, J.C. 1989. *The Historia Regum Britanniae of Geoffrey of Monmouth*, vol. 3: *A Summary Catalogue of the Manuscripts*. Cambridge: Brewer.
- Crick, J.C. 1991. *The Historia Regum Britanniae of Geoffrey of Monmouth*, vol. 4: *Dissemination and Reception in the Later Middle Ages*. Cambridge: Brewer.
- Gunderson, L. 1980. *Alexander's letter to Aristotle about India*. Meisenheim am Glan: Hain.
- Hahn, T. 1980 "Notes on Ross's check-list of Alexander texts". *Scriptorium* 34.275-278.
- Hanna, R. 1986. "Booklets in medieval manuscripts: further considerations". *Studies in Bibliography* 39.100-111.
- Herzog, R. 1989. *Restauration und Erneuerung. Die Lateinische Literatur von 284 bis 374 n. Chr.* (= *Handbuch der lateinischen Literatur der Antike*, 5, ed. by R. Herzog and P.L. Schmidt). München: Beck.
- Hilka, A. 1911. "Studien zur Alexandersage". *Romanische Forschungen* 29.1-71.
- Holder-Egger, O., ed. 1911. *Einhardi Vita Karoli Magni*, 6th ed. (= *MGH Script. rer. Germ. in usum schol.*, 25). Hannover-Leipzig.
- Huisman, G.C. 1984. "Notes on the Manuscript tradition of Dudo of St Quentin's *Gesta Normannorum*". *Anglo Norman Studies VI. Proceedings of the Battle Conference 1983*, ed. by R. A. Brown, 122-135. Woodbridge: Boydell Press.
- Kratz, D.M. 1991. *The Romances of Alexander*. New York, etc.: Garland.
- Kübler, B., ed. 1888. *Iuli Valeri Alexandri Polemi Res Gestae Alexandri Macedonis* Leipzig: Teubner.
- Meister, F., ed. 1873. *Daretis Phrygii De excidio Troiae historia*. Leipzig.
- Mommsen, Th., ed. 1895. *C. Iulii Solini Collectanea rerum memorabilium*, 2nd ed. (reprinted Berlin 1958). Berlin: Weidman.
- Pellegrin E., et al. 1975. *Les manuscrits classiques latins de la Bibliothèque Vaticane*, vol. I. Paris.
- Robinson, P.R. 1980. "The 'Booklet'. A Self-Contained Unit in Composite Manuscripts". *Codicologica 3: Essais typologiques*, ed. by A. Gruys and J.P. Gumbert, 46-69. Leiden: Brill.
- Ross, D.J.A. 1956. "A check-list of mss of three Alexander texts: the Julius Valerius *Epitome*, the *Epistola ad Aristotelem* and the *Collatio cum Dindimo*". *Scriptorium* 10.127-132.
- Ross, D.J.A. 1988. *Alexander Historiatus: A Guide to Medieval Illustrated Alexander Literature*, second ed. Frankfurt am Main: Athenaem.
- Schnell, R. 1989. *Liber Alexandri Magni. Die Alexander Geschichte der Handschrift Paris, Bibliothèque Nationale, n.a.l. 310. Untersuchungen und Textausgabe*. München: Artemis.
- Smits, E.R. 1991. "Editing the *Chronicon* of Helinand of Froidmont: the marginal notes". *Sacris Erudiri. Jaarboek voor Godsdienstwetenschappen* 32.269-289.

- Smits, E.R. 1993. "From the Greek world to Eastern Europe: the changing fortunes of various Alexander traditions resulting in a new text (MS Königsberg UB 334)". *Polyphonia Byzantina: Essays in Honour of Willem J. Aerts*, ed. by H. Hokwerda, E.R. Smits(†) and M.M. Woesthuis, 257-273. Groningen: Forsten.
- Van Houts, E.M.C. 1982. '*Gesta Normannorum Ducum*'. *Een studie over de handschriften, de tekst, het geschiedwerk en het genre*. Ph.D. dissertation. Groningen.
- Van Houts, E.M.C., ed. 1992. *The Gesta Normannorum Ducum of William of Jumièges, Orderic Vitalis and Robert of Torigni*, vol. 1. Oxford: Clarendon Press.
- Voorbij, J.B. 1984. "Additions to Ross's check-list of Alexander texts". *Scriptorium* 38.116-120.
- Voorbij, J.B. 1991. *Het 'Speculum Historiale' van Vincent van Beauvais. Een studie van zijn ontstaansgeschiedenis*. Ph.D. dissertation. Groningen.
- Waitz, G., ed. 1851. "Historia regum Francorum monasterii Sancti Dionysii". MGH Scriptores 9.395-406.
- Walther Boer, W., ed. 1953. *Epistola Alexandri ad Aristotelem* Ph.D. dissertation. Leiden. Reprinted, 1973. Meisenheim am Glan: Hain.
- West, M.L. 1973. *Textual criticism and editorial technique applicable to Greek and Latin texts*. Stuttgart: Teubner.
- Zacher, J., ed. 1867. *Julii Valerii Epitome*. Halle.
- Zarncke, F., ed. 1879. "Der Priester Johannes". *Abhandlungen der Kön. Sächsischen Gesellschaft der Wissenschaften, Philol.-Hist. Classe*, 7.827-1039.

Textual Incompatibility and Many-Pronged Stemmata

William R. Veder
Slavisch Seminarium
University of Amsterdam, NL

The domain of the medieval *Slavia Orthodoxa*, which roughly encompassed present-day Russia, Belorussia, Ukraine, southeastern Poland, eastern Slovakia, Moldova, Romania, parts of Hungary, Serbia, Bosnia, Bulgaria, Montenegro, Macedonia and parts of Albania and Greece, with outposts on Mt. Athos, Mt. Sinai and in the city of Jerusalem, is largely a *terra incognita* as far as the comparative study of palaeography, codicology and the history of texts is concerned. Yet it has significant complements and correctives to contribute to the models of the possibilities and limitations of manuscript transmission of texts evolved from the study of Greek and Latin and, to a lesser extent, from that of vernacular western European literature.

Here, alphabetisation and acquisition of translated texts took place in recent times, starting in 863 with the mission of St. Cyril and St. Methodius (Tachiaos 1989). The faith implanted had a deeply monastic bias, the mission being intimately related to the Triumph of Orthodoxy twenty years earlier (ODB 1991:977,2122-2123), which evidently determined not only the selection of the texts to be translated (Thomson 1978), but the predominance of their closed tradition as well. It may also have determined the conscious randomisation of book structures in transmission in order to cater to meditative reading (Veder 1994a). Further, the texts were all translated in a single artificial language, based on the peripheral southern Slavic dialect of Thessaloniki, the home town of St. Cyril and St. Methodius, namely, Old Church Slavic, equally sacral and vernacular, and equally familiar and unfamiliar to any Slav. Finally, as early as 893, in Bulgaria a process was started to replace St. Cyril's Glagolitic alphabet, a truly original writing system, like the Armenian and Georgian, by an uncial Greek based alphabet which eventually took St. Cyril's name (Schaeken 1993).

I shall try to demonstrate the influence and the interaction of these four

factors in two works, a translation from Greek and one of the few original works of the early period. I shall give randomly selected specimens of their texts in a fully concordant English translation and refer the interested reader to the non-Slavic publications quoted for fuller bibliographic information.

1. The *Scete Paterikon*

The Systematic Collection of *Apophthegmata Patrum* (CPG 1974-1987:5562), compiled before ca. 550, was translated into Old Church Slavic probably by St. Methodius himself from a Greek codex of venerable antiquity in the year before his death in 885, and was later given the name above. Of this work of book length, at least seven different families of texts are known to me (an eighth is as yet, as a whole, inaccessible, due to the difficulties in obtaining access to the MSS and microfilms of them):

3 MSS (12-15th c.)	East Slavic: text family	a
4 MSS (15-16th c.)	East Slavic: text family	b
2 MSS (16 c.)	East Slavic: text family	i
2 MSS (12-16th c.)	East Slavic: text family	o
3 MSS (14-15th c.)	East Slavic: text family	c
2 MSS (14-16th c.)	South Slavic: text family	k
2 MSS (13-14th c.)	South Slavic: text family	w

At the beginning of chapter 22 (cf. the Slavic texts in Veder 1994b), these families present the following texts (synthesised from the MSS according to the procedures discussed below in ‘Incompatible texts’):

<i>t</i>	<i>l</i>	<i>2</i>	<i>3</i>
a Accomplishments of the Holy Fathers:	Once, anchorites wise	and holy and spiritual,	<i>bi</i> in number,
b Accomplishments of the Holy Fathers:	Once, the anchorites wise,	holy, spiritual,	<i>vj</i> in number,
i	Once, the anchorites wise,	holy, spiritual,	<i>vj</i> in number,
o Accomplishments of the Holy Fathers:	Once, the anchorites wise,	holy, spiritual,	<i>vj</i> in number,
c Accomplishments of the Holy Fathers:	Once, the anchorites wise,	holy, spiritual,	<i>vj</i> in number,
k	Once, the anchorites holy,	most wise and spiritual,	numbering <i>vj</i> ,
w Accomplishments of the Fathers:	Once, anchorites wise	and holy and spiritual,	loving God in word,

4	5	6
a having gathered all together,	required themselves to say	how each did in his abode
b having gathered together	and required themselves to say	what each did in his abode
i gathered together,	required themselves to say	how each did in his abode
o gathered together	and required themselves to say	what each did in his abode
c having gathered together,	required themselves to say a word,	who did what, each in his abode,
k having gathered together,	asked one another to say	what each had done in his cell
w gathered together	and required themselves to say	what each did in his abode
7		
a and which spiritual abstinence he practised.		
b and which spiritual abstinence he practised.		
i and which spiritual abstinence he practised.		
o and which spiritual abstinence he practised.		
c and which spiritual abstinence he practised.		
k and which spiritual abstinence he mastered.		
w and which spiritual abstinence he practised.		

2. The Treatise *On the Letters*

The treatise was composed before 921 in Old Church Slavic, partly on the basis of Greek grammatical treatises and scholia, as an exposition of the origin and the dignity of the Slavic alphabet and the Slavic Bible translation. Of this brief treatise, seven families of texts are known to me, apart from four abridgements, to which I will only refer in passing:

8	MSS (14-19th c.)	South and East Slavic:	text family α
2	MSS (15-17th c.)	East Slavic:	text family β
10	MSS (15-17th c.)	East Slavic:	text family δ
53	MSS (16-18th c.)	East Slavic:	text family ϵ
1	MS (15th c.)	South Slavic:	text family s
3	MSS (16-17th c.)	East Slavic:	text family r
2	MSS (13-16th c.)	South Slavic:	text family b

Excluded from consideration are 5 contaminated MSS (16-18th c.); included is the single MS of text family s , customarily considered contaminated (it presents readings of the four text families $\alpha\beta\delta\epsilon$), because the existence of contaminations of four exemplars in Slavic is unattested before the dawn of modern textual criticism.

At the beginning of chapter 8 (cf. the Slavic texts in Džambeluka-Kossova 1980), these families present the following texts (synthesised from

the MSS according to the procedures discussed below in 'Incompatible texts'):

1	2	3
α Now others say:	So what for are Slavic Scriptures?	For neither has God made it,
β Now the others say:	What for are Slavic Scriptures?	For neither has God made it,
δ Now there are some that say:	What for are Slavic letters?	For neither did God make them,
ε There are some that say:	What for the Slavic letters?	For neither God made them,
σ Now the others say:	So what for are Slavic letters?	For neither has God not made it,
ρ Now others say:	So what for is the Slavic Scripture?	For neither has God made it
β For the Greeks say:	The Slavic Scriptures have been taken	And make them did neither God from our Scriptures.
4	5	6
α nor even angels,	nor are they original,	like Hebrew and Roman and Hellene,
β nor apostles it,	nor are they from the outset,	like Hebrew and Roman and Hellene
δ nor angels,	nor are they original,	like Hebrew and Hellene
ε nor an angel,	nor are they canonical,	like the Hebrew and the Hellene
σ nor even apostles,	nor are they primary,	like Hebrew and Roman and Hellene
ρ	from the outset,	but Hebrew, Hellene, Roman
β nor angels	from the outset,	like the Hebrew and the Greek
7		
α which are since the outset and are accepted by God.		
β are from the outset and are accepted by God.		
δ are from the beginning and are accepted by God.		
ε which are from the Law and acceptable by God.		
σ which are since the outset and are accepted by God.		
ρ are from the outset and are accepted by God.		
β but which are from the outset pleasing unto God.		

3. Incompatible Texts

The family texts given above are synthesised from the MSS by strict application of no more than one canon of textual criticism (or 'textology', as the discipline is called in Slavic studies since the works of B.V. Tomaševskij of the 1920's): *utrum in alterum abiturum erat?* The application of this canon appears to be quite sufficient to guarantee a full reconstruction of the respective text families down to the details of their morphology and orthography through an argued choice from among the transmitted variants. Unfortunately, such features, which are distinctive for

each of the text families involved, are largely lost in my translation, but enough remains to demonstrate the problem of incompatibility.

This problem arises when attempting to pass beyond the extant MSS and proceed to the reconstruction of the respective archetypes. It appears that a common archetype must be postulated in the case of both works; it is vouched for not in the last place by the evident kinship of the text families.

The *Scete Paterikon* (henceforth: *SP*), beyond a doubt originally a fully concordant translation from Greek, provides, thanks to the Greek, the possibility of distinguishing correct from deviant readings and of deciding which of the family texts are inferior to others:

- t* **abiockw** none reads '12 Fathers',
- 1 **biock** < **aw** preserve the indefinite form of the adjectives,
- 2 **biockw** < **a**, preserves the indefinite form of the adjectives (2 **w** has a definite form for 'spiritual') and the conjunction,
- 3 **akw** < **bioc** have the correct word order and Cyrillic numeral '12',
- 4 **aiocw** < **bk** preserve the past active participle (**c** has it in a secondary case form) without addition,
- 5 **bockw** < **ai** preserve the aorist without conjunction or addition (as well as the Graecism 'themselves' for 'each other', innovated in **k**),
- 6 **aick** < **bow** preserve the correct interrogative (as well as the Greek aorist and the unspecific 'abode', both replaced in **k**),
- 7 **k** < all but **k** are concordant in the rendering of the Greek.

All of the text families in turn contain inferior readings; not one of them contains a *textus optimus*. Moreover, the direction from good to inferior reading varies per syntagm. Consequently, none of the family texts, as a whole, can be the source of another one.

In the treatise *On the Letters* (henceforth: *OL*), we have no basis for deciding upon the superiority or inferiority of the readings. Still, on the basis of our knowledge of Old Church Slavic, we can point out some elements that betray a direction of change and that relegate some family texts to an inferior position:

- 1 $\beta\delta sb$ < $\alpha\epsilon r$ preserve the indefinite form of the pronoun
- 2 $\delta\epsilon br$ < $\alpha\beta s$ preserve the recessive form of the interrogative 'what for'
- 3 ϵ < $\alpha\beta\delta sr b$ preserve the correct (non-witness) perfect tense
- 4 βsr < $\alpha\delta\epsilon b$ preserve the non-corrupted 'angels'
- 5 $\delta\epsilon$ < $\alpha\beta sr b$ preserve the gender agreement with 2
- 6 $\delta\epsilon b$ < $\alpha\beta sr$ preserve 'Roman', which definitely formed a part of the inscription on the Cross
- 7 ϵb < $\alpha\beta\delta sr$ preserve the correct past active participle

This would seem to suggest that β , δ , ϵ , s , r and b could have been derived from α independently. Consider, however, the significant obstacles to such a directional interpretation: in 1, $\delta\epsilon$ 'there are some that say' can in no way have arisen directly from $\alpha\beta sr$; in 3, $\alpha\beta sr$ 'it' (repeated in β 4) is evidently inferior to the plural form; in 6 and 7, ϵ 'canonical' and 'from the Law' could only be derived from $\alpha\delta$ 'original' and αs 'since the outset', respectively, not from any other of the variants; and, finally, in 1 $\beta\delta s$, 2 $\alpha\beta\delta s$, and 6 $\alpha\beta\delta sr$ incorrectly use the indefinite forms of adjectives, whereas ϵb alone are consistent in their correct use. So here, too, all of the text families in turn contain inferior readings, not one of them contains a *textus optimus*, and not one of them can be determined as the source of another one.

The nature of the variants adduced in the two specimens is clearly unspectacular, to say the least. Most of them could have arisen quite independently and spontaneously as accidents of any link in the chain of transmission. Yet the fact that they are canonised by the extant MSS over as many as six centuries, gives them the status of distinctive features. It also makes it evident that in both cases, as far as the extant MSS are concerned, we are dealing with closed traditions and, consequently, with differences that are so significant as to constitute a veritable barrier to any further synthesis of the text families.

The further we study the transmitted family texts of both works, the greater their incompatibility becomes. Neither the graphic and orthographic (including the punctuational), nor the morphological, nor the lexico-syntactic, nor even the narrative features of the texts can be explained in as straightforward a manner as before in the study of the MSS that form one family, namely, by familiar defects of the exemplar copied, or misreadings, or misspellings, or even by the conscious intervention of an editor in one or more of the extant texts. Let me add that none of the text families of either work can be proven to bear the imprint of dialectal features; it is only the individual MSS that bear such marks on their graphic and orthographic surface.

In the case of the *SP*, we can still rely on the Greek text to guide us to the Old Church Slavic archetype, but in the case of the *OL*, even that is lacking: the archetype could have been preserved in any of the texts, but possibly in none of them. More importantly, in neither case can we provide a satisfactory explanation for either the majority of the variants, or even the relationship between the conflicting family texts. The canon so successfully applied earlier seems to have lost its validity in all but isolated instances; no longer can it be applied to the entire length of the text. We have met a barrier of high textological relevance.

4. Causes of Incompatibility

An explanation for the incompatibility of Old Church Slavic texts has traditionally been sought only in two factors (cf., for the *OL*, Ziffer 1993), which presuppose both a succession of the text families in time, i.e. overlaying new features upon a pre-existent text, and a relatively slow and steady development of the tradition:

- (1) *extra-textual*: intervention of an editor or (in translated works) another translator, and
- (2) *inter-textual*: contamination of different exemplars in copying,

both of which could have been operative simultaneously.

The tradition of the *SP* does exhibit a single instance of the operation of factor (1) in its text family **k**, which inserts some readings of a younger Greek text type into the extant Old Church Slavic translation, but such readings are few and far between. Likewise, the tradition of the *OL* does exhibit some instances of the operation of factor (2), in that 5 of the 79 MSS studied (their total is reported to be well over 100) are contaminated; yet such a factor evidently played no role in the tradition of the *SP*, although, on the level of the structure of the book, conflation can be observed in the form of anthologies and miscellanea.

Thus, for the seven text families of the *SP* and the *OL* as a whole, these factors fail to provide an answer to the most pressing question, why they should have been operative only before the origin of the extant MSS. Why, indeed, should there have been a radical change in transmission principles after the first centuries of textual tradition? Whereas the extant MSS testify essentially to faithful copying in a closed tradition, could the period prior to their appearance have been marked by spontaneous and rather massive intervention in a limited number of extant copies?

Closer examination of the texts reveals that the specimen from the *SP* preserves traces of three different and, in the two positions, incompatible solutions to the problem of rendering the Glagolitic numeral '12' - *bi*, marked as such by a tittle and, therefore, easily confused with the abbreviated adjective 'God's':

- t* **aboc** read it as an epithet and changed it to the less unusual 'holy' : **w** omits it : **ik** omit the entire unusual title,
- 3 **a** preserves the Glagolitic numeral : **biock** replace it with the corresponding Cyrillic numeral : **w** reads it as an epithet together with the first two letters of the instrumental 'in number' and interprets it as an abbreviation of the adjective 'God-loving', while reinterpreting the remainder of the word 'number' as a haplographic rendering of the instrumental 'in word'.

In the *OL*, similar traces are evident in the different and, in the two positions, incompatible misreadings of the Glagolitic letters *ĩ*, which in its earliest form can easily be confused with *o*; *x*, which in its earliest form, can easily be confused with the Glagolitic letter *g*; and *g'*, which in its earliest form, can easily be confused with the Glagolitic letter *p*:

- 3 pronominal form **δeb** plural (reading *ixĩ*) : **αβsr** singular (reading *ogo*),
- 4 **αδeb** 'angel' (reading *ag'l*) : **βs** 'apostle' (reading *apl*) : **r** omitted.

So much for the brief specimens given. Both *SP* and *OL* contain, of course, more ample evidence of individual and incompatible solutions to the problems of (a) reading a text written in Glagolitic script in an artificial language without a formal grammar or lexicon to support its implementation, and (b) rendering the same text in the different Cyrillic script. This evidence, among other things, also proves the independence of the *SL* text families **oc** and the *OL* text families **δε**, which could be in doubt on the basis of the specimens alone.

The barrier we encountered can now be defined as the 'Glagolitic barrier', and we must allow for a third factor of textual change, operative only as long as Glagolitic originals were in circulation and leading to a relatively quick development of the tradition:

- (3) *intra-textual*: accident of transcription from the older Glagolitic into the newer Cyrillic script.

This factor was operative in the 10th century, perhaps starting as early as 893 and even attested as late as 1047, extending over an area from Kičevo in Macedonia (the *Kičevo Triodium* of the mid 11th century) to as far north

as Novgorod in Russia (the *Sixteen Prophets with Commentary* of the priest Upry' Lixyj of 1047), but, in all probability, concentrated in the area where the Cyrillic alphabet originated, namely, in Bulgaria.

5. Textual Palaeontology

By relating the incompatibility of the text families to the accident of transcription from Glagolitic into Cyrillic, we can confidently equate them with just as many independent Cyrillic hyparchetypes of both works in question and thereby extend their closed tradition right up to this barrier. Beyond the 'Glagolitic barrier' we enter the realm of 'textual palaeontology', where no MSS are extant, where the single most reliable canon of textual criticism loses its validity, where other, less reliable canons must be called upon to fill the void, and where *divinatio* reigns. Let us see whether we can find at least some clues to chart this, even to Slavists, *terra incognita*, and whether we can specify the operation of the *intra-textual* factor (3) invoked to explain all the textual variety and incompatibility of the hyparchetypal texts.

Here, once more, the *SP* offers the clearest view of the circumstances of its transcription into Cyrillic, thanks to the fact that we can be reasonably certain of its archetypal text and that we can retrace the codicological history of the Glagolitic original in the relationship of the structural variations of the Cyrillic hyparchetypes. We can easily distinguish three successive stages of deterioration of a Glagolitic codex, which must have been one and the same, probably the autograph of the translator, St. Methodius, as all hyparchetypes individually transmit not only inferior readings, but also archetypal readings superior to those preserved in other hyparchetypes:

Stage 1

The codex contains 35 chapters in the order:

[1] A,1,B (ca.5.5 quires)	[9] 22 (ca.1.5 quire)
[2] C,D,E (ca.2 folia)	[10] H (ca.1 folium)
[3] G (ca.1 folium)	[11] I (ca.1 quire)
[4] 2-4 (ca.4 quires)	[12] J (ca.1 quire)
[5] 5-9 (ca.6 quires)	[13] K (ca.1 folium)
[6] 10-19 (ca.14 quires)	[14] L (ca.1 quire)
[7] 20 (ca.1.5 quire)	[15] M (ca.2 folia)
[8] 21 (ca.1 folium)	[16] O (ca.2 quires)

= hyparchetype **a**

Stage 2

Parts [1], [3], [15] and [16] are preserved, but mobile. This stage has two collateral forms:

Stage 2a

Parts [1]-[2] and [4]-[5] are interchanged, chapter F (ca. 1 folium) is inserted between [2] and [6], chapter N (ca. 1 quire) is inserted between [15] and [16]:

= hyparchetype **b**

Stage 2b

Parts [2], [12] and [14] are lacking, [16]+[1] are inserted after [9]:

= hyparchetype **i**

Stage 3

Parts [1]-[3] and [15]-[16] are lacking

= hyparchetype **o**

This stage has two collateral forms:

Stage 3a

Parts [11]-[12] and [13] are interchanged, loss of substance is evident in [4]-[6] (partly compensated by membra disjecta) and [8], marginalia are integrated into the text:

= hyparchetype **c**

Stage 3b

Parts [13]-[14] and the membra disjecta of [4]-[6] lacking, but traces of texts from parts [1] and [4]-[5] are extant.

= hyparchetype **kw**

The picture that emerges, is one of a Glagolitic codex taken out of its binding for easier transcription and then handed down by the first to subsequent transcribers, who seem to have individually muddled the

sequence of the quires. The transcriptions were, of course, made at the repository of the Glagolitic codex; the marked loss of substance between stages 2 and 3 could point either to transport damage or to negligent storage of a Glagolitic codex already laid *ad acta* after transcription. The mutual linguistic and textual affinity of the hyparchetypes **abio**, on the one hand, and **kw**, on the other, as well as the eccentric position of **c** could point to the fact that (a) they were produced in three independent successive rounds of copying and (b) that in the first and the last rounds the copyists were in contact with each other and had the possibility of asking advice of a predecessor, probably in the same scriptorium. Most important, however, is that we are faced with the evidence of seven successive transcriptions of one and the same work from Glagolitic into Cyrillic, the differences of which can be ascribed neither to the *extra-textual* factor (1) nor to the *inter-textual* factor (2), but only to the differing competence and performance of seven different copyists. What we have are, in fact, seven *idiolectal* versions of one and the same archetypal text.

As far as the *OL* is concerned, we lack such codicological data. We are simply left with the evidence of seven incompatible Cyrillic hyparchetypes, independently transcribed from one (or more ?) Glagolitic originals. We may surmise that they were made in succession, like the transcriptions of the *SP*, but we cannot be certain of the sequence. What we can be certain of is an important additional feature of the crossing of the 'Glagolitic barrier'.

The seven hyparchetypes of the *OL* are concordant in stating the number of letters in the alphabet to be 38. Though this may be true for the Cyrillic alphabet of the time, it most certainly is not true for the Glagolitic alphabet for which the *OL* was originally written (cf. Trubetzkoy 1968; Vyncke and Detrez 1992): the alphabet acrostich, originally written in Glagolitic as a preface to the *Didactic Gospels* by the priest Constantine (10th century), contains no more than 36 letters. This, too, is the number of letters transmitted by two of the abridgements of the *OL*, made independently from the seven hyparchetypes of the full text; the other two abridgements, independent as well, give the numbers 37 and 42, respectively. To my mind, this testifies to concerted *updating* of the text to fit the newer Cyrillic alphabet and points to a significant measure of coordination and organisation in the transcription, for we can distinguish hyparchetypal texts produced within the reach of its control and regulation from those that were not affected by it.

It is important to note that this feature of organised updating somewhat reduces the chances that 'textual palaeontology' should succeed in attaining a single archetypal text. Unless we are able to refine the tools of textual criticism to include the full scope of possible accidents both in the transition

from Glagolitic to Cyrillic script and within the no longer extant Glagolitic MSS themselves, we may, in the case of the *OL*, have to content ourselves with seven incompatible hyparchetypes, which can be reduced to a single archetypal reading only in isolated instances, but not over their entire length. We can, of course, also be satisfied with a consensus text, agreed upon by researchers, but not supported by the texts themselves, as we have done up to now.

6. Multiple Transcription and Many-Pronged Stemmata

What we have is indubitable evidence of two works, originally written in Glagolitic script, which were subjected to multiple transcription into Cyrillic. These multiple transcriptions were evidently performed by copyists of unequal competence, who inevitably produced incompatible idiolectal versions of the original. Their performance obviously also led to the deterioration and eventual disappearance of the Glagolitic original; only their Cyrillic versions survive in the form of the fully reconstructible hyparchetypes of the extant MSS. This, of course, ties the transcriptions intimately to the limited life span of a parchment codex taken from its binding and of one single parchment quire, both intensively handled in copying, and would seem to limit the period of transcription to significantly less than the century and a half indicated above. Researchers of parchment MSS, accustomed to reckoning time in centuries, would consider these transcriptions practically as simultaneous.

We also have evidence that the crossing of the 'Glagolitic barrier' was a concerted campaign in which as many as four copyists in one place could have been engaged simultaneously, and in which as many as seven copyists could have been brought to observe the same conventions in removing obsolete realia and updating the archetypal text. In this campaign, the fundamental fork in the tradition of any of the texts involved must have had multiple prongs; that of the two works in question has at least seven, each of which was faithfully preserved thereafter in a closed tradition. What the tradition may have been before the 'Glagolitic barrier' was consciously and concertedly crossed by the scribes, the texts and their readers, is open to speculation for want of MSS. I should not be surprised, considering the general cultural conditions in the *Slavia Orthodoxa*, if it were the same as thereafter. After all, except for the updating, the intention of the campaign appears not to have been the alteration of texts, but their faithful preservation for posterity, only in a different guise; all else may simply have been incidental.

References

- CPG 1974-1987 = *Clavis Patrum Graecorum*, 1-5, ed. by M. Geerard. Turnhout.
- Džambeluka-Kossova, A. 1980. *Černorizec Xrabăr. O Pismenex* (with Italian translation of the reconstructed text). Sofia.
- ODB 1991 = *The Oxford Dictionary of Byzantium*, 1-3. Oxford.
- Schaeken, J. 1992. "Het ontstaan van het Cyrillische alfabet en zijn verdere ontwikkeling in het Russisch". *Cyrillische Letterkunde*, ed. by M. Delemarre *et al.*, 13-37. Rotterdam.
- Tachiaos, A.-E.N. 1989. *Cyril and Methodius of Thessalonica. The Acculturation of the Slavs*. Thessaloniki.
- Thomson, F.J. 1978. "The Nature of the Reception of Christian Byzantine Culture in Russia in the Tenth to Thirteenth Centuries and Its Implications for Russian Culture". *Slavica Gandensia*, 5.107-139.
- Trubetzkoy, N.S. 1968. *Altkirchenslavische Grammatik*, 2nd ed., 22. Wien.
- Veder, W.R. 1994a. *The Edificatory Prose of Kievan Rus'*. (= *Harvard Library of Early Ukrainian Literature, English Translations*, 6). Cambridge, MA.
- Veder, W.R. 1994b. "Die Collatio XII Anachoretarum. Zur Rekonstruktion einer Methodianischen Übersetzung und ihrer Überlieferung ca. 884-1050". *Anzeiger für slavische Philologie*, 22,2.177-197.
- Veder, W.R. 1994c. "One Translation - Many Transcriptions". *Studies in Slavic and General Linguistics*, 22.433-465.
- Vyncke, F. and R. Detrez R. 1992. "De l'origine et de la structure de l'alphabet glagolitique". *Orientalia Lovaniensia Periodica*, 23.219-250.
- Ziffer, G. 1993. "Sul testo e la tradizione dell' 'Apologia' di Chrabr". *Annali dell'Istituto Orientale di Napoli: Slavistica*, 1.65-95.

PART III

Applications to Manuscript Traditions

Which Variants are Useful in Discovering the Deep Structure of the Manuscript Tradition of a Text?

Contra a so-called Essentially Quantitative Approach

Karl-Heinz Uthemann
Department of Classical Languages
Vrije Universiteit, Amsterdam, NL

In this article, the dialogue with the team of A. Dees which was initiated in 1987 will be continued. The present proposal is in essence the subject of a paper I delivered at the workshop organised by P. van Reenen and M. van Mulken in April 1994; however, instead of enumerating editions belonging to Greek and Latin patristics (ancient Christian literature) as examples to verify my statements, as I did at the workshop, I will now analyse only one edition in detail. In this way I will try to avoid any misunderstanding caused by empty terms and definitions without illustrative material, for as Kant said: "Begriffe ohne Anschauung sind leer."

1. The Main Difference: The Missing Qualitative Evaluation

To begin with, I would like to reaffirm my approval of the terminology of the Three Level Method and particularly of the idea of quadruples, i.e., Dees' "viertallen" (Uthemann 1988; 1994). The list of quadruples is an efficient instrument for discerning problems in the manuscript tradition of a text. A difference, however, arises between Dees' approach and my own concerning the question of an essentially quantitative approach in discovering the deep structure(s) of a manuscript tradition, the so-called first level. If one were to qualify the approach by saying that the application of the method to this first level, as proposed by Dees, cannot be purely quantitative, I would agree, as does van Mulken (1993:49) who argues for a "combination with a qualitative follow-up" and for trying to avoid smoothing down "the contingent irregularities as any purely statistical method would do" (van Mulken 1993:222). Though these two statements mark

progress in the dialogue thus far, no more than a theoretical claim has been presented. For after having applied the method through a list of quadruples, Dees and van Mulken do not move on to a procedure of evaluation and elimination. What I have missed thus far are results based on the quantitative approach applied to complicated manuscript traditions. The contaminated tradition which I published in 1988 is a relatively simple case, because the source of the contaminated tradition, the contaminated manuscript, still exists. The solution first published, together with the text edition (1985), was discovered in 1980 without the use of a quantitative method, and in 1987/88 this solution was confirmed by the application of the computer programme developed by Dees' team, i.e., a programme based on the list of quadruples.

An important hinderance in our discussion is formed by the stemmata proposed by the so-called Lachmannians and based on philological evaluation of the variants. Dees found these to be unconvincing because all contradictory evidence was eliminated by reference to philological judgements. As Dees and van Mulken argue, this procedure seems to be nothing more than a kind of self-fulfilling prophecy. In order to avoid all subjectivity, Dees relies only on the quantitative approach based on the list of quadruples. A quantitative approach, however, has to fight the problem of the "noise" which inconsistent variants produce. Consequently an editor has to explain, i.e., to eliminate, all the variants producing "noise", since they suggest dichotomies which do not reflect the actual deep structure(s) and the actual underlying structure(s). Although Van Mulken claims a "combination with a qualitative follow-up" and Dees accepts this claim, I have seen no one of Dees' team proceeding with this qualitative evaluation. Instead there are continual applications of purely quantitative methods to overcome the "noise" and to discover the structure(s).

2. The Three Level Method: some *pros* and *cons*

It is unfair to challenge the so-called Lachmannian method by adducing falsified stemmata, i.e., by referring to examples of a bad use of this method. Instead, one should look at editions with acceptable results and compare the method applied by the editor. One should study the descriptions of the Lachmannian method, e.g., the description of P. Maas (1950; ³1957), against the background of a broad knowledge of acceptable editions. An acceptable edition is one in which the critical apparatus indicates only few inconsistent variants, i.e., identical readings belonging to different groups, or, in other

words, genealogically independent and, therefore, coincidental readings. Based on such research, I discovered that the difference between the Lachmannian method and the Three Level Method is primarily one of terminology, of greater precision and of attitude. The latter is most important because the approach to the process of understanding a text is dependent on it. Because an explanation of this statement would require many pages, I merely state it without further explanation. With regard to the matter of terminology, it is evident that in order to understand terms, one has to consider what they mean and how they are applied.

Let us consider, for example, the second level of the Three Level Method, the investigation of intermediarity based on the knowledge of the actual deep structure (first level). In Dees' terminology the result would be the so-called underlying structure (second level). Generally, this investigation does not generate problems, at least not in the field of classical and patristic texts. While collating the manuscripts, it is immediately obvious whether or not a manuscript A is a copy of a manuscript B. In the context of an edition concerned with the *constitutio textus* only, "being a copy" does not mean that we know by some evidence that A really was copied from B, but only that A could have been copied from B, i.e., that A (1) is younger than B, (2) has all variants of B and (3) will likely have some readings distinguishing it from B and all other manuscripts. Therefore, having collated all (!) manuscripts, the editor is acquainted with all intermediary ones and he eliminates all copies which are not useful in constituting the oldest form of the text attainable by manuscript tradition. However, if in order to preserve all the information of all the manuscripts until an acceptable deep structure has been discovered, a researcher can store the information in the computer and delay answering the question of intermediarity until the deep structure has been established, he then is obliged to distinguish the actual deep structure from the underlying one, a distinction which a Lachmannian need not make.

An important point of divergence between Dees and the Lachmannians seems to be the use of the term "error": Dees avoids using the terms "error" or "mistake" during the first and second step of investigation. "Error" seems to imply the subjectivity of the editor. Granted that the question of whether the term "error" is useful or not is not merely a theoretical dispute (Uthemann 1989b), if the necessity of distinguishing between variants which are useful for the discovery of the actual deep structure and those which are not is accepted, i.e., if one accepts an evaluation of variants before listing the selected quadruples, then it is possible to deny having implemented the terms "error" or "mistake" as qualitative labels, as we will see below.

Undeniably, the distinction between a structural analysis and the historical approach, in the terminology of the Three Level Method, the distinction between actual deep and underlying structure on the one hand and stemma as (historically) oriented structure on the other, indicates progress in precision. It is necessary to distinguish these two phases as methodologically different approaches. Nonetheless, when applying the method, or as the Lachmannian would say, when one attempts to discover the families of "Variantenträger" and, afterwards, the hyparchetypi and the archetypus or perhaps the archetypi, *and* if one accepts the necessity of evaluation, it is not possible to establish the actual deep and underlying structure(s) without a so-called hermeneutical circle. For in evaluating a single reading, the consequences for the whole deep structure must be anticipated, both for the deep structure itself as well as for the orientation. At any step of the evaluation one must have in view the original text, which is to be discovered or reconstructed. Each reading implies the whole of all readings. An evaluation of a subsequent reading will be a re-evaluation of all previous readings. Handling the evaluation in this way induces a progressively better understanding of the text: "mistakes" and above all readings only once introduced in the process of manuscript transmission, i.e., the readings necessary in order to establish the actual deep structure, are discovered.

This procedure of evaluation is, thus, a process by many steps aimed at preventing the third level of the Three Level Method, i.e., the orientation of the stemma, from being nothing more than a self-fulfilling prophecy. Nonetheless, this process cannot avoid being a hermeneutical circle which intends to understand the whole of a text by progressive steps. While discussing which variants are useful for reconstructing the actual deep structure, the question of orientation presents itself as soon as we begin to evaluate. We have to consider the consequences of each separate evaluation for its effect on the whole, i.e. in order to understand the whole of the text. A more precise view of the whole is our instrument for discovering the "good readings" indicating the deep structure and to prove that indeed they are the "good readings". This is the Lachmannian logic of discovery. After having established the actual deep structure, the question of orientation may be asked, i.e., the determination of which variants are probably original readings. This, however, is done in order to discuss the consequences for the actual deep (underlying) structure. Let us hope that this verifying procedure will be a confirmation of the evaluation. Let us hope that not too many "inconsistent readings" will be found, i.e., readings introduced more than once by different copyists during the manuscript transmission. As can be

seen, the main problem of the Lachmannian method is to develop criteria defining the "good readings" in order to discover the actual deep structure.

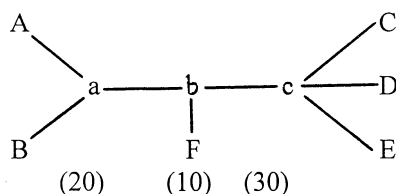
3. *Status quaestionis* - Illustrated by an Fictional Example

That the Lachmannian method of evaluation is unavoidable can be shown by recalling the procedure for transforming variants into quadruples which enables the discovery of quantitatively determined dichotomies as basic elements of actual deep structures. If the relationships are not complicated by contamination or successive changes in relationship or by any other factor producing equivocal quadruples, one gets a "clear" list of quadruples. Such manuscript traditions give no problems for constituting the stemmata. Let us choose as an example six manuscripts A, B, C, D, E and F with the following list of (unequivocal) quadruples:

		the quadruples without "noise"				the quadruples of the fictional example incl. "fictional noise"		
1.	AB/CD	60	0	0		65	7	12
2.	AB/CE	60	0	0		57	8	5
3.	AB/CF	30	0	0		29	8	9
4.	AB/DE	60	0	0		72	5	3
5.	AB/DF	30	0	0		33	5	6
6.	AB/EF	30	0	0		41	12	9
7.	AC/DE	0	0	0		7	6	3
8.	AC/DF	0	0	40		12	7	49
9.	AC/EF	0	0	40		9	13	38
10.	AD/EF	0	0	40		11	7	45
11.	BC/DE	0	0	0		7	5	11
12.	BC/DF	0	0	40		6	10	49
13.	BC/EF	0	0	40		9	12	52
14.	BD/EF	0	0	40		3	7	36
15.	CD/EF	0	0	0		11	2	9

The zero indicates that there is no dichotomic relationship. The first quadruple, for example, shows that AC and BD (or AD and BC) are not opposed as belonging to different genealogical groups. The quantitative distance indicates that AB and CDE form groups (a,c) and that F has a position between both groups dependent on a point "b". As there are three quadruples where F and one manuscript of "c" have a quantitative distance to AB (= a),

F clearly forms a group with CDE (= c). There are, however, also six quadruples where F and a manuscript of "a" have a quantitative distance to two manuscripts of the group "c". Therefore, the position of F has to be between both groups. Considering the numbers, we see that "a" is characterised by twenty readings, "b" by ten and "c" by thirty. This can be illustrated in following way:



A second column of fictional numbers indicates the same relationships, but then with noise induced by genealogically independent coincidences in the manuscript tradition. If these actual quadruples correspond to the first column of quadruples, i.e., represent the same relationships, there will be a reasonable percentage of "inconsistent variants". What, however, is an acceptable percentage of inconsistencies in a manuscript tradition without qualitative evaluation? I will postpone the discussion of an answer to this question until the end.

Contamination creates entanglement. To illustrate this we suppose that a copyist has compared A with C and has changed in A eight readings common to A and B and introducing sixteen readings of C, i.e., five which are common to CDE and eleven which only occur in C. Let us consider the first and second quadruples: instead of "60 0 0", we now have "26 24 5", provided B has been compared with F so that five readings of F entered B and six readings of "a" have been changed which were not altered in A by contamination. The first and second quadruples, therefore, are changed to "20 24 5", the third to "11 5 5", etc. When the normal noise provoked by coincidences is added, it will be difficult to discover a deep structure, and there will be a great number of variants which at first sight do not fit the structure. A good example is the list of quadruples M. van Mulken (1993:66) set up for the manuscript tradition of the "Perceval" of Chrétien of Troyes:

al/fm	253	55	65
ap/fm	213	61	72
fm/pt	207	48	57
fm/ps	229	79	57
fm/lp	170	50	51
fm/pu	224	63	70

Choosing the best supported dichotomies (scil. al/fm; ap/fm etc.) “which presumably represent the general tendencies” (p. 68), she has constructed a tree which is, as she says, “far from reliable, since no less than 87% of the variants must be rejected”, i.e., of the variants which are listed as quadruples. Only 340 of the total 2615 initial variants “do not conflict these dichotomies”, i.e., are the best supported dichotomies. It is thus necessary to try other methods, as van Mulken did. Though she revealed the problems, she did not discover a solution: “The stemmatological investigations show that the family relations among manuscripts are far more complex than we are capable of imagining with our modern concepts” (p. 221). She proposed to combine “both quantitative and qualitative evaluations of the information contained in the list of variants” (p. 221). This consequence is good and even necessary; just how necessary it is, I will explain by looking at a Lachmannian edition.

4. A Test of an Edition as Paradigm

Modern editors following the Lachmannian method often give long lists of variants to convince their readers that they have proved the proposed stemma. They do not discuss the variants in order to evaluate them. Instead of listing such editions, I will here analyse one edition in order to give a paradigm. I have chosen a patristic text (CPG 2246) edited by Henric Nordberg (1962). Some years ago, I evaluated this edition and other editions of texts transmitted by the Ottobonianus gr. 14 (cf. Uthemann 1989a, also for other examples).

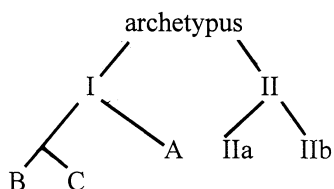
4.1. *The editor's stemma*

The text is a homily of Athanasius Alexandrinus (who died 373 A.D.) transmitted by two different types of manuscripts. On the one hand, we have three liturgical manuscripts containing homilies of various preachers. Nordberg lists twenty-three variants and states “the errors found in” these manuscripts (A, B, C) “demonstrate the unity of the group” (p. 11*). On the

other hand, we have eleven manuscripts, the so-called main group, which transmit collections (Corpora) of Athanasian and Pseudo-Athanasian treatises including some homilies. Nordberg lists "far more numerous common errors" (p. 11*: 85 variants), and he states: "It is evident from these lists of errors that the text has been far better preserved in the" liturgical manuscripts than in the Corpora tradition (p. 13*). He then assumes a connection between the Ottobonianus gr. 14, let us say A, and the main group because they have eight "errors in common" not supported by B and C, and concludes that there have been two common ancestors, one for B and C, and one for A and the main group. Furthermore, he distinguishes two groups of Corpora manuscripts. The first is constituted by five "common errors", the second by "a few more", i.e., by sixteen (pp. 14* - 15*).

I (= ABC)	II (= main group)	IIa	IIb	(A/II)#BC
23 "errors"	85 "errors"	5	16	8 "errors"

How does Nordberg reconcile the first statement (I= ABC) with the statement "A/II # BC"? Though he does not discuss the problem, he give the following stemma (35*):



4.2. Distinguishing types of variants

Here we will not discuss the notion of "common error". We will only evaluate the list of variants adduced by Nordberg to establish the groups. "Evaluating" means to accord to all variants a weight by combining two perspectives, the text-internal perspective, which depends on the text-internal change of content, and the perspective focussed on the copyist and his psychology. The best argument will be the discovery of tendencies linking the common variants of manuscripts (Uthemann 1990), the second best a variant which is a single change in content without being in opposition to the whole context and, therefore, without any text-internal necessity for

making improvements. All variants which can be interpreted as intentional¹ could have been introduced into the text by different copyists independently, and, therefore, have to be eliminated by the process of evaluation. Furthermore, all variants which can be explained by the context, or a paleographical argument, the codicology, the history of mentality, etc., will have little genealogical weight because the changes have occurred independently more than once during the process of manuscript transmission.

By evaluating variants, we try to discover two kinds of changes. The first group consists of variants separating manuscripts, i.e., of variants the philologist qualifies as changes which probably will not occur a second time within the same context. It is possible that these are historical entities which have come into being only once, readings which no copyist would "correct" or alter deliberately. If such variants (*lectiones separativae*) occur in several manuscripts, these manuscripts in principle are regarded as a group or family.

The second category of variants consists of changes which only connect manuscripts (*lectiones distinctivae vel coniunctivae*). They are defined negatively: although there are reasons to admit a genealogical connexion, because there seems to be no intentional or unconscious reaction to text or script, the variants cannot be qualified as *lectiones separativae*; the philologist has no strong arguments to exclude the possibility that (1) there are other manuscripts which belong to the same group but no longer transmit the connecting variant, and that (2) the same variant occurs in more than one group.

4.3. *Application to the paradigm*

I have argued (Uthemann 1989a:303-307) that no variant in the list of twenty-three variants which are common to the liturgical manuscripts (A, B and C) separates these manuscripts as a group from the archetype. In principle, they could all be the original reading, and, therefore, A, B and C would be three copies of the archetype. This, however, does not exclude a dichotomy "ABC/main group" defining the actual deep structure; therefore, we have to reconsider the lists of variants.

¹ In my view, a first step towards acknowledging the importance of the elimination of all intentional variants in order to find out the deep structure is the discussion of a dialectological approach by van Mulken. For, if you eliminate one sort of intentional variation, why not eliminate all sorts of intentional variation?

Let us start with the eight variants which Nordberg has interpreted as errors indicating a common ancestor of A and the main group. In principle, they all make good sense, each is a correct text, although the common readings of B and C are “better”.² In other words, as far as the eight variants are concerned, the readings of BC correspond to the tendency of the long list of individual variants of B and C in the critical apparatus. The copyist of the ancestor of B and C had the intention of “improving the text”. None of these eight variants, however, “has the weight” to distinguish groups of manuscripts or even to connect them.

The unity of the main group can be demonstrated, although most of the eighty-five variants listed by Nordberg prove absolutely nothing. At first sight, we ascertain a general deterioration of the text. Some readings of the main group, however, may be archetypal or even original. Comparing, for example, all passages where Matt. 21,2 is quoted, we should at least say that the archetype transmitted “Poreuthentes” instead of “Poreuesthe” (1,2; 5,9; 7,1; 14,11; Uthemann 1989a:306). Some variants can only be interpreted as connecting ones. In 4,7-8, for example, the quotation of Matt. 10,39 is missing in the manuscripts of the Athanasian Corpora. Because the text which follows presupposes a similar quotation, it could have been introduced by the copyist of the ancestor of the liturgical tradition. Vice versa, in 13,3-4 one misses a necessary quotation of a *logion* of Christ in A, B and C, whereas the Corpora manuscripts cite Joh. 5,39 which may be a second-hand improvement dependent on 13,7-8.³ Some variants prove nothing at all, some may perhaps connect manuscripts. More or less obvious errors challenge the intelligent copyist to improve the text, and there is no criterion

² Two variants of B and C listed by Nordberg are redundant (2,9; 18,6). The missing word in 4,3 is a *homoioteleuton* and, therefore, it cannot connect manuscripts genealogically. The article in 13,2 and 15,6 can be the original reading, but precisely this article is also missing (1) in all manuscripts except B in 2,3-4 and 16,18 and (2) in all manuscripts except A, B and C in 11,5. Of course, in 15,8 and 16,20 the text of B and C is more correct than the text of A and the main group, although their reading is not a mistake.

³ There are more analogous cases: in 5,5, further in 1,4, resp. in 7,6-9, where is quoted Phil. 3,19, resp. Hebr. 12,18.22, and the reading of the main group is wrong, finally, in 7,12 where the editor has preferred *exeblethesan* as reading of Hebr. 11,15 instead of the lectio communis *exebesan* which one finds in the liturgical tradition of the text and which is probably the original reading. Neither 6,1-2 (*ton paronton*] *ton poulou touton*) nor 7,10-11 nor 8,9 nor 15,4 nor the fact that only the liturgical manuscripts contain a smoothing *toutestin* in 2,1 are distinguishing variants. Cf. 3,11 where you find *toutestin* in the same exegetical function.

for deciding whether such an improvement happened only once or twice or even more often.

Are there any readings of “better” quality in the list of eighty-five variants? In 4,13 the variants of both traditions (“apobalesthai” - “apolesthai”) can be accepted in the context, although after the quotation of Eccl. 3,6 in the proem the preacher permanently uses “apollymi”. Has the copyist of the ancestor of A, B and C guessed that “apolesthai” was replaced unconsciously by the more common “apobalesthai”? Was an original “apobalesthai” (unconsciously) changed on account of the context? A more evident situation, i.e., two readings which both are not contested by the context, occurs in 11,14; 12,5-6; 13,12; 15,2-3; 17,11-12 and 18,14-15. Particularly 11,14 is a good example, whether one follows the reading “kechenasi” of the Corpora manuscripts or the simple “echousin kai” of the liturgical manuscripts, which fits the context (Uthemann 1989a). Both readings form an opposition. If “echousin kai” was the original text, probably a change to “kechenasi” happened only once; if “kechenasi” is accepted as original reading, but transmitted, e.g., by an unreadable, faded manuscript, it would be an unlikely coincidence that two or more copyists guessed the reading to be “echousin kai”. In 18,14-15 (Uthemann 1989a:305), only the Corpora manuscripts add a sentence of four words which in principle is redundant because it repeats the idea of the preceding sentence. This redundant sentence, however, does not fit the following word introducing a *logion* of Christ (Joh. 14,31). Yet a convenient sentence of six words is found in the liturgical manuscripts: after having freed the dead from hell (18,12-14), Christ raised “the others” (*tous allous*) who are still alive, saying what he said in Joh. 14,31. Nevertheless, there remains a problem. In the liturgical tradition, the phrase is introduced by a particle read in the following sentence in both traditions. Here, however, we are not yet, or not directly, concerned with the question of orientation. The last-mentioned six variants (*lectiones separativae*) seem to be enough to prove a dichotomy “Liturgical mss (ABC) / Corporal mss”, and here in our context we do not have to decide whether there was redactional work on the level of the ancestor of the liturgical manuscripts.

After having proved this dichotomy with some *lectiones separativae*, one can say that it is unlikely that all of the *lectiones distinctivae* mentioned above, which do not fit the context and are, therefore, obvious errors, occurred more than once by coincidence in the course of manuscript transmission. This argument is only a logical consequence which gains in strength, as all Corpora manuscripts except the Laurentianus IV,23 (G) and the Coislinianus gr. 45 (S) are *codices eliminandi* or copies. G may even be

the *codex unicus* of our text in the Corpora tradition, but we are not concerned with the question of intermediarity. This argument ("it is unlikely that all of the *lectiones distinctivae*", etc.) will not introduce a statistical method because it is not a purely quantitative one; rather, it presupposes an evaluation and elimination of variants. It only argues for using *lectiones distinctivae* insofar as they are errors, i.e., not fitting the context. It states that the groups connected by such readings can be larger than the group of manuscripts actually presenting the error. It does not prove the actual deep structure. If a text tradition contains no *lectio separativa*, we cannot reconstruct an actual deep structure because we have no criterion for distinguishing groups. It is doubtful whether it is possible to define an exact quantitative rate of variants which only connect manuscripts in order to exclude coincidence in such a manuscript transmission. Were that possible, it would still not be likely that genealogical dependance could be fixed.

Finally, Nordberg is wrong in constituting two subgroups of the Corpora tradition. The five variants mentioned prove nothing.⁴

Nordberg listed 108 variants (85 + 23) in order to demonstrate the dichotomy "Liturgical mss / Corpora mss". After evaluation, only six variants remain as *lectiones separativae* which equally fit the context (text-internal perspective) and which lack a mark of intentional or unconscious alteration (perspective focussed on the copyist).

References

- Maas, P. 1950; ³1957. *Textkritik*. Leipzig.
- Mulken, M. van. 1993. *The Manuscript Tradition of the Perceval of Chrétien de Troyes. A stemmatological and dialectological approach*. PhD dissertation, Vrije Universiteit. Amsterdam.
- Nordberg, H. 1962. *Athanasiana. Five Homilies - Expositio fidei - Sermo maior, Part I: The Texts* (= *Societas Scientiarum Fennica. Commentationes Humanarum Litterarum* XXX.2). Helsinki.
- Uthemann, K.-H. 1988. "Ordinateur et Stemmologie. Une constellation contaminée dans une tradition grecque". *Distributions spatiales et temporelles, constellations des manuscrits. Etudes de variation linguistique offertes à Anthonij*

⁴ Regarding the readings in 1,7 and 2,15, even Nordberg admits that they have been corrected by two manuscripts belonging to the group he wants to define by these variants. Differences that are nothing else than orthographical mistakes based on an identical pronunciation prove nothing (6,1-2; 11,13). Finally, it is easy to correct 7,7, because *kai* is *lectio facilior*: *Sion oros kai polis theou* ("the mountain of Sion and the city of God").

- Dees à l'occasion de son 60me anniversaire*, ed. by P. van Reenen and K. van Reenen-Stein. Amsterdam - Philadelphia.
- Uthemann, K.-H. 1989a. "Die Lazarus-Predigt des Leontios von Arabissos (BHG 2219u)". *Byzantion* 59:291-353.
- Uthemann, K.-H. 1989b. "Editionstechnik ohne 'Trenn- und Bindefehler'? Zur Rekonstruktion von Stammbäumen mittels des Computerprogramms von A. Dees". *Jahrbuch der Österreichischen Byzantinistik* 39:49-54.
- Uthemann, K.-H. 1990. "Codex recentior, non deterior? Zur Überlieferung des Hodegos im Codex Vindobonensis theol. gr. 40". *Jahrbuch der Österreichischen Byzantinistik* 40:129-143.
- Uthemann, K.-H. 1994. *Die Pseudo-Chrysostomische Predigt In Baptismum et Tentationem* (BHG 1936m; CPG 4735) (= *Abhandlungen der Heidelberger Akademie der Wissenschaften, Philos.-hist. Kl.*, Jahrgang 1994, 3. Abhandlung). Heidelberg.

From Variant to Pedigree in the *Charroi de Nîmes*

A Typology of Variants

Pieter van Reenen

Department of General Linguistics
Vrije Universiteit, Amsterdam, NL
and

Lene Schøsler

Department of Language and Communication
Odense Universitet, Odense, DK

The analysis of the relations between manuscripts,¹ for the purpose of preparing text editions or for linguistic, literary or historical ends, has been carried out by means of fundamentally different approaches, ranging from Lachmann's method of common error, aimed at detecting the original readings, to Bédier's respect of the variation found in single manuscripts, with little hope of discovering the original. We have put forward our own position on previous occasions (see Schøsler 1988; 1989; 1995) and will not repeat our arguments here.

In this study we want to stress the following point. The choice of method, i.e., Lachmann versus Bédier, is not always a matter of the individual taste of the scholar, for it is also dependent on the type of text tradition under examination. Consequently, the scholar cannot always base his analysis exclusively on the most revealing relations in the manuscripts to the neglect of other information. The most revealing relations between manuscripts are the so-called type-2 oppositions, i.e., oppositions of the type

¹ We follow the traditional, ambiguous use of the term manuscript referring to the text in a volume and to the volume as a whole.

AB/CD, as has been put forward by Greg and Dearing (see Salemans 1990).² An example of a type-2 opposition is given in (1), vs. 14 of the *Charroi de Nîmes*:³

(1)

ms A Ce fu en mai, el novel tens d'esté

ms B Ce fu en mai, el novel tens d'esté

'It was in May, the new summer time'

ms C Che fu en mai, que on dit en esté

ms D Ce fut en mai, que l'an dit en esté

'It was in May, that one says in the summer'

Recently it has been defended, for instance, by Salemans (1990) and Dees (1988a), that no other relations are to be taken into account. This restriction, however, has a serious drawback: not all text traditions offer sufficient type-2 oppositions at all levels, and some text constellations lack type-2 oppositions per definition. The lack of type-2 oppositions leaves us with the choice of either giving up our stemmatological research or accepting the use of other oppositions which form a sufficiently reliable complement. We will illustrate this problem in a general way on the basis of the nine manuscripts of the *Charroi de Nîmes*.⁴ The variants between the four families, A, B, C and D, exhibit considerable differences. Especially one family, D, is so distinct from the other three that it is often difficult to determine whether it has any relationship with the others at all.

The goal of this study is to discuss a set of parameters for registering all kinds of differences among manuscripts, i.e., all types of variants which are stemmatologically relevant. These parameters not only concern type-2 oppo-

² It is essential that the formula contains one slash. On either side of it are at least two elements: AB/CD, ABC/DE, AB/CDE or ACD/BE. See Wattel and van Mulken "Weighted formal support of a pedigree", in this volume.

³ The verse numbers correspond to the edition McMillan 1972.

⁴ The assonanced poem of the *Charroi de Nîmes* is found in the following manuscripts:

A1 Paris, Bibliothèque Nationale, f.fr. 774;

A2 Paris, Bibliothèque Nationale, f.fr. 1449;

A3 Paris, Bibliothèque Nationale, f.fr. 368;

A4 Milano, Biblioteca Trivulziana 1025;

B1 London, British Museum, Royal 20 D XI;

B2 Paris, Bibliothèque Nationale, f.fr. 24369-24470;

C Boulogne-sur-mer, Bibliothèque Municipale, 192;

D Paris, Bibliothèque Nationale, f.fr. 1448;

F (Fragment), Paris, Bibliothèque Nationale, nouv. acq. fr. 934.

sitions but also cover a wide range of variations between manuscripts and manuscript families, including distant ones whose relationships may be hardly detectable. The types of variations may exhibit different qualities: some serve in the first place to build stemmas, others contribute primarily towards establishing the dialect properties of the language in the manuscripts, still others do both.

In section 1 we present and discuss the problem of the type-2 oppositions. In section 2 we discuss a set of external parameters for comparing manuscripts, i.e., parameters concerning the physical appearance and the history of the manuscripts. In section 3 we distinguish a set of internal, i.e., linguistic parameters.

1. Type-2 Oppositions in the Manuscripts of the Four Families of the *Charroi de Nîmes*

As observed above, in the *Charroi de Nîmes* we distinguish four families of manuscripts: A, B, C and D (see Schøsler 1995). Since we have not encountered data indicative of the structure AC/BD, the underlying structure of the families of manuscripts has been determined by means of type-2 oppositions indicating a choice between two out of three possible structures: either AB/CD or AD/BC.

The text of the *Charroi*, comprising about 1500 verses, provides us with a rather short list of type-2 oppositions at family level, as appears from Table 1, in which the manuscripts of the A family (A1, A2, A3, A4 and the fragment F) and the B family (B1, B2) are each considered as one group, unless otherwise mentioned. In producing type-2 oppositions such as AB/CD, we have distinguished differences at four levels, only two of which are directly stemmatologically relevant: levels 3 and 4. The difference between the levels 3 and 4 concerns the degree of relatedness. At level 3 there is no doubt that the passages to be compared are *related*, and within it we note usually a content difference (see section 3.1.3). At level 4 we consider the question whether versions which are different are indeed *related*, and not completely *independent* from each other (see section 3.1.4). (The levels 1 *graphic / phonological* and 2 *morphological* will be discussed in sections 3.1.1 and 3.1.2, respectively.)

Table 1 shows that, in a text of 1500 verses, at level 3 we find only ten clear formulas in favour of AB/CD (or AB1/CD) and five clear formulas in favour of AD/BC (or AD/B2C). If we consider level 4 oppositions, we are left with only three formulas, twice AB/CD and once AD/BC.

Table 1: *The complete list of type-2 oppositions of the manuscript families of the Charroi de Nîmes at content level (level 3) and the level of relatedness of the variants between versions (level 4)*⁵

verse	level 3	level 4
0014	AB/CD	ABCD
0029	AB/C/D	AB/CD
0054	AD/BC	ABCD
0127	AB/CD	ABCD
0161	AB/CD	AB/CD
0246	AB/CD	ABCD
0313	AB/CD	ABCD
0466	AB/CD	ABCD
0483	AB1/CD	ABCD
0509	AD/BC	ABCD
0765	AD/BC	AD/BC
0766	AB/CD	ABCD
0903	AB/CD	ABCD
0903	AD/BC	ABCD
0923	AB/CD	ABCD
1136	AB/CD	ABCD
1234	AD/B2C	ABCD

The difference between the levels 3 and 4 can be illustrated by means of verse 29, see (2) below. That there is a relation between C and D in verse 29 is beyond reasonable doubt at level 3; nonetheless, there is a difference which should be noted at this level: AB/C/D. The presence of *descendi* 'he dismounted' in C and of *est descendus* 'he has dismounted' in D provides the basis for grouping them together, in spite of considerable differences. In A and B, which are virtually identical, no mention is made of 'dismounting'.

⁵ In verse 0483, the formula AB1/CD at level 3 is valid only on the condition that we consider B1 to represent the B family; in verse 1234, the formula AD/B2C is valid only on the condition that we consider B2 to represent the B family.

(2) vs. 0029 AB/CD

A1 Li cuens G[uaillaume]⁶ fu molt gentix et ber

A2 Li quens G[uaillaume] fu molt gentis et ber

A3 lacuna

A4 Li quens Guill[aume] fu molt gentil et ber

B1 Li quens G[uaillaume] fu molt gentiz et ber

B2 lacuna

AB 'The count Guillaume was very noble and brave'

C A son ostel descendi au degre

'He dismounted beside the steps of his home'

D En la grant place est descendus li ber

'At the great square the brave one has dismounted'

A complete list of all verses containing a type-2 opposition at family level is found in Appendix 2. For a discussion of many of these type-2 oppositions, we refer to Schøsler (1995).

Since the AB/CD and the AD/BC formulas cannot both be stemmatologically valid,⁷ either the ten AB/CD or the five AD/CD formulas are due to chance. Closer inspection of the manuscripts has convinced us that only the AD/BC formulas can be explained as due to chance. Consequently, we believe that the type-2 opposition AB/CD provides us with the relevant information concerning the underlying structure of the stemma.

Important as this result may be, at this juncture the question remains as to why the total number of fifteen type-2 oppositions at level 3 is so small, especially since there are many more oppositions of other types at this level. We assume that this has something to do with the distance between the main families within the *Charroi* tradition. Considering the verses 1081-1215, the only passage in which the fragment F is present, we see (cf. table 2) that the co-occurrence of all manuscripts is extremely low.⁸ Many verses or even complete passages are lacking, or have been added by the scribe, in one or more of the four families. As a consequence, we have only 48 verses out of 200 in which A, B, C and D co-occur, i.e., only 48 verses in which all four families can be compared. This provides us a priori with a potential maxi-

⁶ Expansions of abbreviations are given in square brackets.

⁷ Unless there is question of contamination, which is not the case in our text tradition, see Wattel and van Mulken "Weighted formal support of a pedigree", in this volume.

⁸ Table 2 is derived from Appendix 1, columns 1 and 3.

num of only 48 type-2 oppositions, assuming that there is only one opposition per verse which, of course, is not necessarily true (see verse 903 in table 1 above). In fact, at level 3 we find one single genuine type-2 opposition (type AB/CD) in the whole passage, the main reason being that C, and especially D, are often both deviant from A and B and deviant from each other.

Table 2: *Co-occurrence of verses in the four manuscript families of the Charroi de Nîmes (vss. 1081-1215) and their formulas⁹*

The total number of verses (interpolations included) is 200 of which:

152 do not co-occur in all families, i.e.

48 co-occur in all families.

The 48 verses in the families relate as follows (at level 3):

15	ABC/D
12	AB/C/D
7	ABCD
5	ACD/B
4	A/B/C/D
2	ABD/C
1	AB/CD
1	AC/B/D
1	AD/B/C
0	AD/BC, AC/BD, A/BCD, A/BC/D, A/BD/C

The following two facts can be derived from table 2:

(a) the formulas ABC/D and AB/C/D are the most frequent ones (15+12=27 occurrences, i.e. 56%) and

(b) there is only one type-2 opposition: AB/CD (verse 1136).

We venture the following explanation for these two facts. Being written down only later (see Tyssens 1967: Introduction), the *Charroi de Nîmes* was originally an oral legend. It is hardly plausible that A and B go back to different oral traditions, it is even probable that the manuscripts of the AB family have a common written ancestor, possibly different from that of C and D, since there are more similarities between A and B (15+12+7+2+1=37) than between any other pair of families. On the other hand, there is some plausibility that C and D stem from two different oral traditions. In

⁹ The complete list of manuscript variations of the passage are presented in Appendix 1.

particular, the similarities between D and the other families are low (DB: $7+2=9$; DC: $7+5+1=13$; DA: $7+5+2+1=15$, see table 2). Consequently, there may have been two or even three independently written ancestors derived from different oral versions of the same legend. This line of reasoning accounts for the impression we have that the same story is told in all manuscripts, but that it is only in the case of AB that we are able to compare the precise wording of the legend. It also explains why we find so many formulas in which D is lacking and why, among the complete formulas, we encounter quite a number of formulas of the type ABC/D and AB/C/D. We will probably never be able to prove beyond doubt the existence of two or three independently written ancestors, and our explanation will remain open to discussion. Indeed, it can never be excluded that all manuscripts have had one and the same common, rather remote written ancestor (see Régnier 1966:28).

Whatever the family relationship may be, the similarity of A and B has an interesting implication: in terms of type-2 oppositions we have no basis of comparison between AB on the one hand and C and D on the other hand. Since there are rather few and small differences among the four A manuscripts and the F fragment forming the A-family, the two B manuscripts, and even between the A and B families, they can easily be considered as one family with one common ancestor. Traditionally this common ancestor has been called *x* (see Tyssens 1967 and McMillan 1972:20). After having replaced AB by *x*, the formula AB/CD is reduced to *x*/CD, and it has thus become logically impossible to derive type-2 oppositions for this text tradition at family level.

If type-2 oppositions are chosen as the only basis for stemmatological classification, it is by definition impossible to examine family relations of a text tradition with three branches. Since in our view such a restriction is unpalatable, we will carry out an analysis of the relations between manuscripts by using a supplementary set of internal parameters, instead of refraining from further research into the properties of such trees.¹⁰ Besides the internal parameters we will discuss a set of external parameters, which may be relevant for the classification of manuscripts. The latter will be dealt with first.

¹⁰ For more on the role of these internal parameters, see Wattel and van Mulken "Weighted formal support of a pedigree", in this volume.

2. A Set of External Parameters for Comparing the Manuscripts of the *Charroi de Nîmes*

We distinguish five external aspects of manuscripts: the level of codicology (section 2.1), the cyclic level (section 2.2) and mutilations (section 2.3). Two other relevant aspects concern the handwriting (section 2.4) and the relatedness of hands (section 2.5).

2.1. Codicology

The manuscripts of the *Charroi de Nîmes* are found in eight volumes containing collections of Old French literature. We refer to these volumes as A1, A2, A3, A4, B1, B2, C and D. They have been described in Tyssens (1967). F, a fragment mentioned by De Poerck et al. (1970) and McMillan (1972), does not form part of a volume, though it may have done so at some point. A3 and D are mutilated (see Tyssens 1967:361,392).

Tyssens (1967) has provided the most thorough description of this level; we are unable to provide further relevant stemmatological information at this level with respect to the *Charroi de Nîmes*.¹¹

2.2. Cycles

There are eight volumes combining the two cycles of *Guillaume* and *Aimeri* (see table 3). The volumes, however, do not always contain the same number of *chansons de geste* (see table 3), and they are not composed in the same way (see table 4).

At first sight, we might conclude from the tables 3 and 4 that the A and C volumes are related as opposed to the B and D volumes, since the former contain almost the same series of texts and almost exclusively those of the cycle of Guillaume, whereas the latter contain several, but not the same, chansons from the cycle of Aimeri. The cyclic manuscripts of the B and maybe the D families contain the largest number of texts, even without taking into account the possibility that D originally has been longer, as it appears from table 4.

¹¹ For more about this parameter and its relevance for stemmatological investigation, see Voorbij, "Medieval Dossiers and Modern Stemmas. An Exploration of Manuscripts of the *Epistola Alexandri ad Aristotelem*", in this volume.

Table 3: List of 20 *chansons de geste* in eight volumes (A1, A2, A3, A4, B1, B2, C and D) combining two cycles: that of Guillaume (1-10) and that of Aimeri (11-20).¹² The absence of a *chanson* in a volume is marked by the symbol #. Fragments, such as F, are not mentioned.

cyclic manuscripts								<i>chanson de geste</i>
A1	A2	#	A4	B1	B2	C	D	Enfances Guillaume
A1	A2	A3	A4	B1	B2	C	D	Couronnement de Louis
A1	A2	A3	A4	B1	B2	C	D	Charroi de Nîmes
A1	A2	A3	A4	B1	B2	C	D	La Prise d'Orange
A1	A2	A3	A4	B1	B2	C	D	Enfances Vivien
A1	A2	A3	A4	B1	B2	C	D	Chevalerie Vivien
A1	A2	A3	A4	B1	B2	C	D	Aliscans
#	A2	A3	A4	B1	B2	C	D	Bataille Loquifer
A1	#	A3	A4	B1	B2	C	D	Moniage Rainouart
A1	#	A3	A4	B1	B2	C	#	Moniage Guillaume I + II
#	#	#	#	B1	#	#	#	Garin de Monglane
#	#	#	#	B1	#	#	D	Girart de Viane
#	#	#	#	B1	B2	#	D	Aimeri de Narbonne
#	#	#	#	B1	B2	#	#	Narbonnais
#	#	#	#	B1	B2	#	D	Siège de Barbastre
#	#	#	#	B1	B2	#	#	Guibert d'Andrenas
#	#	#	#	#	#	#	D	Prise de Cordres et de Seville
#	#	#	#	B1	B2	#	#	Mort Aimeri
A1	#	#	#	B1	#	C	#	Foucon de Candie
#	#	#	#	#	B2	#	#	Renier

A more careful look at the number and the succession of the texts¹³ in tables 3 and 4 reveals that the opposition AC/BD is far from evident. We consider first the volumes A and C:

- A3 and A4 are in fact the only ones containing the same texts in the same order (we assume that the mutilated part of A3 has had the same beginning as the other A volumes);
- A2 comes close to A3 and A4, but at its end two texts (*Moniage Rainouart* and *Moniage Guillaume II*) are lacking;

¹² This table is based on table 3 in Tyssens 1967.

¹³ We use the word "text" to indicate a single "story", i.e., "récit", in a volume.

Table 4: Succession of the texts in the eight volumes¹⁴

Legend:

#	text not present
mu	mutilation (beginning of ms A3 and end of ms D)
XX/ /XX	text in which one or two other texts are inserted (Na, EV, MR in B1 and B2)
XX- -XX- -XX	texts merged (CL, CN, PO in D)

A1	A2	A3	A4	B1	B2	C	D	
#	#		#	GM	#	#	#	Garin de Monglane
#	#		#	GV	#	#	GV	Girart de Viane
#	#		#	AN	AN	#	AN	Aimeri de Narbonne
#	#		#	Na/	Na/	#	#	Narbonnais
EG	EG		EG	EG	EG	EG	EG	Enfances Guillaume
#	#	mu	#	/Na	/Na	#	#	Narbonnais
CL	CL	CL	CL	CL	CL	CL	CL-	Couronnement de Louis
CN	CN	CN	CN	CN	CN	CN	-CN-	Charroi de Nîmes
PO	PO	PO	PO	PO	PO	PO	-PO	La Prise d'Orange
#	#	#	#	#	EV/	#	#	Enfances Vivien
#	#	#	#	#	SB	#	SB	Siège de Barbastre
#	#	#	#	#	#	#	PC	Prise de Cordres et de Seville
#	#	#	#	#	GA	#	#	Guibert d'Andrenas
EV	EV	EV	EV	EV	/EV	EV	EV	Enfances Vivien
CV	CV	CV	CV	CV	CV	CV	CV	Chevalerie Vivien
AL	AL	AL	AL	AL	AL	AL	AL	Aliscans
FC	#	#	#	#	#	#	#	Foucon de Candie
#	BL	BL	BL	BL	BL	BL	BL	Bataille Loquifer
MR	#	MR	MR	MR	MR/	MR	MR	Moniage Rainouart
#	#	#	#	#	MA	#	mu	Mort Aimeri
#	#	#	#	#	/MR	#		Moniage Rainouart
#	#	#	#	#	Re	#		Renier
#	#	#	#	#	#	FC		Foucon de Candie
#	#	#	#	#	#	MG1		Moniage Guillaume I
MG2	#	MG2	MG2	MG2	MG2	MG2		Moniage Guillaume II
#	#	#	#	SB	#	#		Siège de Barbastre
#	#	#	#	GA	#	#		Guibert d'Andrenas
#	#	#	#	MA	#	#		Mort Aimeri
#	#	#	#	FC	#	#		Foucon de Candie

¹⁴ This table is based on table 4 in Tyssens 1967.

- A1 comes also close to A3 and A4 but contains one text more, *Foucon de Candie*, whereas *Bataille Loquifer* is lacking;
- C also comes close to A1; but it contains one text more, *Moniage de Guillaume I* (C is the only volume containing this text), and *Foucon de Candie* occurs at a different place.

Consequently, the opposition at this level is A3A4/A2/A1/C.

In the volumes B and D we observe that:

- the scribe of B2, as opposed to B1 and D, makes an effort to distribute the texts genealogically and chronologically and cuts off *Enfances Vivien* in order to insert *Siège de Barbastre*, and cuts off *Moniage Rainouart* in order to insert *Mort Aimery* (see Tyssens 1967:367-368).
- B1 and D, as opposed to B2, contain *Girart de Viane*;
- B1 and B2, as opposed to D, share the same text insertion in *Narbonnais*;
- B2 and D, as opposed to B1, have *Siège de Barbastre* at the same place in the sequence of texts;
- D is the only volume containing the *Prise de Cordres et de Seville*; B1 is the only volume containing *Garin de Monglane*; B2 is the only volume containing *Renier*.¹⁵
- D is the only volume which has reduced its version of *Couronnement de Louis*, merging it and *Charroi de Nîmes* into a single text: the scribe of this volume starts with a previously established kernel of the texts, *Couronnement de Louis*, *Charroi de Nîmes*, *Prise d'Orange*, and he inserts this kernel as a whole into the composition by means of transitional *laisses* (Tyssens 1967:447). All volumes, however, have the succession CL-CN-PO.¹⁶

Consequently, at this level we find the following oppositions: B1D/B2, B1B2/D, B2D/B1 and B1/B2/D.¹⁷

In conclusion, at cyclic level many groupings and oppositions can be found. Except for A3 and A4, no volumes have exactly the same distribution. Thus, the information for building stemmatological trees is scarce at this level in the *Charroi de Nîmes*.

¹⁵ D is the only volume containing a transitional text *Département des fils d'Aimerie*.

¹⁶ We do not discuss ms. E since it does not contain *Charroi de Nîmes*.

¹⁷ The preceding observations are limited to the most important oppositions and exclude those considered to be less important. For instance, we have not mentioned the possible role of the *vers orphelin* in the cyclus. For discussion of this point we refer again to Tyssens (1967:163-176).

2.3. *Mutilation*

By mutilation we mean the disappearance of part(s) of a text by external circumstances, with material traces left in the extant copy. We have some important cases of mutilations in the manuscripts of the *Charroi*:

- F is a fragment, corresponding to the vss. 1081 to 1215 in the other manuscripts of the A family;
- one folio of A1, the vss. 717-877, is lacking (cf. McMillan 1972:14);
- the beginning of A3 is mutilated: 263 verses are lacking (cf. Tyssens 1967:361);
- the first 70 verses of B2 are lacking, corresponding to one folio, including the preceding end of the Couronnement de Louis (cf. McMillan 1972:17).

These mutilations have occurred at the level of the *extant* copies which have not been copied *afterwards*.

When mutilations have *preceded* further copying, it is possible that the newly transcribed text no longer exhibits traces of the external aspect of the mutilation. Instead we may come across abrupt transitions, interpolations, lacunas or reshuffling. It seems to us that such cases can better be classified among the internal parameters to be dealt with in section 3.

2.4. *Handwriting*

A last possible external parameter is the hand of the scribe. If two manuscripts are written by the same scribe, information becomes available on the nature of the variants a scribe considers to be acceptable within his idiolect (for details we refer to Busby 1994). The manuscripts of the *Charroi de Nîmes* are all written by different hands.

2.5. *Relatedness of Hands - Atelier*

Tyssens (1967: ch. XIV) has presented arguments suggesting that A1, A2 and A4 have been written at the same moment in the same atelier, one of the arguments being that the hands are related. It is not impossible that this is right at least in the case of A1 and A2, although some arguments may point in another direction. (In the case of A4, McMillan 1972:15 disagrees with Tyssens.) It is, however, interesting to discuss further the question of "same atelier" and the possible implications of such a statement. We do not know much about the functioning of ateliers. Groups of scribes may have worked together with illuminators in ateliers. The scribes of the same atelier

may have come from the same or from different dialectal areas of France. Thus, their copies may or may not differ with respect to dialectal features, but all copies would have the same common source (except when different sources, that is, more models, reveal themselves in a contaminated manuscript).¹⁸ It may also have been the case that scribes from other areas came to an atelier in order to copy a certain, well-known source. Both of these cases could have resulted in copies copied at the same atelier, but possibly differing dialectally for the reasons proposed. A copying activity like the one suggested here would probably not differ greatly from another possible copying activity consisting in a common source being transported from one place to another.

The problem is, as mentioned above, that we have no evidence for any of the hypotheses mentioned. The only thing of which we are certain is that the mss. of the family A have a common ancestor and that those of the family B have another common ancestor, both going back to a common ancestor x. Thus, if Tyssens' hypothesis concerning the same atelier for the whole A-family and another atelier for the B-family (Tyssens 1967: ch. XV) conforms to what is suggested here, we totally agree with her. If, on the other hand, Tyssens means that coming from the same atelier implies the same date of transcription and the same dialect, we disagree for the following reasons: first, the language of A1 and F points to Haute-Marne, that of A2, A3 and A4 to Bourbonnais (and not to the Ile-de-France as is claimed by Tyssens 1967:328 for all manuscripts of the A-family); second, the text of A4, which contains forms such as *pensser* and *forssenez*, has to be dated as later than A1, A2 and A3 (see van Reenen 1994).

3. Internal Parameters

External parameters of the types discussed in section 2 contribute to the understanding of stemmatological relations without looking at the variants within the manuscripts. We now turn our attention to the variants within the manuscripts.

¹⁸ Wattel and van Mulken "Shock waves in text traditions", figs. 3 and 4, in this volume have shown that there is no contamination in the extant manuscripts of the *Charroi*.

Copying a text from a Vorlage may involve different scribal activities:¹⁹

- (1) precise, unchanged copying by a mirror copyist (see Benskin and Laing 1981:58 for this term);
 - (2) translation into the scribe's own dialect;
 - (3) partial adaptation, i.e., a compromise between (1) and (2).
- In addition a scribe may suppress or add passages.

Even though we are not certain of the Vorlage(s), we can say that all manuscripts of the *Charroi de Nîmes* are cases of (3), partial adaptations, and that within the A-family we are often close to mirror copying. All replacements of forms in the Vorlage by forms proper to the language of the scribe are translations - some suitable, some not - which may include newly created assonances or rhymes. These highly interesting replacements provide us with stemmatological and dialectal information. All correspondences between text and Vorlage are the result of mirror copying. Even if we were to assume that we know which is the text and which is the Vorlage, an assumption which usually is unrealistic, a problem still remains. Since the text is partially adapted, we cannot be certain whether the mirror copying of a word or a passage is the result of the scribe sticking to the Vorlage, in spite of the fact that the source dialect of the Vorlage and the goal dialect of the text are not the same, or, alternatively, whether the mirror copying is due to the fact that source and goal dialects share the same features. Although it may be interesting from a dialectal point of view, mirror copying is not interesting from a stemmatological point of view, since it does not result in variation. Alternatively sticking to the Vorlage and translating from the Vorlage into another dialect within the same text may lead to a mixture of different dialects, a so-called *Mischsprache* (see Benskin and Laing 1981).

When a text is rhymed or assonated, there is an extra dimension which may throw light on both stemmatological and dialectological aspects. Though variants occur in rhyme or assonance, usually more variants are to be found in the body of the text. The recreation of assonances and rhymes in the dialect of the scribe being a more difficult activity than the adaptation of linguistic units occurring in the body of a text, it is more important for a scribe to conserve the original assonances and especially the original rhymes of a Vorlage, since assonances, rhymes (and metre) constitute the funda-

¹⁹ A typology of the scribe's activity has been presented in Schøsler 1995, section 3.1. We have found a similar distinction in Benskin and Laing (1981). See also van Mulken (1993:195-196) on this subject.

mental structure of a poem. The tendency of rhymes and assonances to be more stable between manuscripts than the body of the manuscripts tends to be, requires that special attention be paid to variants in rhyme and assonance.

. Variants found by comparing manuscripts may be of three types. There are local variants, non-local variants and interpolations / lacunas. Local variants are discussed in subsection 3.1, non-local variants in subsection 3.2. Interpolations and lacunas concern passages and their presence in one text and absence or reduction in another (see subsection 3.3). In some relevant subsections attention is paid to the distinction between variants in the body of the text and variants in rhyme and assonance.

3.1. *Local variants*

Variants are local when they are the result of a comparison between passages which occur at the same place in the manuscripts under investigation. They may concern differences in spelling (phonology) (see 3.1.1), morphological properties (see 3.1.2), content differences in clearly related passages (see 3.1.3) and relatedness versus nonrelatedness of passages (see 3.1.4). The linguistic expressions forming the variants may concern a word, a phrase, or even a clause, and, in poetry, two succeeding verses.²⁰

3.1.1. *Spelling and phonology (level 1)*

In a previous study on *Narcisse* (see Schøsler 1988), graphic or phonological variation, variation at level 1, was dismissed as irrelevant for stemmatological investigations, as the information obtained at this level seemed random. A closer inspection of this level, however, has led us to a more qualified opinion. Although some of the variations seem in fact to be completely irrelevant for the relation between manuscripts, e.g., the alternation of *q* and *c* in:

vs. 1172, ms. B2: *cuidez*, ms. C: *quidies*

other variants may reveal important dialectal distributions and thus provide valid information about the transmission of the text through time and space. An example is, in the same verse 1172, the final *z* / *s* distinction, corresponding to the southern and the northern dialects, respectively (see map 266

²⁰ See Van Hoecke (1989) for a method of noting variants even below the level of the hemistich, and van Mulken (1993:36-38) for a typology of local variants.

in Dees et al. 1980). Modifications in assonance due to dialectal differences are even highly relevant from a stemmatological point of view (see Schøsler 1995 and van Mulken 1993). Consequently, it is important to distinguish random graphic variation with no stemmatological relevance from variation that discloses dialectal, hence possible stemmatologically relevant, information. In Schøsler (1995, section 3) some phonological properties have been detected which could be assigned to the original or at least to a Vorlage.

In poetry scribes tend to choose the extreme solutions in rhyme and assonance: they either skip or mirror copy the rhymes and assonances of the Vorlage. Consequently, they can provide us with information about the region of origin of the text tradition.

An example is found in the verses 300 and 309 of the A and B manuscripts of the *Charroi*. The body of the texts contains *preu*, whereas the assonance is *prouz*. We can be almost certain that the Vorlage had *prouz*, and that in the body of the text the scribes adapted this older form to their more recent and often geographically different dialects which no longer contained the *ou*. They have, however, left the assonance untouched. This is one of the reasons why we can argue that the Vorlage was written in a geographically different and / or at least in an older stage of the dialect.

Sire Guillelmes, dit Looys li *prouz* (verse 300 A1; A4 has *proz*)
 'Sir Guillelmes, says Looys the brave'

Prenez la terre au *preu* conte Foucon (verse 309 A and B)
 'Take the land from the brave count Foucon'

The scribe of C has adapted the assonance as well and writes:

Loeys li *preus* (verse 359).
 'Loeys the brave'

The result is a vowel which is incompatible with the *o(u)*-assonance within the *laisse* (xv).

The Normand (Eure) dialect of B2 can by no means be considered to be the original dialect of the text, although its dialect is rather pure. The assonances in this manuscript, just as the assonances of the other manuscripts of the *Charroi*, point rather to Haute-Marne and, consequently, the geographical diffusion of the manuscripts of the *Charroi* must have been as indicated in Figure 1 (further arguments are found in Schøsler 1995).

Consequently, dialectal purity is not necessarily a stemmatologically relevant argument. When translating, an adept scribe will produce a linguistically homogeneous text, regardless of what he is copying: a Vorlage which is already a mirror copy, a pure translation or a partial adaptation.

In summary, it is possible to find information at the graphic / phonological level which is dialectally, hence possibly stemmatologically, relevant, especially in rhyme and assonance.

3.1.2. Morphology (level 2)

On the morphological level, level 2, we have the same two types of variants as in subsection 3.3.1: apparent random variations on the one hand and those revealing dialectal features of the Vorlage and of the scribes on the other hand. One dialectally relevant feature is case marking. Many features have not (yet) been examined with respect to dialectal distribution, e.g., the use of auxiliaries, as seen in *Charroi* v. 509 (*chevaucheraï / chevaucher vueil*), a potential type-2 opposition AD/BC (see table 1 above). Word order, studied at this level, is certainly a dialectally relevant feature (see Dees 1980, carte 276).

Information at the morphological level may contribute to the identification of the dialect both of an original and of a copy (see van Mulken 1993: 177-185; Busby 1993:34). From a stemmatological point of view it may thus be relevant.

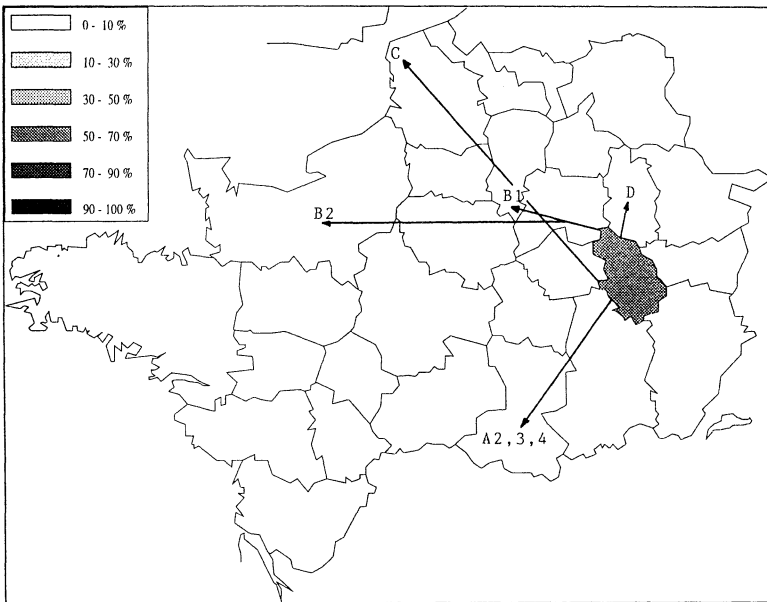


Figure 1. Geographical diffusion of the manuscripts of the *Charroi* de Nîmes

3.1.3. *Content differences in related passages (level 3)*

At the level of the content, level 3, we have the same two types of variants: apparent random variations on the one hand and those revealing dialectal features of the Vorlage and of the scribes on the other hand. Nonetheless, we know but little about dialectal lexical variants (see, however, Roques 1980; 1988; Möhren 1986). As for non-dialectal variations, it is at the level of content variation that we have proceeded to group the manuscripts according to the typology of Greg and Dearing, in order to find stemmatologically relevant evidence. This grouping has provided us with the type-2 oppositions discussed in section 1 of this paper. Here we also discuss other types of oppositions, for instance, oppositions of the type AB/C/D.²¹

At level 3 information can be gathered in two ways: by looking at what is the same within the formulas; and by considering what is different within the formulas. We have done so by making different kinds of groupings, which give some idea of the relations between the families. The results are given in tables 5 and 6, based upon the complete formula list of variants in the *Charroi* from vss. 1081 to 1215 (see appendix 1, col. 3). They concern both type-2 and other oppositions of level 3.²²

Table 5 shows that the number of correspondences per pair of manuscript families is always low in the case of D, for instance, between F and D it is not more than sixteen. According to this criterion, with ten to sixteen correspondences to the other families, D is almost equally removed from each of them and is distant from all of them. Table 6 shows another pattern. Although between D and the other families there are as many as 44 to 47 differences, here it is C which is more divergent from all the other families, with a score of 65 in the case of B. However, table 6 contains a bias. Since D is often not present at all, as we will see in table 8 (section 3.3 below), it is a priori unavoidable that D cannot show differences from the other families on such occasions. It would be better to count the differences of table 6 and the lacunas of table 9 together (see section 3.3 below).²³

²¹ Cf. Wattel and van Mulken ("Weighted formal support of a pedigree" in this volume) who provide suggestions as to how all types of correspondences and oppositions (including type-2 oppositions) can be stemmatologically dealt with.

²² Although F does not belong to a separate family, we have placed it in the tables as if it did.

²³ The best approach however to solve this problem has been developed by Wattel, "Shock waves in text traditions", section 3, in this volume.

Table 5: Number of correspondences (local variation)

	F	A	B	C	D
F	-	126	86	58	16
A		-	83	57	15
B			-	45	10
C				-	13
D					-

Table 6: Number of differences (local variation)

	F	A	B	C	D
F	-	6	28	60	47
A		-	31	62	48
B			-	65	46
C				-	44
D					-

3.1.4. *Relatedness versus nonrelatedness (level 4)*

In order to make a distinction between passages that show some resemblance and passages which are completely independent, we have established a fourth level of comparison between manuscripts, the level of relatedness (see appendix 1, coll. 4).²⁴ We need such a level in the *Charroi*, where the comparison between the families is often coarse-grained, whereas within each family it is possible to distinguish fine-grained variations, thus causing an uneven degree of comparison between closely and distantly related manuscripts. The introduction of level 4 is meant to repair this imbalance by allowing a final degree of comparison. As mentioned in section 1, at this final level almost all type-2 oppositions of the *Charroi* have been cancelled.

²⁴ In the study of Narcisse (Schösler 1988), this level did not seem necessary.

manuscripts H,L Deu vos dont, fait il, bone aventure
Beals sur tote creature
'That God may give you good adventure, he says,
(God) nicer than any creature'

The opposition ABCE/HL is best integrated on the level of content, level 3, as there is no doubt of the relatedness of the passage in the mss., i.e., there is no opposition at level 4.

Several reshufflings and even complete wandering passages are found in D as compared to the other families (see Tyssens 1967:113-114). This can be illustrated on the basis of the passage from vss. 1081 - 1215 of the *Charroi*, the only passage also found in F. Table 7, providing corresponding passages, shows that passages in D have “wandered”. D is written by an independent type of scribe who has reshuffled verses or series of verses four times with respect to what we consider as the “norm”.²⁵

Table 7: *Wandering passages in the Charroi de Nîmes, vss. 1081-1215*

ABC	Corresponding Passages in D
vss. 1133.1-1138	vss. 1121-1126
vss. 1165-1172	vss. 1126.1-1126.11
vs. 1177	vs. 1350
vss. 1185-1204	vss. 1326-1345

If we do not replace these wandering passages in the position corresponding to the “norm”, they will be considered as interpolations or lacunas (see section 3.3 below) and will thus fail to be compared to the corresponding passages of the “norm”. Since this is not a satisfactory solution, we prefer to put them back into the setting which gives us the possibility of comparing the passages. By replacing them, the modern scholar creates the possibility of considering such passages as potential candidates for establishing local variants (see section 3.1.1-4 above). For instance, it is possible to replace verse 1124 in D as verse 1136 as seen below:

- 1136 A Biau tres douz sire, Tiacre voiremant (transcription F)
 1136 B Biau tres douz sire, Tyacre vraiment (transcription B2)
 ‘Very, very good lord, Tiacre truly’
 1136 C [Tr]iacle ai nom, sacies a essiant
 ‘My name is Triacle, you must know this for certain’

²⁵ We are aware that to consider ABC as the “norm” is an arbitrary choice, made here for the sake of the comparison only.

1124 D J'ai nom Triacle par lou mien esciant
 'My name is Triacle, as far as I know'

Thus, we have created a type-2 opposition (AB/CD) at level 3 (see table 1 above: *biau tres douz sire* versus *ai nom*). This local variant can only be accepted on condition that the replacement of the verse has been accepted, i.e., that which was first a non-local variant has been accepted as a local variant. Without the replacement of verse 1124, verse 1136 would be analysed as AB/C and D lacunary.

In terms of the frequency of non-local variants, D differs from all other manuscripts. On the basis of this parameter alone we obtain the following opposition: ABC/D.

In summary, reshuffled passages or verses are non-local variants. Non-local variants are stemmatologically relevant, insofar as they can be put back in the position where they belong according to the "norm". The "norm" is arbitrarily chosen. Being made comparable, the reshuffled passages are no longer interpreted as interpolations or lacunas (see section 3.3), but they form newly created local variants. The creation of such local variants is dependent on the modern scholar.

3.3. *Interpolations and lacunas*

From the point of view of variant classification, the difference between an interpolation and a lacuna is arbitrary. If we consider D as the norm, the other manuscripts are often deficient. If we consider A as the norm, C and D are often interpolated. What is relevant, however, is that we register the differences, not whether we label them as interpolations or lacunas. The stemmatological results obtained do not depend on the choice of these labels.²⁶

²⁶ It is important to distinguish lacunas which are internal, linguistic phenomena from mutilations which are external, physical phenomena (see section 2.3 above). The difference is not always clear-cut. If an external mutilation has occurred in the Vorlage of a single copy, the mutilation may appear to the scholar as an internal lacuna in the extant manuscript. Whether or not this lacuna is the consequence of a mutilation may be difficult to detect, especially when the Vorlage no longer exists, as is the case with the *Charroi*-tradition where no manuscript has probably been copied from another of the extant manuscripts. The lacuna may have the length of, for example, a page, but external traces are lacking. Mutilations may also be more or less repaired. For lack of external evidence it will be difficult to recognise such spots in the text.

In assonating texts such as the *Charroi*, lacunas and interpolations usually consist of minimal units of one verse. Verses 1131-1133 of the *Charroi* illustrate this type of variation involving several cases:

1131

- A Dient paien qui les vont esgardant
'Say the pagans who look at them'
- F Dient paien, (continues 1132)
- B Dient paien, (continues 1132)
- C (see 1133)
- D Dient paien, (continues 1132)
'Say the pagans'

1132

- A A grant merveille avez or beaus enfanz
'A great miracle, you have nice children now'
- F mult avez beaus enfanz
'Very nice children you have'
- B il sunt molt bel enfant
'They are very nice children'
- C (see 1133)
- D mult par sont (continues 1133)
'They are very'

1133

- A S'il se seussent vestir avenament
- F S'il se seussent vestir avenament
'If they were able to dress charmingly'
- B S'il se voussissent vestir avenamant
'If they would dress charming'
- C Se il fussent vestu par avenant
'If they were dressed very charmingly'
- D avenant
'charming'

We find three verses in A, two in F and B, one in C and D, but none exactly identical. Since it is problematic to establish the identity of what is not present, it may be safe to start with the longest version and to consider

lacunas as being always independent from each other. The result is A/F/B/C/D.

Another point of view is to assume that lacunas of the same length are identical. It follows that the lacunas in B and F are the same and form a group, as opposed to C and to D which have shortened the passage, each in a different way and different from F and B. The result is A/FB/C/D. The question whether this assumption is justified can be confirmed only a posteriori. Since complete analysis of the data has taught us that F is closer to the original than to the other members of the A family, it might be the case that FB have the original reading. It follows from this that the A family, minus F, has added an interpolation; however, it does not follow necessarily that the scribes of C and D have shortened the passage independently from each other. By assuming that two lacunas of the same length are identical, it becomes possible to decide whether we have a lacuna or an interpolation. This assumption allows one to analyse the above passage as containing the following lacunas / interpolations, to which we have added the local variants.

Lacunas / interpolations	Local variants
1131 A/FBD/C	
1132 A/FB/D/C	A/F/B/C/D
1133 AFBC/D	AFB/C/D

Except in the evident cases, however, we refrain from interpreting lacunas and consider all lacunas as independent from each other.

Table 8: Number of lacking or interpolated verses in the Charroi the Nîmes

	F	A	B	C	D
F	-	3	47	45	87
A		-	50	46	90
B			-	70	112
C				-	116
D					-

In table 8 we present the frequencies of lacunas or interpolations encountered in the passage 1081-1215 of the *Charroi*. We have decided to consider the wandering passages in D (see table 5 above) as comparable to the corresponding passages in the other manuscripts, which diminishes the

number of lacunas and interpolations between D and the other families. The unit of comparison is the verse.

The frequencies in table 8 show the following: when we compare manuscripts in terms of absences and presences of verses, the A manuscripts and F are most closely related, B and C are more distant, also from each other and D is remote from all others (data come from Appendix 1, col. 1). The conclusion arrived at in this section is independent from the one in section 3.1, table 6, as it is based on completely different information. As was observed there, lacunas and oppositions have to be considered together. This has been done in table 9, in which data of tables 6 and 8 have been combined in order to obtain a more balanced picture.

Table 9: Number of local differences plus interpolations or lacking passages

	F	A	B	C	D
F	-	9	75	105	134
A		-	81	108	138
B			-	135	158
C				-	160
D					-

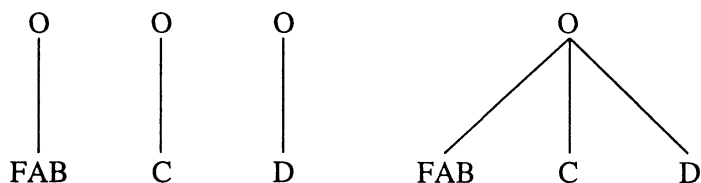
Table 9 is more or less a mirror image of table 5. D is further from all the other families than any of the other families. B is closer to the A family than C. Tables 5 and 9 point to the same stemmatological state of affairs.

3.4. *Synthesis*

In this section three main types of variants have been discussed. In 3.1 we have distinguished four levels of local variants, two of which are directly relevant for stemmatology: the content differences in related passages at level 3, and the differences in terms of relatedness or unrelatedness at level 4. In 3.2 we have introduced the notion of non-local variant. The third category distinguished, in 3.3, concerns interpolations and lacunas. What all the variants have in common, is that they contribute towards finding the right tree(s).

Tables 5 and 9 show that A and B are closely together and that C and D are distant from all others. This result of the evaluations of oppositions in the preceding section suggests the existence of two possible stemmata: a tri-

partite stemma, as in Figure 2, and three independent versions as in Figure 3. In Figure 4 we propose a partial stemma corresponding to the FAB-family, of which x is the hypothetical common ancestor.²⁷



*Figures 2 and 3: Possible partial stemmata of the Charroi de Nîmes
Right: tripartite; left: three independent versions.*

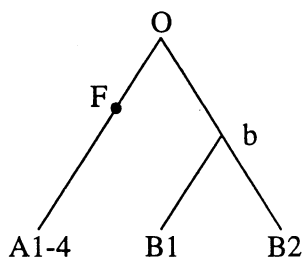


Figure 4: Possible partial stemma of the FAB-family of the Charroi de Nîmes

²⁷ Arguments for this stemma are advanced in Schøsler (1995).

4. Conclusion

Due to our uneasiness concerning the lack of formulas in terms of type-2 oppositions for establishing a stemma for the *Charroi de Nîmes*, we felt it necessary to look for other types of reliable and relevant information. This information can be provided by using external stemmatological arguments and especially by distinguishing different kinds of internal variation, besides type-2 oppositions. We have introduced the notion of *non-local variant* (reshuffling and wandering passage), and have distinguished interpolations and lacunas from local variations.

The study of variants at different levels provides independently substantiating evidence of the representation of the stemmata in Figures 2, 3 and 4, or is at least consistent with it. In our text tradition, manuscripts FAB are the most closely related, C is a more remote parent and D is even more distant.

In section 1 we assumed (1) the possibility of an archetype copied in only three, not four independent versions, giving birth to three, not four main families, and (2) the possibility of more than one independent ancestor. We might conclude that family D is the first candidate to represent an independent family, and that we may have some doubts in the case of C. As far as AB is concerned, there can hardly be any doubt that they all descend from the same common written ancestor.

References

- Benskin, M. and M. Laing. 1981. "Translations and *Mischsprachen* in Middle English Manuscripts". *...so meny people, longages and tonges. Philological Essays in Scots and Mediaeval English Presented to Angus McIntosh*, ed. by M. Benskin and M.L. Samuels, 55-106. Edinburgh.
- Busby, K. 1993a. "Variance and the politics of textual criticism". *Towards a Synthesis? Essays on the new philology*. Amsterdam: Rodopi. 29-45.
- Busby, K. 1993b. "Doin' philology while the -isms strut". *Towards a Synthesis? Essays on the new philology*, 85-95. Amsterdam: Rodopi.
- Busby, K. 1994. "The Scribe of MSS T and V of Chrétien's *Perceval* and its Continuations". *Les manuscrits de/The manuscripts of Chrétien de Troyes*, ed. by K. Busby, T. Nixon, A. Stones and L. Walters, 49-65. Amsterdam: Rodopi.
- Cerquiglini, B. 1989. *Eloge de la variante. Histoire critique de la philologie*. Paris: Editions du Seuil.

- Comfort, T.E. 1954. *The Charroi de Nîmes: An Old French Chanson de Geste Edited from the Manuscript of Boulogne-sur-mer* (= *Doctoral Dissertation Series* 7837). Ann Arbor, Michigan.
- Dees, A. 1984. "Regards quantitatifs sur les variations régionales en ancien français". *Dialectology, Quantitative Linguistics*, vol. 21, ed. by H. Goebel, 102-120. Bochum: Studienverlag Dr. N. Brockmeyer.
- Dees, A. 1988a. "Analyse par l'ordinateur de la tradition manuscrite du Cligès de Chrétien de Troyes". *Actes du XVIIIe Congrès International de Linguistique et de Philologie Romanes*, tome VI, 62-75. Tübingen: Max Niemeyer Verlag.
- Dees, A. 1988b. "Analyse des rimes dans la Bible de Macé de la Charité, vol VI et VII". *Aspects de Linguistique Française, Hommage à Q.I.M. Mok*, 91-106. Amsterdam: Rodopi.
- Dees, A., avec le concours de P. van Reenen et J. de Vries. 1980. *Atlas des formes et des constructions des chartes françaises du 13e siècle* (= *Beiheft zur Zeitschrift für romanische Philologie*, Band 178). Tübingen: Max Niemeyer Verlag.
- Dees, A., avec le concours de M. Dekker, O. Huber et K.H. van Reenen-Stein. 1987. *Atlas des formes linguistiques des textes littéraires de l'ancien français* (= *Beiheft zur Zeitschrift für romanische Philologie*, Band 212). Tübingen: Max Niemeyer Verlag.
- Dees, A., et P. van Reenen. 1980. "L'interprétation des graphies -o- et -ou- à la lumière des formes trouvées dans les chartes françaises du 13e siècle". *Linguistic Studies offered to Berthe Siertsema*, ed. by D.J. van Alkemade et al., 269-276. Amsterdam: Rodopi.
- Dekker, M. 1986. "Intermediariteit". *Jaarboek 1986, Corpusgebaseerde Woordanalyse*, 61-69. Amsterdam: Faculteit der Letteren, Vrije Universiteit.
- De Poerck, G., R. van Deyck and R. Zwaenepoel. 1970. *Le Charroi de Nîmes, chanson de geste*, 2 vol. Saint-Aquilin-de-Pacy.
- Foulet, A., and M.B. Speer. 1979. *On Editing Old French Texts*. Lawrence: The Regents Press of Kansas.
- Frappier, J. 1955, 1967. *Les Chansons de Geste du cycle de Guillaume d'Orange*, 2 vol. Paris.
- Jeanroy, A. 1935. "Review of Lange-Kowal (1934)". *Romania* 61.112-114.
- Jodogne, O. 1956. "Le manuscrit de Boulogne du 'Charroi de Nîmes'", *Publicaciones de la Facultad de Filosofía y Letras* II, 17, 301-326. Zaragoza.
- Lange-Kowal, E.-E. 1934. *Das altfranzösische Epos vom Charroi de Nîmes: Handschrift D, herausgegeben mit sprachwissenschaftlichem Kommentar und Glossar*. Jena.
- Löfstedt, L. 1995. "Identifying an Old French text with the help of dialect analysis". *Historical Linguistics 1993*, ed. by H. Andersen, 251-258. Amsterdam & Philadelphia: John Benjamins.
- McMillan, D. 1972. *Le Charroi de Nîmes*, éditée d'après la rédaction AB, avec introduction, notes et glossaire. Paris: Editions Klincksieck.

- Mok, Q.I.M. 1992. +/- "Nouvelle philologie?". *Neophilologus* 508-518.
- Mulken, M. van. 1993. *The Manuscript Tradition of the Perceval of Chrétien de Troyes. A stemmatological and dialectological approach*, dissertation. Amsterdam: Vrije Universiteit.
- Möhren, F. 1986. *Wort- und Sachgeschichtliche Untersuchungen an französischen landwirtschaftlichen Texten, 13., 14. und 18. Jahrhundert Seneschauie, Ménagier, Encyclopédie*. Tübingen: Niemeyer.
- Quentin, Dom H. 1926. *Essais de critique textuelle*. Paris: Editions Auguste Picard.
- Reenen, P. van. 1982. "Voyelles nasales en ancien français non suivies de consonne nasale". *Rapports/Het Franse Boek* 52, numéro 3.132-144.
- Reenen, P. van. 1988. "AN/EN en ancien français: distributions (géo)graphiques". *Aspects de Linguistique Française, Hommage à Q.I.M. Mok*, 141-160. Amsterdam: Rodopi.
- Reenen, P. van. 1989. "La pertinence linguistique des rimes en EN/AN dans la Bible de Macé de la Charité". *Actes du Colloques sur l'ancien provençal, l'ancien français et l'ancien ligurien, Bulletin du Centre de Romanistique et de Latinité Tardive*, vol. 4-5.247-266.
- Reenen, P. van. 1994. "Les premières (?) voyelles nasales en ancien français et le rapport avec la non prononciation du *r*, -ss- intervocalique dans *pensser* et *perssone*". *Diachronie et variation linguistique, Les nasalisations dans le monde roman. Communication & Cognition* 27.111-121.
- Reenen, P. van, and K.H. van Reenen-Stein (eds.). 1988. *Distributions spatiales et temporelles, Constellations des manuscrits. Etudes de variation linguistique offertes à Anthonij Dees à l'occasion de son 60me anniversaire*. Amsterdam & Philadelphia: John Benjamins.
- Reenen-Stein, K.H. van. 1988. "Ne weil passer sanz paier men winage". *Jaarboek 1987-88, Corpusgebaseerde Woordanalyse*, 111-113. Amsterdam: Faculteit der Letteren, Vrije Universiteit.
- Régner, C. 1966. *Les rédactions en vers de la Prise d'Orange*. Paris: Klincksieck.
- Régner, C. 1968. "A propos de l'édition du "Charroi de Nîmes"". *L'information Littéraire* 20.32-33.
- Roques, G. 1980. *Aspects régionaux du vocabulaire d'ancien français*, thèse inédite de Strasbourg.
- Roques, G. 1988. "Quelques mots régionaux dans le poème de la Vie de saint Silvestre et de l'Invention de la sainte Croix". P. van Reenen and K.H. van Reenen-Stein (eds.) 1988.177-185.
- Salemans, B.J.P. 1990. "Text Genealogical Remarks on Lachmann, Bédier, Greg and Dearing". *Leuvense Bijdragen* 79.4.427-468.
- Schösler, L. 1988. "La constellation de Narcisse". P. van Reenen and K.H. van Reenen-Stein (eds.) 1988.247-263.
- Schösler, L. 1989. "Problèmes de stemmatologie, illustrés par le cas de Narcisse". *Jaarboek 1988-89, Corpusgebaseerde Woordanalyse*, 167-174. Amsterdam: Faculteit der Letteren, Vrije Universiteit.

- Schøsler, L. 1995. "New Methods in Textual Criticism: The Case of the *Charroi de Nîmes*". *Medieval Dialectology*. ed. J. Fisiak, 225-276. Berlin: Mouton de Gruyters.
- Tyssens, M. 1967. *La Geste de Guillaume d'Orange dans les manuscrits cycliques*. Paris: Les Belles Lettres.
- Van Hoecke, W. 1989. *Aspects de la critique textuelle et de l'édition critique de textes français du moyen âge*. (= *Preprint* no 63). Leuven: Faculteit van de Letteren en de Wijsbegeerte, Katholieke Universiteit.
- Weber, C. 1912. *Die Prosafassungen des "Couronnement de Louis", des "Charroi de Nîmes" und der "Prise d'Orange"*. Inaugural-Dissertation. Halle.

Appendix 1: Variations of form and meaning of the passage vss. 1081 to 1215, i.e., the only passage containing all manuscripts of the *Charroi*

LEGEND:

- column 1: formal presences/absences:
 + simple presence of a manuscript lesson with no respect to the content of the verse
 % transfer of a lesson from another part of the manuscript to this place where it is thought to fit better
 - lacuna in relation to the norm (arbitrarily chosen, here manuscript A is chosen as the norm)
 # lacuna of the norm (manuscript A) accompanied or not accompanied by a lacuna in the other manuscripts
 column 2: verse number
 columns 3 and 4: content variation (level 3) in column 3 and variation of relatedness (level 4) in column 4

COL. 1	COL. 2	COL. 3	COL. 4
F A B C D	VERSE	FORMULAS/3/	/4/
++++ -	1081	FABC	
++++ -	1082	FAB/C	FAB/C
### + -	10821		
++++ -	1083	FABC	
++++ -	1084	FABC	
## + + -	10841	B/C	B/C
## + + -	10842	B/C	B/C
## + + -	10843	B/C	B/C
## + + -	10844	B/C	B/C
laisse			
++ - + -	1085	FAC	
++ - + -	1086	FAC	
++ - + -	1087	A/FC	FAC
++ - + -	1088	FA/C	FAC
++ - + -	1089	FA/C	FAC
++ - + -	1090	FAC	
++ - + -	1091	FA/C	FA/C
++ - + -	1092	F/A/C	FA/C
++ - ++	1093	FAC/D	FAC/D
++ - ++	1094	FA/C/D	FA/C/D
++ - ++	1095	FA/C/D	FAD/C

++-++	1096	FA/C/D	FA/C/D
++-++	1097	F/A/C/D	FAD/C
++-++	1098	FA/C/D	FA/C/D
++-++	1099	FA/C/D	FA/C/D
##-+#	10991		
##-+#	10992		
##-+#	10993		
##-+#	10994		
##-+#	10995		
##-+#	10996		
##-+#	10997		
-+---	1100		
laisse			
####	11011		
+++++	1101	FAB/C/D	FABCD
+++ - +	1102	FAB/D	FABD
+++++	1103	FACD/B	FABCD
+++++	1104	FABCD	
+++++	1105	FAB/C/D	FABC/D
#### +	11051		
+++++	1106	FACD/B	FABCD
+++++	1107	FAC/B/D	FACD/B
#### +	11071		
+++++	1108	FACD/B	FACD/B
+++++	1109	FA/B/C/D	FAC/B/D
+++ - +	1110	FA/B/D	FA/B/D
+++++	1111	FABD/C	FABCD
## + # +	11111	B/D	B/D
## + # +	11112	B/D	B/D
#### +	11113		
#### +	11114		
laisse			
++++ -	1112	FABC	
++++ -	1113	FAB/C	FABC
## # + -	11131		
++++ -	1114	FABC	
++++ -	1115	FABC	
++++ -	1116	FAB/C	FAB/C
++++ -	1117	FABC	
## # + -	11171		
++++ -	1118	FAB/C	FABC
+++ - -	1119	FAB	

# # + # -	11191		
# # + # -	11192		
# # + # -	11193		
# # + # -	11194		
# # + # -	11195		
# # + # -	11196		
# # + # -	11197		
# # + # -	11198		
+ + + + +	1120	FACD/B	FABCD
+ + + + +	1121	FABC/D	FABC/D
# # + # #	11211		
+ + + + +	1122	FABC/D	FABC/D
+ + + + +	1123	FABC/D	FABC/D
# # + # #	11231		
+ + + + +	1124	FAB/C/D	FABCD
+ + + + +	1125	FABC/D	FABCD
# # # # +	11251		
+ + + + +	1126	FABCD	
+ + + + +	1127	FABCD	
+ + + + +	1128	FABC/D	FABCD
+ + + + +	1129	FABCD	
- + - + -	1130	AC	
+ + + - +	1131	A/FB/D	A/FBD
- + - - -	1132		
+ + + + +	1133	FABC/D	FABCD
# # # # %	11331		
+ + + + %	1134	FABC/D	FABCD
+ + + + %	1135	FABD/C	FABCD
+ + + + %	1136	FAB/CD	FABCD
+ + + + %	1137	FABCD	
+ + + + %	1138	FAB/C/D	FABCD
# # # # +	11381		
# # # # +	11382		
# # # # +	11383		
# # # # +	11384		
+ + + + +	1139	FABC/D	FABC/D
+ + + + +	1140	FAB/C/D	FABCD
+ + + + +	1141	FABC/D	FABC/D
+ + + + -	1142	FAB/C	FABC
+ + + + -	1143	FAB/C	FAB/C
+ + + + -	1144	FA/B/C	FA/B/C
# # + + -	11441	B/C	B/C
# # + + -	11442	B/C	B/C
# # + + -	11443	B/C	B/C

# # + + -	11444	B/C	B/C
# # # + -	11445		
+ + + + -	1145	FAB/C	FAB/C
+ + + + -	1146	FA/B/C	FAB/C
# # + # -	11461		
+ + + + -	1147	FABC	
# # + + -	11471	B/C	BC
+ + + + -	1148	FABC	
+ + + + -	1149	FABC	
+ + + + -	1150	FABC	
+ + + + -	1151	FA/B/C	FABC
# # # + -	11511		
# # # + -	11512		
+ + + + -	1152	FAC/B	FABC
+ + + + -	1153	FAC/B	FAC/B
# # + + -	11531	B/C	B/C
# # + # -	11532		
laisse			
# # + # -	11541		
+ + + + -	1154	FA/B/C	FAB/C
# # # + -	11541		
+ + + + -	1155	FABC	
+ + + + -	1156	FAC/B	FAC/B
+ + + + -	1157	FAB/C	FAB/C
+ + + - -	1158	FAB	
+ + + - -	1159	FAB	
+ + + - -	1160	FAB	
+ + + + -	1161	FA/B/C	FABC
+ + + + -	1162	FABC	
+ + + + -	1163	FABC	
+ + + + -	1164	FA/B/C	FABC
+ + + + -	1165	FAB/C	FAB/C
# # # # %	11651		
# # # # %	11652		
# # # # %	11653		
+ + + + %	1166	FAB/C/D	FABCD
+ + + + %	1167	FABCD	
# # # # %	11671		
+ + + + %	1168	FAB1/B2/C/D	FAB1CD/B2
+ + + + %	1169	FAB/C/D	FAB/C/D
+ + + + %	1170	FAB/C/D	FAB/C/D
+ + + - %	1171	FA/B/D	FAD/B
+ + + + %	1172	FAB1/B2C/D	FABC/D

++++ -	1173	FA/BC	FABC
++++ -	1174	FABC	
++++ -	1175	FABC	
++++ -	1176	FABC	
++++ %	1177	FABC/D	FABCD
+++++ -	1178	FABC	
++++ -	1179	FA/B/C	FABC
++++ -	1180	FA/B/C	FA/B/C
++++ -	1181	FA/B/C	FAB/C
+++ - -	1182	FAB	
# # + - -	11821		
++++ -	1183	FABC	
++++ -	1184	FAB/C	FABC
laisse			
++++ %	1185	FABC/D	FABC/D
++++ %	1186	FA/B/C/D	FAC/B/D
++++ %	1187	FACD/B	FABCD
+++ - -	1188	FAB	
++++ + %	1189	FAD/B/C	FABCD
++++ -	1190	FA/BC	FABC
++++ %	1191	A/FBC/D	FABC/D
++++ %	1192	FABC/D	FABC/D
++ - - -	1193	FA	
++++ %	1194	FAB/C/D	FABCD
++++ %	1195	FABC/D	FABC/D
++++ %	1196	FABC/D	FABC/D
++++ %	1197	FAB/C/D	FAB/C/D
++ - + %	1198	FA/C/D	FA/C/D
++ - + %	1199	FA/C/D	FA/C/D
++++ %	1200	FA/B/C/D	FA/B/C/D
+++ - %	1201	FAB/D	FAB/D
+++ - %	1202	A/FBD	FABD
# # # %	12021		
# # # %	12022		
++++ %	1203	FA/B/C/D	FA/B/C/D
++++ %	1204	FABCD	
laisse			
++++ -	1205	FAB/C	FAB/C
# # # + -	12051		
# # # + -	12052		
# # # + -	12053		
# # # + -	12054		

# # # + -	12055		
+ + + + -	1206	FAB/C	FAB/C
+ + + + -	1207	FAB/C	FAB/C
+ + + - -	1208	FAB	
+ + + + -	1209	FABC	
+ + + + -	1210	FAB/C	FABC
+ + + + -	1211	FAB/C	FAB/C
# # # + -	12111		
+ + + + -	1212	FAB/C	FABC
+ + + + -	1213	FABC	
+ + + + -	1214	FAC/B	FABC
# # + # -	12141		
# # + # -	12142		
+ + + + -	1215	FAB/C	FABC

Appendix 2: Type-2 oppositions

VERSE	LEVEL 3	LEVEL 4
0014	AB/CD	ABCD
0029	AB/CD	AB/CD
0054	AD/BC	ABCD
0127	AB/CD	ABCD
0161	AB/CD	AB/CD
0246	AB/CD	ABCD
0313	AB/CD	ABCD
0466	AB/CD	ABCD
0483	AB1/CD	ABCD
0509	AD/BC	ABCD
0765	AD/BC	AD/BC
0766	AB/CD	ABCD
0903	AB/CD	ABCD
0903	AD/BC	ABCD
0923	AB/CD	ABCD
1136	FAB/CD	FABCD
1234	AD/B2C	ABCD

OPPOSITIONS AT CONTENT (USUALLY LEXICAL) LEVEL 3

vs. 0014 AB/CD

- A1 Ce fu en mai, el novel tens d'esté
- A2 Ce fu en mai, el novel tens d'esté
- A3 lacuna
- A4 Ce fu en moi, ou novel tens d'esté
- B1 Ce fu en may, ou nouvel tans d'esté
'It was in May, the new summer time'
- B2 lacuna
- C Che fu en mai, que on dit en esté
'It was in May, that one says in the summer'
- D Ce fut en mai, que l'an dit en esté
'It was in May, the new summer time'

vs. 0054 AD/BC

- A1 Puis en monta tot le marbrin degré
- A2 Puis en monta tot le marbrin degré
- A3 lacuna
- A4 Puis en monta tot le marbrin degré
'After that he went up the marble staircase'

- B1 Puis en monta enz ou pales listé
 'After that he went up in the framed palace'
 B2 lacuna
 C Et puis monta sus ei palais listé
 'End after that he went up in the framed palace'
 D O palais monte les maberins degrés
 'In the palace he climbs the marble staircase'

vs. 0127 AB/CD

- A1 Que les troncons en volent trusqu'as trez
 A2 Que les troncons en volent trusqu'as trez
 A3 lacuna
 A4 Que les troncons en volent jusqu'au trez
 B1 Et li troncon en volent jusqu'a tref
 B2 Et li troncon en volent jusqu'a tre
 'That the splinters fly to the tents'
 C Que les esclices en fist en haut voler
 'That the fragments fly in the air'
 D Et les esclices en vollent vers lou trer
 'That the fragments fly to the tents'

vs. 0246 AB/CD

- A1 Tel li donai de mon tranchant espié
 A2 Tel li donai de mon tranchant espié
 A3 lacuna
 A4 Tel li doné de mon tranchant espié
 'Thus I gave him with my sharp spear'
 B1 Tel li donnai de mon trenchant espiél
 B2 Tel li donai de mon trenchant espié
 C Tel li donai de l'espée d'achier
 D Tel li donoï li de m'espée d'acier
 'Thus gave him with my spear of steel'

vs. 0313 AB/CD

- A1 Qui bien la terre maintenir en porront
 A2 Qui bien la terre maint[enir en por]ront
 A3 Qui bien la terre maintenir en porront
 A4 Qui bien la terre maintenir en porront
 'Who shall be able to maintain the land well'
 B1 Qui sa grant terre maintenir bien porront
 B2 Qui sa grant terre bien maintenir porront
 'Who shall be able to maintain the great land well'
 C D'ore en avant bien lor terre tenront
 'Henceforth they will keep their land well'

- D D'or en avent lor terre maintenront
 'Henceforth they will keep their land'

vs. 0466 AB/CD

- A1 Puis li a dit: Guillaume, quar seez
 A2 Puis li a dit: Guillaume, quar seez
 A3 Puis li a dit: Guillaume, quar seez
 A4 Puis li a dit: Guillaume, quar seez
 B1 Puis li a dit: Guillaume, car seez
 B2 Et puis a dit: Guillaume, ca sees
 'Then he said to him: Guillaume, sit down!'
 C Sire Guillaume, dist il, c'or vos sees
 'Sir Guillaume, he said, sit down please'
 D Sire Guillaume, dejoste moi saez
 'Sir Guillaume, sit (down) next to me'

vs. 0483 AB1/CD

- A1 Si vos demant Nymes cele cité
 A2 Si vos demande Nymes cele cité
 A3 Si vos demant Nymes cele cité
 A4 Si vos demant Nymes cele cité
 B1 Si vous demant Nymes celle cité
 'So I ask you Nymes that city'
 B2 Si vous demant Nimes la fort cité
 'So I ask you Nymes that strong city'
 C Si demant Nimbres cele bone chité
 'So I ask Nymes that good city'
 D Les tors de Nimes cele bone cités
 'The towers of Nymes, the good city'

vs. 0509 AD/BC

- A1 Chevaucherau au soir et a la lune
 A2 Chevaucherau au soir et a la lune
 A3 Chevaucherau au soir et a la lune
 A4 Chevauchere au soir et a la lune
 'I shall mount in the evening and in the moonshine'
 B1 Chevaucher vueil au soir et a la lune
 B2 Chevauchier veull au soir et a la lune
 'I will mount in the evening and in the moonshine'
 C Chevalchier voel au jor et a la lune
 'I will mount in the day and in the moonshine'
 D Chevalcherau au soir et a la lune
 'I shall mount in the evening and in the moonshine'

vs. 0766 AB/CD

- A1 lacuna

- A2 Calices d'or et messeaus et sautiers
- A3 Calices d'or et messauz et sautiers
- A4 Calices d'or et messiex et sautier
- B1 Calices d'or et messeux et sautier
- B2 Calices d'or et messel et sautier
- 'Golden chalices and missals and psalters'
- C Calices d'or et crois et encensiers
- 'Golden chalices and crosses and censers'
- D Galices porte et crois et encensiers
- 'He bears chalices and crosses and censers'

vs. 0903 AB/CD

- A1 Di *va*, vilain, par la loi dont tu vives
- A2 Di *va*, vilain, par la loi dont tu vives
- A3 Di *va*, vilain, par la loi dont tu vives
- A4 Di *va*, vilain, par la loi dont tu vive
- 'Well, peasant, according to the religion by which you live'
- B1 Di *va*, vilain, nel me celer tu mie
- B2 Dy *va*, vilain, et nel me celer mie
- C Di *moi*, vilain, nel me choiles tu mie
- 'Tell me, peasant, do not hide it for me'
- D Di *moi*, vilains, par les deus ou tu pries
- 'Tell me, peasant, according to the god to which you pray'

vs. 0903 AD/BC (for translations, see above)

- A1 Di *va*, vilain, *par* la loi dont tu vives
- A2 Di *va*, vilain, *par* la loi dont tu vives
- A3 Di *va*, vilain, *par* la loi dont tu vives
- A4 Di *va*, vilain, *par* la loi dont tu vive
- B1 Di *va*, vilain, nel me *celer* tu mie
- B2 Dy *va*, vilain, et nel me *celer* mie
- C Di *moi*, vilain, nel me *choiles* tu mie
- D di *moi*, vilains, *par* les deus ou tu pries

vs. 0923 AB/CD

- A1 Qui avroit ore .m. tonneaus de tel guise
- A2 Qui avroit ore .m. tonneaus de tel guise
- A3 Qui avroit ore .m. toneaus de tel guise
- A4 Qui avroit ore .m. tonniaus de tel guise
- B1 Qui avroit ore .m. tonniax d'itel guise
- B2 Qui avroit ore .m. tonniaus de tel guise
- 'Who would have now 1000 barrels of such kind?'
- C Qui aroit ore .xx. tonnes a teil guise
- 'Who would have now 20 casks of such kind?'

- D Qui averoit .ix. tonnes prises
 'Who would have taken 9 casks?'

vs. 1136 FAB/CD

- F Biau tres douz sire, [Ti]acre voiremant
 A1 Beau tres dolz sire, Tiacre voirement
 A2 Biau tres dolz sire, Tiacre voirement
 A3 Biau tres douz sire, Fiacre voirement
 A4 Biau tres douz sire, Fiacre voiremant
 B1 Biaus tres dous sire, Tyacre vraiment
 B2 Biau tres douz sire, Tyacre vraiment
 'Kind, very good lord, Tiacre truly'
 C [Tr]iace ai nom, sacies a essiant
 'Triacle my name is, this know well'
 D J'ai nom Triacle par lou mien esciant
 'My name is Tiacre, as far as I know'

vs. 1234 AD/B2C/B1

- A1 Quant je fui juvenes, meschins et bachelers
 A2 Quant ge fui juesnes, meschins et bachelers
 A3 Quant ge fui juvenes, meschins et bacheler
 A4 Quant je fu jeune, meschins et bacheler
 'When I was young, young knight and bachelor'
 B1 Quant je fui jeunes et nouviex adoubez
 'When I was young, and newly armed as knight'
 B2 Quant g'iere joene et en meilleur aé
 C Quant je fui juvenes et de petit aé
 'When I was young, and in the best age'
 D Cant j'estoie anfes, meschins et bacheler
 'When I was a child, young knight and bachelor'

OPPOSITIONS AT LEVEL 4

vs. 0029 AB/CD

- A1 Li cuens Guillaume fu molt gentix et ber
 A2 Li quens Guillaume fu molt gentis et ber
 A3 lacuna
 A4 Li quens Guillaume fu molt gentil et ber
 B1 Li quens Guillaume fu molt gentiz et ber
 B2 lacuna
 AB 'The count Guillaume was very noble and brave'
 C A son ostel descendi au degre
 'He dismounted beside the steps of his home'

- D En la grant place est descendus li ber
 'He, the noble, has dismounted at the great square'

vs. 0161 AB/CD

- A1 S'il le deffent, bien en doi avoir blasme
 A2 S'il le deffent, bien en doi avoir blasme
 A3 lacuna
 A4 S'i le deffant, bien en doi avoir blame
 B1 S'il s'en deffent, bien en doi avoir blasme
 B2 S'il s'en defent, bien en doi avoir blasme
 'If he forbids, I am to blame'
 C S'il le desdist, prest sui que m'en combate
 'If he refuses, I am prepared to fight'
 D S'o volt noier, pres suis d'a lui conbatre
 'If he wants to deny, I am prepared to fight him'

vs. 0765 AD/BC

- A1 lacuna
 A2 Bien vos sai dire que porte li premiers
 A3 Bien vos sai dire que porte li premiers
 A4 Bien vos sai dire que porte li premier
 'I am able to tell you well what the first contains'
 B1 Bien vos sai dire qu'en portent li princier
 B2 Bien vos sai dire qu'en portent li princier
 C Bien vos sai dire que mainent li princhier
 'I am able to tell you well what the princes bear'
 D Plaist vos oïr que porte li prumiers
 'Do you want to hear what the first contains?'

Index

- Accidence 240, 241
Adequacy
 grammatical 29, 32, 36
 historical-philological 36
Alexander the Great 210, 211,
 213, 216, 227, 228
Algorithm to build chains with 41
Alphabet
 alphabet acrostich 243
 Cyrillic 241, 243
 Glagolitic 233, 243
 Slavic 235
Alteration
 intentional 260
 unconscious 260
Apparatus of variants 12
Approach
 dialectological approach 257
 historical approach 252
 quantitative approach 249, 250
Archetype 123, 127, 135, 140, 237,
 239, 252, 256-258
 archetypal reading 241, 244
 archetypal text 241, 243, 244
Assonance 264, 276, 277, 278,
 279, 282, 284
Authenticity 170, 172ff, 176ff,
 186

Bédier 98
Best fit 136, 137
Bitwise and 166, 167
Bitwise or 166, 167
Branch 138

Canterbury Tales Project 80, 94
Cardiograms 117
CD-ROM and network publication
 74, 75
Centrality 127, 129
 centrality score 127
Chain
 bundle 186-188, 192, 197, 202,
 203
 definition 170, 184ff
 nodes 191, 192
Change score 115
Character state(s) 61-65, 67-70
Cladistics (or: phylogenetic
 systematics) 3, 43-46, 50,
 76-78, 80, 88
 (cf. also: Parsimony; PAUP)
Cladogram 69
Cluster 130
Codices deteriores 224
Codices recentiores 224
Codicising 222
Coincidence 254, 259, 260
 coincidental readings 251
Collate 74, 75, 94
Collation 71-75, 80, 81, 84, 90,
 91, 99, 101
Combinations (of texts) 210,
 218-220
Common link
 divers under the common link
 201, 202
 theory 189, 192, 193, 197,
 202-204
Competence 243, 244

- Computer corpus 181-184, 204
- Conflation 239
- Connection (internal, terminal) 138
- Consistent 151
 - Consistency Index 48, 49, 69
- Contamination (or: hybridisation)
 - 9, 41, 43, 44, 46, 69, 70,
 - 105, 123, 235, 239, 253, 254
 - contaminated manuscript 250
 - contaminated tradition 250
- Context 251, 256, 257, 259, 260
 - contextualisation 221
- Contingent irregularity 249
- Contradict 151
 - contradicting percentage 158
- Copy 251
 - copying 239, 243, 244
 - physical copy 210, 220
- Corpora 256
 - Athanasian Corpora 256, 258
 - Corpora manuscripts 256,
 - 258-260
 - Corpora tradition 256, 260
- Cumulative formula weight 112
- Cycle 270, 271
- Data structure 108, 124, 139
- Dead end 154, 155
- Degree 138
- Dialect
 - dialect properties 265
 - dialectal area 275
 - dialectal distribution(s) 277,
 - 279
 - dialectal feature(s) 275, 279
 - dialectal information 276
 - dialectal purity 278
- Dichotomy 149, 150, 155, 250, 255,
- 257, 259, 260
 - best supported dichotomies 255
 - dichotomic relationship 253
 - dichotomy separators 166
 - dichotomy support 159, 161
 - quantitatively determined
 - dichotomies 253
- Difference formula 108
- Discovery 249-257
 - logic of discovery 252
- Distance, quantitative 253, 254
- Dossier 210, 220-225, 227-230
 - dossier manuscript 221
 - dossier transmission 222
- Double(s) 153, 154, 155
- Edge 123, 138, 150
 - edge contraction 137, 157, 161
- Elimination 250, 251, 257, 260
 - codices eliminandi* 259
- Error(s) 251, 255, 256, 258-260
 - common errors 256
 - mistake 251, 252, 258, 260
- Evaluation 5, 250-252, 255-257,
- 260
 - judgements of variants 14, 18,
 - 29, 34-36, 37
 - objectivity and subjectivity 14,
 - 23
 - philological evaluation 250
 - philological judgements 250
 - qualitative evaluation 250, 254,
 - 255
 - quantitative evaluation 255
 - subjective judgements or
 - interpretations 7
- Evolutionary systematics 42, 43
- Exchanging attachments 146, 147
- Family 257
 - family relations 255
- Formula weight 112, 124, 127, 131,
- 139
- Goodness of fit 136
- Graph, graph theory 123, 138
- Grid 111, 115
- Group 250, 253-258, 260

- Handwriting 270, 274
 different hands 274
 related hands 270, 274
 Heading line 108, 124, 139
 Hermeneutical circle 252
 Homology 42, 43
 (cf. also: Parallelism)
 Homoplasy 8
 Hyparchetype 241-244, 252
 hyparchetypal text 241, 243

 Inconsistency 254
 inconsistent readings 252
 inconsistent variants 250, 254
 Information
 biographical 181-183, 198, 199
 content 110, 126, 141
 geographical 199
 information entity 181, 182
 reliability of 180-182, 199
 Intentional alteration 257, 258, 260
 Intentional variants 257
 Intermediate (manuscript, node)
 137, 144, 160
 intermediarity 251, 260
 Internal connection 138, 150
 Interpolation 17-19, 32, 33, 35, 36, 268, 274, 277, 283-287, 289
 Inversion 19, 32, 33, 35, 36, 282
 Irregularity, contingent 249

 Juynboll 175

 Lachmann 71, 72, 99
 Lachmann
 Lachmannian common errors 14, 43
 Lachmannian logic of discovery 252
 Lachmannian method 4, 14, 22, 45, 46, 51, 250, 251, 253, 255
 Lachmannian notion 12
 Lacuna 274, 277, 281, 283-287, 289, 293
 Lectiones coniunctivae 257
 Lectiones distinctivae 257, 259, 260
 Lectiones separativae 257, 259, 260
 Level
 first 249, 251
 second 251
 third 252
 Three Level Method 249-252
 Level of comparison
 content 265, 287, 293
 morphology 277, 279
 phonology 277-278
 relatedness 265, 287, 293
 spelling 277-278
 Levelling 111
 Lumping, lump 131

 MacClade 5, 46
 (cf. also: PAUP)
 Manuscript
 composite manuscript 219-222, 225
 dossier manuscript 221
 manuscript collation 214
 manuscript sources 210
 manuscript tradition 218, 220
 manuscript transmission 218
 manuscript volume 218, 220, 230
 Manuscripts
 Cambridge, Gonville and Caius College, 177/210 215
 Leiden, University Library, B.P.L. 20 215
 London, British Library, Cotton Nero D viii 215
 London, British Library, Cotton Vitellius A viii 225

- Madrid, Biblioteca Nacional,
 9783 212
 Oxford, Lincoln College, 96
 215
 Paris, Bibliothèque nationale,
 lat. 5062 215
 Paris, Bibliothèque nationale,
 Nouv. acq. lat. 310 221
 Rouen, Bibliothèque
 Municipale, 1173 – Y.11
 228
 Vaticano, Ottob. lat. 909 225
 Vaticano, Reg. lat. 339 228
 Monophyletic group 20, 43-45,
 62
 (cf. also: Type-2 variations)
 Mutilation 272, 274, 284

 Node 123, 138
 Noise 250, 253, 254
 Normative message 177

 Ólafsson, Gunnar 81, 83
 Ólafsson, Helgi 83
 Omission 16-20, 27, 32, 33, 35,
 36
 Orientation 252, 259
 oriented structure 252

 Paleographical argument 257
 Parallelism 8-10, 12, 15, 27, 29,
 37, 41, 43, 44, 46, 50, 56,
 69, 70
 (cf. also: homoplasy; coincident
 variation; accidental variation)
 accidental 8, 30
 diachronical 11, 16, 30, 31
 idiolectical 10, 16, 30, 31
 inflectional 11, 30
 orthographic 30
 regional 10, 30
 synonymous 9, 10, 27, 29, 31
 Parameters
 external 265, 269, 270, 275
 internal 265, 269, 274, 275
 Parsimony, principle of 44-46,
 50, 76, 80
 Passages
 reshuffling of 282-284, 288,
 289
 wandering 282, 283, 286, 289
 Path 138
 PAUP 3, 4, 22, 28, 41, 42, 45-50,
 77, 84, 86, 87
 Pedigree 123
 Performance 243, 244
 Periphery 129
 Perspective
 focussed on copyist 256, 260
 text-internal perspective 256,
 260
 Phenetics 43, 44
 Philologist, role of 32-34, 36
 Pilot tree 142, 146
 Possible weight 115

 Quadruple(s) 145, 153, 154, 155,
 249-251, 253-255
 equivocal quadruples 253
 unequivocal quadruples 253
 Qualitative
 evaluation 249, 250, 254, 255
 follow-up 249, 250
 label 251
 Quantitative
 approach 249, 250
 distance 253, 254
 evaluation 255
 method 250, 260
 quantitatively determined
 dichotomies 253
 rate of variants 260

 Reading 151
 (cf. also: Variants)
 Reconstruction 252, 260
 Redactional work 259
 Rhyme 276, 277, 278, 279, 282

- Rhyming
 conventions 33
 position of variants 24, 25, 29,
 30, 32, 33
- Shock waves 117
- Similarity score 109, 126, 140, 141,
 149
- Slavia Orthodoxa* 233, 244
- Spelling regularisation 75
- Spread 113
 spreading 111
- Statistics
 statistical method 249, 260
- Stemma(ta) 250, 252, 253, 255, 256
 (cf. Chain, bundle)
 oriented 288
 unoriented 288
- Stemma codicum* 209, 210,
 214-218, 223, 224, 228
- Stemmatics 72, 76, 78, 81, 88,
 89, 96, 97
- Structure 250, 254
 acceptable deep structure 251
 actual deep (underlying)
 structure 252
 actual deep structure 250-253,
 257, 260
 actual underlying structure
 250, 252
 deep structure 249, 251, 252,
 254, 257
 oriented structure 252
 structural analysis 252
 structural variations 241
 underlying structure 251
- Subject matter 210, 219-221, 227,
 230
- Successive changes in relationship
 253
- Superiority / inferiority of readings
 237
- Support (dichotomy) 149, 151
 supporting percentage 159
- Synoptic text edition 12, 35, 36
- Systematics 42, 43
 biological 42, 49, 51
 evolutionary 43
 phylogenetic 44
- Taxonomy 60
 numeral taxonomy 43
- Terminal manuscript 137, 138
- Text
 family 234-241
 textological relevance 239
 textology 236
 textual witness 210, 215, 217,
 220, 225
 textual palaeontology 241, 243
 transmission 210
- Texts
 *Abbreuiatio Gestorum Regum
 Franciae* 215, 228, 229
 alphabet acrostich 243
 Apophthegmata Patrum
 (Systematic Collection) 234
 Bede, *De Natura Rerum* 226,
 227
 Callisthenes (Pseudo),
 Alexander Romance 211,
 212
 Charroi de Nîmes 283, 286
 Chaucer (Geoffrey), *Canterbury
 Tales* 89, 90, 95
 Chaucer (Geoffrey), Wife of
 Bath's Prologue 72, 74, 75,
 77, 89-92, 95-98, 100
 *Collatio Alexandri cum
 Dindimo* 210-213, 218, 222
 Collatio II 212
 Collatio III 212
 Dares Phrygius, *De Excidio
 Troiae* 218, 223
 Didactic Gospels 243
 Dudo of Saint-Quentin, *Gesta
 Normannorum* 226, 228,
 229

- Einhard, *Vita Karoli Magni*
215, 216, 218, 223, 226, 228
- Epistola Alexandri ad Aristotelem* 210-219, 222, 223, 225-230
- Epistola II* 212
- Epistola Presbyteris Johannis*
223, 224
- Genealogia Comitum Flandriae*
215, 226
- Geoffrey of Monmouth,
Historia Regum Britanniae
215, 218, 223, 224, 227, 228, 229
- Gildas (Pseudo), *Historia Brittonum* 215, 225-229
- Hegesippus, *Historiae* 216, 226
- Helinand of Froidmont,
Chronicon 210
- Historia Appollonii Regis Tyrii*
223
- Historia de Preliis* 211-213
- Iter ad Paradisum* 213
- Iter Ierosolimitanum Caroli Magni* 228
- Julius Valerius, *Alexander Romance* 211-213
- Leo Archipresbyter of Naples,
Nativitas et Victoria Alexandri Magni 211
- Ordericus Vitalis, *Historia Ecclesiastica* 215, 227
- Oxford-Montpellier Epitome*
213
- Paulus Diaconus, *Historia Langobardorum* 216, 218
- Seneca, *Epistolae Morales ad Lucilium* 223
- Seneca (Pseudo), *Epistolae ad Paulum et Pauli ad Senecam*
218, 223, 224
- Solinus, *Collectanea Rerum Memorabilium* 223, 224
- Svipdagsmál* 72, 80-84, 86, 87, 89, 95-98
- Treatise On the Letters* 235, 237
- Turpinus (Pseudo), *Historia Karoli Magni et Rotholandi*
223
- Vincent of Beauvais, *Speculum Historiale* 210, 213, 218
- William of Jumièges, *Gesta Normannorum Ducum* 215, 223, 224, 226, 228, 229
- Zacher-Epitome* 210-213, 215, 216, 218, 223, 226-229
- Tradition, closed 233, 238, 239, 241, 244
- Traditions
definition of 170
description of 178ff
Muslim study of 172
variants 171, 179, 180, 188, 198
Western study of 175
- Transcription 72, 74, 81, 90, 94, 99, 100, 240-244
- Transmission
aspects of 194-196
context 209, 217, 223, 230
dossier 220, 222, 224
lines 224
manuscript 218
oral and written 171, 172
process 221, 222
relations 185-187
terms 184
text 210, 224
time 186ff, 194
types 173ff
- Transmitter
definition of 170, 184ff
identification of 180ff, 198
- Tree 106, 123, 130, 138, 144
- Triple(s) 153, 154, 155
- Triumph of Orthodoxy 233

- Type 1 variant 155
- Type 2
 - difference 109, 125, 140, 155
 - limitation 19, 21, 22
 - opposition 263-268
 - variations and groups 4, 17, 19-21, 26, 29, 33, 34
(cf. also Monophyletic group)
- Utrum in alterum abiturum erat?* 236
- Variantenträger* 252
- Variants 250-260
 - archetypal readings 258
 - coincidental readings 251
 - common readings 258
 - common variants 256
 - distinguishing variants 258
 - genealogical variants (or: relationship revealing variants) 6, 29-32, 41, 43, 61, 62
 - good readings 252, 253
 - groups connected by readings 260
 - identical readings 250
 - inconsistent readings 252
 - inconsistent variants 250, 254
 - individual variants 258
 - intentional alteration 257, 258, 260
 - intentional variants 257
 - local 277, 283, 284, 286, 287
 - non-local 277, 282, 284, 287
 - original reading 259
 - original readings 252, 257, 258
 - original text 252, 259
 - original variants 258
 - orthographical mistakes 260
 - trivial differences 6, 10, 35
(cf. also: Parallelism)
 - unconscious alteration 257, 260
 - variants connecting manuscripts 257, 258, 260
 - variants separating manuscripts 257
- Variation
 - accidental 8
 - coincident 8
 - complex 19-22, 24
 - directional 20
 - fundamental 19
 - genealogical 32
 - graphical 11
 - multi-state 22
 - place of 15, 19, 21, 24, 29, 31, 33, 36, 59
 - semifinal 40
 - simple 19, 20
 - split-up 46
- Version formula 124, 125, 139
- Viertallen* 249
- Volume 218, 220-222, 225, 230, 270, 271
- Wagner network method 4, 41, 44, 62-70
- Weight factor 155
- White noise 119, 120
- Witness against contraction 160
- Word
 - boundaries 15, 16, 18, 30
 - categories 16, 24-30
 - classes 16, 17, 25, 29, 30
 - order 4, 10, 16, 17
 - types 4, 24, 25, 29