Tools, Evaluation and Preprocessing for Stemmatology

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 12
der Johann Wolfgang Goethe -Universität
in Frankfurt am Main

von
Armin Hoenen
aus Krefeld

Frankfurt (2018)
(D 30)

vom Fachbereich 12 der
Johann Wolfgang Goethe - Universität als Dissertation angenommen.

Dekan: Prof. Dr. Uwe Brinkschulte

Gutachter: Prof. Dr. Alexander Mehler, Prof. Dr. Visvanathan Ramesh

Datum der Disputation: 19. 12. 2017

Contents

Abstract

Die vorliegende Arbeit beschäftigt sich mit dem Thema Stemmatologie, d.h. der Rekonstruktion und geeigneten visuellen Darstellung der Kopiergenealogie handschriftlich fixierter Dokumente. Im Mittelpunkt dieses Wissenschaftszweiges steht die Frage des Autorenoriginals, falls ein einziges solches existiert haben sollte (s.u.) und die Frage der Rekonstruktion des autoreigenen Textes. Dieser durch für manuelle Kopierprozesse kennzeichnende Abweichungen zunehmend abgewandelte Originaltext ist meist nicht direkt überliefert. Dies impliziert die Menge T aller Manuskripttexte der Tradition[1], sowie die Teilmenge T' der überlieferten Texte.

Ziel der Arbeit ist es, die semi-automatische Stemmatologie, die bereits kurz nach der kommerziellen Verfügbarkeit von Computern in den späten 50er Jahren begonnen hatte (Ellison, 1957) und in den späten 80er Jahren durch einen Technologietransfer aus den Biowissenschaften transformiert wurde, siehe auch Robinson and O'Hara (1996), durch analytische Verfahren (technisch

[1]Eine Tradition umfasst alle in der Überlieferungsgeschichte enstandenen Handschriften eines Textes.

und hermeneutisch) weiterzuentwickeln, indem Entwicklungsmöglichkeiten aufgezeigt und technische Hilfsmittel entwickelt und vorgestellt werden. Beim letzteren dieser Punkte zielt die Arbeit darauf ab, evaluierbare Ergebnisse vorzuweisen und bedient sich deshalb einer Reihe vorhandener Datensätze für die die Kopiergeschichte bekannt ist.

Die Arbeit beginnt mit einer allgemeinen Einführung in die Entstehungsgeschichte des Fachgebietes. Die Stemmatologie als genuine Wissenschaft, insbesondere die digitale, blickt auf eine vergleichsweise kurze Geschichte zurück, deren Wurzel in etwa mit dem Beginn des europäischen Printzeitalters korreliert. Sehr frühe textgenealogische Untersuchungen wurden vor allem in der jüdischen Gemeinschaft durchgeführt (Wegner, 2006). Mit dem Beginn des Printzeitalters mußte eine Entscheidung bzgl. derjenigen Version getroffen werden, welche als immergleiches Printexemplar fürderhin die Tradition repräsentieren sollte. Um den dafür bestmöglichen Text zu erfassen (aus der Menge T' heraus oder aber mit Hilfe von T' rekonstruiert) hat die Editionsphilologie verschiedene Ansätze entwickelt, die eng mit der Stemmatologie in Verbindung stehen. Nach einer Rekapitulation dieser historischen Entwicklungen wird auf neue Möglichkeiten digitaler Stemmavisualisierungen eingegangen und ein Tool vorgestellt, welches Bäume im Newick Format,[2] welches ein verbreitetes Exportformat bio-informatischer Software ist, mittels einer Kombination aus Java und LaTeX in eine Baumdarstellung transformiert bei der im Gegensatz zu bioinformatischen Standardbaumvisualisierungen die Blätter nicht im Fokus stehen. Des Weiteren ist die Software in der Lage dynamische Darstellungen zu produzieren. Auf andere Möglichkeiten der Stemmavisualisierung, z.B. als *Circular Tree Map* mit unterlegter Karte wird ebenfalls eingegangen. Die Grundlagen zum Abschnitt sind als Hoenen (2016a) publiziert.

Aufgrund der besonderen Relevanz der stemmatologischen Debatte um Joseph Bédiers Beobachtung, dass die meisten rekonstruierten Stemmata wurzelbifurkativ waren (in seiner und anderen Kollektionen) (Bédier, 1928) geht das nächste Kapitel genauer auf die Debatte ein. Für Bédier war das inhärente methodologische Prozedere zur Stemmaerstellung die plausibelste Erklärung für die für ihn überraschende Vielzahl wurzelbifurkativer Stemmata. Einerseits nahm er an, dass Editoren sich durch das Postulat von nur genau zwei Unterfam-

[2] `http://evolution.genetics.washington.edu/phylip/newicktree.html`

ilien eine Freiheit in der Wahl der Varianten des Urtextes schaffen, andererseits, dass eine starke Tendenz zur Überseparation bestünde; d.h. dass beispielsweise statt einer Trifurkation häufig eher zwei Bifurkationen angenommen werden, da empirisch zu erwarten ist, dass ein Paar aus drei Kindern des Elternknotens sich einander am ähnlichsten ist und fälschlicherweise im Stemma damit ein neuer Zwischenknoten für diese beiden postuliert werden kann. Daher stellte Bédier stemmatologische Ansätze grundsätzlich in Frage und initiierte eine Editionsmethode, die auf der Wahl und Edition eines besten Manuskriptes beruhte. Die Debatte darum, wie methodologisch kontaminiert Stemmatologie ist, wird seitdem geführt und hat viele Publikationen stimuliert, von denen die wichtigsten oder einflussreichsten im Kapitel zusammengefasst werden. Ein Argument, welches u.a. von Fourquet (1946); Maas (1937); Weitzman (1982); Castellani (1957); Whitehead and Pickford (1973); Trovato and Guidi (2004) vertreten wird, besagt, dass aufgrund der mathematischen und historischen Natur des Problems eine große Zahl an wurzelbifurkativen Stemmata erwartbar seien und dass damit Bédier's Beobachtung nicht so sehr verwundern sollte, dass dadurch die Stemmatologie gänzlich abgelehnt wird. Die vorliegende Arbeit untersucht die Proportion wurzelbifurkativer Stemmata in gewurzelten bezeichneten Bäumen der Größe n allgemein (diese stellen alle möglichen wahren und vollständigen Manuskriptstammbäume dar und werden von Fourquet (1946) als arbre réel, kurz Arbre bezeichnet). Für historisch relevante n ist die Proportion tatsächlich groß(ca. 0.373 bei $n = 100$) und die wurzelunifurkativer Bäume ebenfalls (ebenfalls ca. 0.373 bei $n = 100$), womit wurzelmultifurkative Bäume je nach Furkation deutlich unwahrscheinlicher sind. Neben den Proportionen für Arbres, erarbeitet der Autor die Proportionen von Wurzel-k-Furkationen für sog. *Greg Bäume* (Flight, 1990). Greg Bäume (so benannt nach dem Philologen W.W. Greg) sind solche Bäume, die alle möglichen Rekonstruktionen von Genealogien bei m überlebenden Manukripten darstellen. Es gibt dabei zwei Arten von Knoten, bezeichnete (überlebende Manuskripte) und unbezeichnete (hypothetische, verloren gegangene Vorlagen). Es gilt die philologische Praxis, möglichst keine *codici interpositi* zu postulieren, siehe auch Haugen (2015), d.h. keine unbezeichneten Knoten zu postulieren, die einen Eingangsgrad und Ausgangsgrad von jeweils 1 haben. Ein Greg Baum definiert sich also als Baum, der aus m bezeichneten und $n = 0 \ldots m-1$ unbezeichneten Knoten besteht, wobei unbezeichnete Knoten mindestens Grad 3 oder als Wurzel im gewurzelten Greg Baum

mindestens Grad 2 haben müssen. Die Arbeit leitet eine Formel zur Berechnung wurzel-k-furkativer Greg Bäume her, tabuliert und zeigt, dass für diese Bäume die Proportion der Wurzelbifurkativen im Rahmen der historisch relevanten Größen mit 0.606 bei $m = 100$ noch größer als bei den Arbres ist.

Anschließend erarbeitet der Autor eine Massensimulation von Stemmata, um die Auswirkungen von starkem Manuskriptverlust in der Geschichte mit möglichen unterliegenden Distributionstypen zu erörtern. Bereits Greg (1931) hatte starken Manuskriptverlust mit Furkationsmustern in Verbindung gebracht und wurde u.a. durch Weitzman (1982); Trovato and Guidi (2004) darin weitgehend bestätigt. Der Einfluss von Distributionstypen der Manuskriptausgangsgrade (der historischen Kopierdynamik) wurde zwar in vielen philologischen Publikationen beschrieben, jedoch noch nicht Subjekt einer systematischen Untersuchung. Mittels der Generierung zufälliger Werte aus 7 Distributionstypen (binomial, exponential, geometrisch, hypergeometrisch, normal, zipfsch, zufällig) durch die Softwareumgebung R werden Ausgangsgrade für Knoten in Arbres ermittelt, diese Arbres aufgebaut (Java) und einem von 3 Verlustszenarios unterzogen. Die übrig gebliebenen Knoten werden dann Grundlage des zu rekonstruierenden maximal richtigen Stemmas, welches keine *codici interpositi* und ansonsten alle Knoten auf dem Pfad zwischen dem letzten gemeinsamen Vorfahren aller Überlebenden (Archetyp) und diesen als hypothetische Knoten enthält. Für ein solches Stemma werden Furkationen gezählt. Die Verlustrate berut auf den Zahlen aus Trovato (2014). Durch den starken Verlust ($> 73\%$) zeigt sich unabhängig von der zu Grunde liegenden Distribution eine erwartbare Vielzahl an Unifurkationen, die zahlreicher als Bifurkationen sind, welche wiederum zahlreicher als Trifurkationen sind usw. Werte aus Kollektionen an Stemmata, die von Haugen (2015) untersucht wurden, stimmen zwar weitgehend mit diesem Muster überein, enthalten aber wenige Unifurkationen. Die Erklärung hierfür könnte methodologische oder historische Ursachen haben und wird diskutiert. Wurzelfurkativität wird zudem vorraussichtlich durch baumtopologische Faktoren wie Symmetrie mitbestimmt, wie schon Trovato and Guidi (2004) andeuten und wie weitere Ergebnisse der Simulationen ebenfalls vermuten lassen. Die Simulation ist eine Erweiterung der Publikation Hoenen (2016b).

Der zweite Teil der Arbeit beginnt mit der Beschreibung der 4 Datensätze, die zur Evaluation in den folgenden Kapiteln herangezogen werden. Diese heißen *Parzival* (in englischer Sprache) (Spencer et al., 2004a), *Notre Besoin* (Französisch)

(Baret et al., 2004), *Heinrichi* (Altfinnisch) (Roos and Heikkilä, 2009), und *TASCFE* (Hoenen, 2015a). Letzteren hat der Autor im Zusammenhang dieser Arbeit produziert und digitalisiert und beschreibt ihn *en detail* im letzten Teil des Kapitels. Im nächsten Kapitel wird zunächst die Hypothese geprüft, die besagt, dass durch visuelle Verwechslung entstandene Abweichungen hauptsächlich für die Genealogie verantwortlich seien, vgl. Reynolds and Wilson (2013, S.223ff) und Wegner (2006, S.45). Dabei erkennt man die Vorlage-Kopie Paare innerhalb der Menge aller möglichen Manuskriptpaare dadurch, dass die darin vorkommenden Abweichungen, vor allem die besonders leicht verwechselbaren Buchtsabenpaare erkennen lassen. Dies geschieht einerseits anhand einer qualitativen Fehleranalyse des Parzival und des Notre Besoin, sowie andererseits durch die Implementation eines Algorithmus, der eine Vielzahl psycholinguistisch gewonnener Buchstabendistanzmatritzen (akustisch, visuell und motorisch) zur Gewichtung von Manuskriptdistanzen einsetzt. Bei einem Manuskriptpaarvergleich wird hierbei jedes sich entsprechende Wortpaar beider Manuskripte verglichen und aligniert. Mit Hilfe einer psycholinguistisch gewichteten (Levenshtein) Distanz (Levenshtein, 1965) werden pro Wortpaar die visuelle und motorische Distanz berechnet, während die akustische Distanz durch einen zusätzlichen Mechanismus berechnet wird. Eine Abbildung der paarweise alignierten Grapheme auf alle durch sie möglicherweise dargestellten Phoneme wird vorgenommen, wobei die Graphem zu Phonem Relationen aus der einschlägigen Fachliteratur des Englischen, Französischen, Finnischen und Persischen stammen. Für jede Paarung möglicher Phoneme wird dann durch Rückgriff auf eine ebenfalls psycholinguistisch gewonnene akustische Phonemdistanzmatrix (Cutler et al., 2004) oder durch Extrapolation ein Distanzwert ermittelt und über alle Paare gemittelt, um den akustischen Verwechslungswert zu konstituieren. Ein Gewichtungsparameter gewichtet und summiert dann die modalen Distanzen gegeneinander, wobei das Mittel für jedes Wortpaar den Distanzwert und das Mittel über alle Wortpaare den Manuskriptpaar Distanzwert konstituieren. Aus der Matrix aller Manuskriptpaardistanzen wird dann ein Stemma automatisch generiert (Neighbour Joining Algorithmus (Saitou and Nei, 1987)). Evaluiert wird mittels eines graphbasierten Baumvergleiches, der Average Sign Distance (ASD), die von Roos and Heikkilä (2009) entwickelt wurde. Die Ergebnisse weisen darauf hin, dass die durch multimodale Ähnlichkeiten verursachte Verwechslung zwar tatsächlich den größten Teil der genealogischen Information

widerspiegeln, dass aber die Modellierung der nicht abgedeckten Prozesse noch weitere Verbesserungen versprechen. Eine Korrelationsanalyse der Parameterwerte und der Evaluationergebnisse zeigt, dass die Korrelation des akustischen Parameters, wobei im Akustischen als einziges Relationen auf Basis von längeren graphemischen Einheiten (Länge > 1) trainiert werden, für das Englische und Französische besonders stark positiv ist und somit auf einen Einfluss der orthographischen Tiefe (Katz and Frost, 1992) schliessen lässt, da diese längeren Einheiten in diesen beiden Sprachen verbreitet sind, während im Finnischen selten mehr als 1 Buchstabe für mehr als 1 Phonem steht, so dass hier die psycholinguistischen visuellen und motorischen Matritzen, welche $1:1$ Verwechslungen befassen besser greifen können.

Schließlich werden besonders gute Ergebnisse durch die Anwendung von Minimum Spanning Trees (MST) auf aus dem implementierten Leitfehlerkalkül (Roelli and Bachmann, 2010) abgeleiteten Distanzmatritzen gewonnen. Wendet man den Kruskal (Kruskal, 1956) oder Prim (Prim, 1957), Algorithmus zur Generierung eines MST auf Basis der durch den Leitfehler generierten Distanzmatritzen an, so generiert er für den Parzival ein Ergebnis, bei dem nur eine einzige Kante falsch erkannt wird, was nach bestem Autorenwissen das bisher beste Ergebnis auf den artifiziellen Datensätzen wäre. Die Methode geht aus der Kombination philologischen Kalküls mit informatischen Methoden hervor. Im Kapitel zu den MSTs wird ein umfassender Vergleich von 4 Methoden zur Distanzmatrixgenerierung und 4 verschiedenen baumgenerierenden Algorithmen vollzogen, der dieses Resultat bestätigt.

Terminology, symbols and conventions

The term *letter* refers to a basic unit of a(n alphabetic) writing system, compare Costard (2011, p.12). This implies that diacritics are no letters and that lower case and upper case letters are different letters. Coltheart (1981) introduced the term *abstract letter identity* which refers to cross case, cross font representations of the same letter (for instance <A> and <a>) in the sense of functional systematic unit. According to Costard (2011, p.16), there are different inconsistent usages of another additional term in the literature, the term *grapheme*, which is in close parallel to the term phoneme. As in Cook and Bassetti (2005, p.4), following Sproat (2000), the term grapheme is used similar to the term letter as term for the smallest unit of a writing system without a connection to phonemes. When referring to a group of letters, which are used as a symbol for a certain phoneme, the compound *graphemic unit* may be used.

As a convention, when writing about phonemes, the signs used to represent them fixed in/as the Interantional Phonetic Alphabet (IPA) are enclosed in slashes. The phoneme /p/ is written in this way. Graphemes and grapheme sequences are usually enclosed by pointed brackets.

Examples:

- letter: a, A
- abstract letter identity: (A,a), (B,b)
- phoneme: /p/
- sequence of letters under discussion: <abc>

Rounding to the second or where values were small to the third digit is implicitly used. Clarifying translations from French, Italian or German into English have all been made by the author.

Acknowledgements

First of all, I want to thank the Persian students, who were endlessly kind in providing the basis for the TASCFE corpus, in Tehran and in Frankfurt demonstrating the exceptional hospitality and politeness of their people:

مرحمت شمَا خيلي زيَاد است

Without their work, a substantial part of this thesis could not have been completed. Along goes a thank you for the German volunteer, who provided the sample of a copied text from a writing system unknown to him.

Furthermore, I want to thank my supervisors. I want to thank Prof. Dr. Alexander Mehler for his aid in the design of the MMD formula; the impulse to use Minimum Spanning Trees was due his input. His feedback and innumerable comments and provisions of references have helped improving the work in substantial ways. I want to thank Prof. Dr. V. Ramesh for his readiness to supervise this thesis and the honest discussions on style and content which hopefully likewise found a way into effect for the thesis and some of its chapters in the final form.

I also want to thank my colleagues from the TTLab and the empricial linguistics department (Dr. A. Lücking, Dr. Md. Z. Islam, P. Sauer, Dr. Z. Pourtskhvanitse, R. Gleim, G. Abrami and many others), who provided resources, always had an open ear for and actively engaged in questions and discussions about individual topics, provided organisational help and advice and encouraged me frequently to continue. I am especially deeply indepted to Dr. S. Eger and Dr. R. Gehrke in this respect. Without their suggestions, help and openness, the current thesis may not have come into being. A special thanks goes to the student assistent Rashedur for helping to extract some of the tables and performing various other tasks. Especially Prof. Dr. Dr. h.c. mult. Jost Gippert, who was also member of the examination commitee, provided access to valuable resources, rec-

ommended literature and encouraged and supported me continuously through letting me participate in debates and the writing of applications on a larger, pan-European scale especially in connection with Corpus Avesticum, the members of which another special thanks goes to; above all to Prof. Dr. A. Korn, who gave me the chance to visit Persia. At the more difficult points of the writing process, consulting with him calmed me and helped me go on. I also want to thank him for being ready to supervise the thesis, which also applies to Prof. Dr. C. Chiarcos, even though in the end this did not come into effect.

Colleagues I met at conferences and those, I only know from email contacts have provided data and critical comments confirming the best ethics and highest moral standards of science, and I want to thank them, especially C. Macé, R. Wiley, P. Roelli and his colleague D. Bachmann, as well as T. Roos and T. Heikkilä and their team.

I want to thank the state of Hesse, the Federal Ministry of Education and Research and the Goethe University for providing the financial backbone of the projects connected with this thesis.

Last, but not least, I want to thank my family, my parents, who have sustained me and brought me to where I am. My wife Nicole, who stood with me and encouraged me in the harder times, when one questions if it is possible at all to achieve a completion. She helped me cope with an illness acquired recently more than anyone else. Our daughter Benazir sustained me with her smile more than I can express in words.

Publication Statement

Substantial parts of this thesis have been published. In no case does the section entirely overlap with the publication which is due to space limitations in the publications and posterior reformulations.

In the introduction (literature and metadata) material overlapping with the publication Hoenen (2014b) is to be found. Stemma visualisation is connected with Hoenen (2016a) and the simulation of distributions of outdegrees of nodes in arbres is partly congruent with Hoenen (2016b). The combinatorial section has been published in extended form as Hoenen et al. (2017). The TASCFE corpus' description is an extended version of the publication Hoenen (2015a). In

conjunction, the corpus presented there is publicly available.[3] The section on Multi Modal Distance is presented in Hoenen (2018) and the chapter on Minimum Spanning Trees is being prepared for publication together with Alexander Mehler. The thesis was generated using the LaTeX software.

[3] https://www.hucompute.org/ressourcen/corpora

Preface

The present thesis is concerned with problems in the analyses of historical documents. Primarily it focusses on the question how to automatically reconstruct the copy history of an ancient tradition of digitized hand written manuscripts. Alongside this aim, the thesis presents tools for stemmatological and prestemmatological processing.

A stemma as the visual representation of the copy history of a manuscript tradition displays two basic things: the manuscripts and the copy processes. In graph theoretical terms, a stemma is a tree very appropriately comparable to a family tree of an aristocratic family, which was the first application, for which such ”trees of history”, as O'Hara (1996) has called them, have been in use. The nodes represent manuscripts and the root node itself the authorial original, provided there is such. The edges map to copy processes. Many ancient texts have come down to us in a staggering variety of versions, others only in one or two manuscripts in which case only the last chapter of the thesis may be interesting. To disentangle the complexity hidden within this variety, to point to the genealogical supremacy of certain manuscipts over others of one and the same text - genealogically (not from the point of view of the real historical authoritativity that has been excerted by them)- and to enable a modern editor to provide a representative text; these are the benefits of the stemmatic endeavour. A crucial danger on the other hand lies in the very nature of historical data. The ground truth is forever hidden. This makes historical corpora themselves dangerous objects for the evaluation of algorithmic solutions. This is not to say that an expert working in a thorough way cannot achieve reasonable results without commiting a cyclic reasoning, but if one uses artificial traditions, benchmark data sets - that is data sets which have been created by volunteers in recent times under record of the true ground truth - then, one can be sure not to commit such an error and as a novice, the author chose this line of approach.

The current thesis will recapitulate the remote and recent history of stemmatology, its visualization component and summarize the theoretical debate initiated by Joseph Bédier (Bédier, 1928) which calls into question the whole stemmatic endeavour. This first part, entitled 'Literature and Theory' closes with a simulation of manuscript trees and aludes to lessons that can be learned from the theoretical and historical perspective.

The thesis proceeds introducing the artificial data sets used for evaluation and then present approaches to the digital reconstruction of the copy history. This part is entitled 'Attempts at Reconstruction'. It closes with a summarizing reflection. Among the texts, the thesis aims at are the oldest non silent witnesses of mankind (or of their respective cultures).

Part I

Literature and Theory

General Introduction

The general overarching field this thesis is located in, is computer-assisted stemmatology. By this, in the author's opinion, one refers to all methods, processes and approaches using the computer in any task directly connected with the solution to the question of how to deduct and (visually) display the copy history of manuscripts or rather their texts. In mathematical and computer-scientific terms, the preferred way of modelling the copy history is by means of graph theory, especially but not exclusively as trees. As the article by den Hollander (2004), which is published within a volume on the science of (computer-assisted) stemmatology, shows, computer-assisted stemmatology extends to tasks related to identifying and inferring copy relationships between texts. Computer-assisted stemmatology is a field of applied computer science on the one hand and a theoretical melange of philological and mathematical reasoning on the other. Therefore in its character, it might justifyably be located in what is used to be called 'digital humanities' or 'humanities computing'. In this chapter, especially in subchapters 1.4, 1.7 and 1.8 some phrases may overlap with material in Hoenen (2014b), while subchapter 1.8 through 1.12 are an elaboration and translation into English of Hoenen (2016a).

Chapter 1

Literature

1.1 Introduction

In this chapter, the theoretical prequisites of stemmatology, the historical development and the existing literature are summarized ending in a survey on recent developments in the digital sphere. The chapter begins with a theoretical explanation of the historical circumstances surrounding stemmatology, followed by a brief historical sketch of the philosophical roots of stemmatology in antiquity, its early development and emancipation as a branch of science and closes with a summary of the recent developments in the digital medium. This section itself is kept as narrative and ultimately quite 'wordy' not unlike some philological discourse. It tries to construct a brief but holistic picture of the field's history avoiding too coarse definitory delimitations. More operational definitions are to be found in the following chapters.

1.2 Preliminaries

A *stemma codicum*[1] is a visualisation of the copy history of manuscripts containing one text.[2] In terms of graph theory, typically the manuscripts are nodes

[1] As for the term stemma, both stemmas and stemmata (this alone extends to stemmata codici) are possible plurals. Throughout this thesis they will be used interchangeably for stylistic or other purposes. Given, that some publications use the one others the other, it is inevitable for a thesis on stemmatology to employ both.

[2] This entails any graphical representation of manuscript genealogies including bracketing structures.

and a *vorlage-copy*[3] relationship is represented by an arc, the overall topology is typically that of an acyclic rooted tree. The root node represents the authorial original or the latest common ancestor of all surviving manuscripts, then called the archetype. Usually all nodes are labelled. However, surviving manuscripts are often labelled by Latin, hypothetical ancestors by Greek letters. The visualisation has been developed by philologists in order to better overview, understand and evaluate the body of manuscripts at their disposal and to inform the reader about the basis of editions.[4]

The focus of this thesis is in principle on stemmatology for historical manuscripts, that is handwritten documents (chirography). Chirography is to be understood within a broader context (see Ong (2012); Foley (2002)) baring several implications. The context is a theory of succession of language technological stages, which crucially affected the transmission of texts. The four assumed stages are:

1. orality
2. chirography
3. print
4. digital.

Orality has shaped certain text types, typically verse text, which served as role models for the first hand written fixations, compare Lord (1960). However, the impact of orality is much greater than that. New text types and modes of presentation developed after the introduction of script and constantly mixed old oral and new written techniques (Amodio, 2004, p.93). Print technology causes the end of dominance of hand writing as the main means of diffusing knowledge, but the shift from handwriting to print was not abrupt and displays various facets, compare Reynolds and Wilson (2013, pp.155). Print technology stays thus implicationally relevant to those handwritten works which were written after the invention of locally accessible print technology.

Each technological stage has its own dynamics and follows a schema. In terms of media theory, for instance, certain remediation processes happen with the invention of a new medium. Bolter and Grusin (1999, p.45) identify the spectrum of remediation from an emphasis on the old (imitation of the remediated)

[3]Vorlage is an English loan from German and means original, or model, of a copy.

[4]Because of the nature of historical data, where the ultimate truth is not verifiable, a stemma usually represents a hypothesis, comp. Bordalejo (2015).

where the new medium imitates to an emphasis on the new. Both ends can be attested in the case of the transition from manuscripts to printed books. Consider also Amodio (2004) who describes in detail the transition of oral to written for medieval England. For each culture, the time of introduction (invention) of script and print differs (partly drastically) and so does the absolute temporal localisation of the technological stages. [5] The most important aspect for stemmatological analysis is the impact of remediation from oral to written on the deviations between the manuscripts. Namely, the kinds of deviations or errors which occur are in many different subtle ways influenced by the aforedescribed constantly developing technological pretext. For instance, inspired by orality, where each oral poet presented a song in his own way, early scribes may have felt freer to improve a rhyme or wording when copying a manuscript, whereas later, monchs rooted in the study of texts deeply immersed in literacy had been taught to preserve the text as well as they could. Also, early scribes may not have been aware of how much copy errors can change a text over longer copy chains, which simply did not exist. The practices of commenting, writing variants into the margins and so on are all secondary phenomena.

1.3 Antiquity and the Middle Ages

Stemmatology proper does not come into being as a scientific discipline before the onset of the print age, compare Reynolds and Wilson (2013); Timpanaro (2004); O'Hara (2006); Bordalejo (2015). This is unsurprising given the large spread of orality and given that availability of manuscripts was more important than detail.[6]

Casson (2002) shows that 'Systematically organized, institutionally spon-

[5]For East Asia and especially China for instance, woodblock printing is at least around 1500 years old (Taylor and Taylor, 2014), and so the print age with printed books in large numbers started much earlier there than in Europe where only around 1450, Gutenberg invented the printing press and unleashed a new age. For yet other cultures, there may be a *jump* or transition directly from the oral into the print age.

[6]Whereas a printed page is identical to the same page in the subsequent print of the same edition and volume (sufficient ink and intact types provided), a perfect manual copy for longer texts is empirically close to impossible. While each printed <e> looks the same and has the same proportions, lengths, angles etc., each handwritten <e> will differ within certain limits from any other in proportions, line width etc.

sored comparison of manuscripts expressly for this purpose' dates back at least to the founding of the library at Alexandria, see also Cisne et al. (2010). Casson (2002); Dearing (1970) mention Zenodot, the first director of the Great Alexandrian Library under Ptolemaios the 1st, who compiled a standard version (edition) of the Homeric works 'Odyssee' and 'Ilias' thereby encountering, enumerating and describing variants in the manuscripts. If a stemma was involved in his works cannot be said, since evidence has not come down to us.[7] Discussions on how to deal with errors, which are similar to later genuinely stemmatological discussions certainly had been held, a particular case being the Hebrew scriptures. The consonantal Hebrew alphabet first represented mostly consonants (Wegner, 2006, p.65) which allowed ambiguity (just as <rd> could stand for 'read', 'red', 'rod' etc.). This was especially relevant for holy texts which some of were to be read out aloud. In consequence, extensions to this so-called abjadic writing system were developed, fixing short vowels in order to generate authoritative, enduring and detailed versions of the holy texts (Wegner, 2006). Along with this, awareness of errors increased and the discourse induced strict and stringent rules for copyists in the Tannaim group (Wegner, 2006, p.73). Such and similar discussions, especially around sacred texts but not less for classics continued during the middle ages also for Latin and Greek and eventually led to all kinds of strategies of coping with textual drift. One particular phenomenon arising in this vast time span as a strategy for coping with variation was so-called *contamination*, an umbrella-term under the hood of which many phenomena are summarized where more than one manuscript (or memorized text) was consulted when copying to a new exemplar. Some scribes, when encountering more than one possible 'readings' that is more than one plausible word/sequences of words to put into the copy (for instance in case there was a hole in the original) for one original word/sequence of words, they wrote one reading into the main text and another into the margins of the manuscript, sometimes commented upon, sometimes not. One prominent guideline for copying or devising editions dealing with variation in the philological discourse of the classics seems to have been the intuition on what the author himself would have considered the most appropriate version of the text, compare Najock (1995) on principles for Zenodot to the editing of

[7]Because of the oral character of the works, this would have probably looked somewhat different than modern stemmas or been prohibited altogether; the invention of one-original-based stemmatology may only arise in later deeply literate societies.

Homers works and arguments to this end.

1.4 The Advent of the Print Age - Just Before Stemmas

Shortly after the invention of the printing press, "from the seventies of the fifteenth century onwards", Reynolds and Wilson (2013, p.155), the first print editions of chirographically transmitted texts appeared. In the beginning, those were reproductions of regionally available manuscripts (Reynolds and Wilson, 2013; Cameron, 1987). The printer thus acted like an ordinary chirography age reader, he grabbed the closest available copy: "Accident and chance often determined what manuscript was used for early printed editions.", Cameron (1987, p.235). And by conjecture accident and chance had determined what version of a text a medieval reader had read. But, gradually philologists got aware of and began to systematize dealing with variation present in nearly all of the manuscript corpora of contemporary interest. The variation was often such, that for particular places in the text, two or more versions existed, which were sufficiently different to spark controversies about the original authorial intention. A or the key question, which brought into existence modern stemmatology was *which version or which text to base a print edition on.*

Above all, the sacred text of the Bible, which was/is at the same time the most widespread hand copied book, fuelled discussions on the correct editing techniques for print editions and presumably led to favourable financing conditions for such research. The systematic research on the transmission history however needed some 200 years to develop.

As was declared in 1734 by a biblical scholar named Bengel: "a perfect edition of the New Testament would propose a classification of the codices for their genealogical relations" (Pasquali, 1988, p.9).

In fact, since those times or shortly after, most editions started to give a stemma in their preface or appendix. Timpanaro (2005, p.92) attributes the first stemma in a modern sense to Carl Johan Schlyter, a Swedish scholar, who, nearly a century after Bengels claim published an edition of ancient Swedish legal texts with a stemma, see Figure 1.1. The earliest stemmas already carry all main attributes of their modern successors.

Since after 1500, methods for textual criticism were more and more discussed,

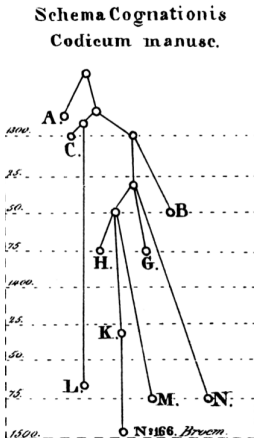


Figure 1.1: First modern stemma by Schlyter, 1827, from O'Hara (1996).

developed and refined in scientific and editorial contexts culminating in the 19th century (Reynolds and Wilson, 2013, pp.203). In the 1850ies, Karl Lachmann published his edition of Lucretius (Lachmann, 1853) which had a very influential preface, wherein a method is described, which Lachmann applied for deducing genealogical relationships and which is henceforth deeply connected with his name, even though some of the principles he employed were already in place when he published, compare the title of Timpanaro's book *La genesi del metodo del Lachmann* (Timpanaro, 2004). Under more Lachmann relies on significant errors (Leitfehler). The concept of error implies a hypothesis about the original and is not to be understood as the modern aberration from the orthographic norm. An error is an innovation within a copy, a substitute which can be equally orthographically acceptable, consider exchanging <clash> for <dash>. If such an innovation is found in a leaf of the stemmatic tree, that is a manuscript which has no surviving descendents, then it is not particularly genealogically informative since it cannot indicate any groups of manuscripts which go back to a common ancestor. Thus, a leitfehler as defined by Roelli and Macé (2015) as 'genealogically significant error' has to have some operationalizations on how to find such indicative errors (see for instance Salemans (2000)) but the procedures have also been called into question several times, compare Tov (1982); Andrews (2014). If any two manuscripts do share the same leitfehler, they can be assumed to go back to a shared common ancestor. The opposite are variants which arise accidentally in different copy processes representing noise for the genealogical indicativeness of textual similarity, see Roelli and Macé (2015). Slightly different

approaches have also been elaborated, see for instance Quentin (1926), who like Lachmann uses the construction of a stemma, but who proposes another schema of variant evaluation, which can lead to a rooted tree via an unrooted one. However, once in place, Lachmanns paradigm can be described as the dominant one for textual criticism of the short time period after 1850 until roughly 1928.

1.5 Stemmatology in crisis - the debate initiated by J. Bédier

The French philologist Joseph Bediér had looked at a collection of manually construed stemmas and observed a large amount of bifurcations at the root that is the assumed archetype or original having only 2 children from which all surviving manuscripts derive. Now, a bifurcation below the root node brings the editor into a dilemma if he or she tries to reconstruct the original authorial wording (associated with the root node). This is so, because if one child node would have a text contradicting the text of the other child node, the editor would have to personally choose one of the versions, comp. Bédier (1928, p.9) "à choisir par intuition, [...] de son gôut.".[8] Taken to the extreme, the criticism ultimately endangers (as does contamination) the whole stemmatic endeavour; alternatively choosing a best base text without compiling a stemma seems possible through for instance identifying and counting older variants, investigating the age of the manuscripts and so forth.

Spanning from the 19th century to the time, this thesis is compiled, a debate is ongoing about bifurcativity in manually reconstructed stemmas and the implication for editing and stemma building. Arguments range from purely statistical/combinatorial ones to purely epistemological ones. Many articles try to assess the manifold possibilities of an interpretation of the givens of history and their implications for Bédiers observation, hence for the amount of bifurcativity in constructed stemmata. The discussion is not concluded and it is questionable if this can be done, (Trovato, 2014; Haugen, 2015; Stussi, 1994; Balduino, 1989). The following sections will outline important contributions subsumed under contributing authors. The publications are made in different languages (often without translation) which is one of the characteristics complicating any

[8] to choose via intuition, [...] according to his taste

holistic assessment of the debate.[9]

1.5.1 Joseph Bédier

After some time of application of Lachmanns method (for instance by Bédier's supervisor), in 1928, the philologist Joseph Bédier reedited a medieval work, the "L'ai de l'ombre" a third time (after 1890 and 1913). He gave the edition an expressive subtitle: "réflexions sur l'art d'éditer les anciens textes".[10] In the preface, he gives account of a "surprising law" (loi surprenante), which he had discovered. Looking at 110 stemmas of different works of French medieval texts, he found (p.11) that:

> Dans la flore philologique il n'y a arbres que d'une seule essence: toujours le tronc s'en divise en deux branches maîtresses, et en deux seulement.[11]

105 out of the 110 stemmas he looked at had a root bifurcation,[12] whilst the remainder would be the exception that proves the rule. Although he acknowledged, that through loss, a third branch (below root) or family can be occasionally lost in its entirety, overall, the numbers he observes, in his opinion were far too large to be explained by this mechanism alone. Bédier describes how the search for the (one) correct against the (rest of) faulty reading(s) (force dichotomique) drives overseparation and leads to bifurcativity through not leaving the philologist at peace until the final task to establish the archetype text be achieved.

Similarly, one could assume overseparation due to humans superior ability to compare pairs of manuscript texts, rather than triples, quadruples or larger tuples. Pairs are not only easier to process but also easier to quantify in terms of similarity/distance.[13] Comparing three manuscript texts, one can compare three pairs and there is a large probability that if one applies this mode of pairwise

[9] The publication by Bédier mostly influenced French editing (Castellani, 1957). Consequently this led to French publications, which due to the time period and the prospective audience have partly never been translated. The same goes for some German and Italian articles.

[10] Reflections on the art of editing the ancient texts.

[11] In the philological flora there are none but trees of one single nature: the trunk always divides into two and only two major branches.

[12] Some authors use the terminology *bifid, trifid* etc. for root bifurcating, trifurcating etc. Apart from this, dichotomy comes to be used for bifurcating.

[13] The method described in Quentin (1926) for instance relies on pairwise comparisons.

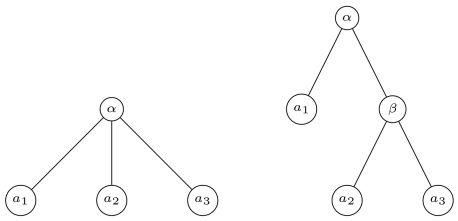


Figure 1.2: Left: true stemma. Right: Probable philological reconstruction due to a tendency to overseparate. Depiction similar to Figures in Hering (1967); Timpanaro (2004); Haugen (2010).

comparison, one will spot one pair of texts to be most similar even if all three have descended from the same parent. Consequently one would be tempted to postulate a common ancestor. The effect of any of the putative explanations is overseparation illustrated in Figure 1.2.

In this case, the philologist would postulate a common ancestor for the closer pair and attach the third sibling together with this ancestor to the parent node producing two bifurcations, where there was a trifurcation originally. Such is the reasoning for instance in Haugen (2002, 2015). Apart from this, Bédier wondered if a bifurcation at the root would not stem from the wish for freely chosing - de gusto - the base text. He had seen, that the stemmas of recensors were less bifurcating than those of editors.

In summary, Bédier observed a large amount of bifurcations and counted the root bifurcating stemmas in a collection of 110 stemmas, where he counted 105 root bifurcations. He believed this to be extremely strange, or a *Silva portentosa* (a strange forest) and blamed overseparation induced by the search for the correct variant and the wish of the recensors to choose after taste.

1.5.2 Implications of Bédiers criticism

Bédier's criticism was acknowledged by the community and inspired the invention of new edition types. Bédier himself would choose one manuscript, he deemed best as a base text and base his edition on this. According to Haugen (2002, p.9), this was a recurse to the old method of editing, which had prevailed before Lachmann and others had outlined the stemmatological method. However, if a best manuscript would be far removed from the authorial text through loss of the intermediaries, then a reconstructed text may be prefered by the user.

Hybrid approaches, which combine a reconstructed with all manuscript texts into one edition exist and finally Lin (2016) mentions narrative textual criticism as a way to keep track of the effect of all variation present. Nonetheless, Bédiers critique was rather polarising and many authors in the subsequent debates position themselves on one or the other side. The terms *neo-Lachmannian* and *Bédierist* editing have become prominent since then, compare Robinson (2000).

How important Bédier's criticism is, is easily conceivable if one imagines that in case Bédiers *worst nightmares* (of fraud[14] and fallacy[15]) were true, the vast majority of our editions (not only) of classics in Greek and Latin alike (!) would have to be recompiled, whilst all conclusions and all history written on the bases of their editions, even all school books would be at least in subtle and certainly sometimes important details subjectable to substantial criticism. Is what we think to know about history from those texts a chimera produced by philological taste? It is needless to say that a reedition of all those classics would be a mammoth task that could take centuries. Accordingly, many philologists did not stay mute about Bédier's argument. In the following, the most famous reactions will be outlined.

1.5.3 Paul Maas

Paul Maas directly reacted as one of the first to Bédier (Maas, 1937). He confirmed that the high rates of root bifurcativity in manually constructed stemmas are valid for Classical Greek traditions as well (p.293). However, he supplied two arguments why this bifurcativity was rather expectable. The first one builds on metadata and contextual knowledge as well as on aspects of editorial practice and depicts bifurcativity as the expectable outcome of those factors. In detail, he argued, that in those medieval traditions, which were read by few, the archetype would rather seldomly have been copied three times (rather fewer times) and the survival of all three copies would be unlikely, and even the loss of all manuscripts in a subtree would not be unlikely. The implications of this type of argument are further elaborated and generalized by Trovato (2014). Maas extends the argument and states that for traditions read by many contamination would naturally be abundant and thus strict stemmatology in principle could not be applied. He admits, that for the youngest manuscripts trifurcations could have emerged and

[14] The recensor/editor chooses de gusto.

[15] The editor overseparates subconsciously or because of method-intrinsic factors.

preserved easier, but that editors could have omitted the reconstruction of unimportant hypearchetypes.

A second argument of Maas is rather numerical, it is based on the manual inference of (in his view) all possible stemmas for two and three surviving manuscripts. Due to some aspects, such as the omission of internode chains of hypothetical manuscripts, explained later, he counts 3 possible stemmas for 2 surviving manuscripts and 22 possible stemmas for 3 surviving manuscripts. The method of counting is outlined roughly in Maas (1937). A generalization, which found the number Maas (1958) give (a translation of an extended and reframed republication of Maas (1937)) for possible stemmas for 4 nodes to be wrong has been elaborated by Flight (1990). Maas states about Bédiers observation, that due to his two reasons (the numerical argument, that a trifurcation is only one of 22 cases in three manuscripts[16] and the historical argument) he believes the observation to be less offending (p. 293). In other words, instead of expecting editorial fraud and fallacy behind the large numbers of root bifurcations observed, much of bifurcativity would rather be expectable/normal and should not be viewed as surprising.

1.5.4 Sebastiano Timpanaro

Timpanaro (2004, pp.129) (first edition in 1963) reacts to Bédier and especially to Maas' arguments. Firstly, he acknowledges a large amount of root bifurcations in actual stemmas of classical texts. However, he disagrees with Maas in some profound ways. He is not the only one criticising Maas for his manner of counting (Haugen, 2015). He recounts the 22 possible stemmas Maas postulates as:

- 6 combinations of chains of all three manuscripts
- 3 combinations where two manuscripts derive from one at the root
- 3 combinations where the two other manuscripts derive from the third as

[16]Maas leaves the number of unifurcations present in his own reconstructed stemmas with at least 3 undiscussed at this point. However, Bédier had counted solely root bifurcations to arrive at 105 in 110. Maas counted all non trifurcations finding a proportion of 1 in 22, which was a multiple of Bédier's own count of 5 non root bifurcating stemmas in 110. However, this count was based on different furcation types. Had Maas counted just the number of root bifurcating stemmas of 3 surviving nodes, he had found 12. Thus, actually 12 in 22 corresponds to Bédiers 105 in 110.

root via an intermediary

- 3 combinations where an hypothetical root gives rise to one of the manuscripts and a second hypothetical node, which fathers the remaining two manuscripts
- 6 combinations, where a hypothetical root fathers two extant manuscripts and the third is a copy of any of the two
- 1 combination with a trifurcation below a hypothetical root

For an illustration, see Figure 1.3.

Timpanaro (2004, pp.138) critcises Maas for mixing different classes of stemmas here, namely, some where no loss is assumed at all, some where loss is assumed to be a quarter of the original size and some, where loss is assumed to be $\frac{2}{5}$.He states that with an imprecise amount of loss infinitely many more stemmas would be possible. Timpanaro (2004) argues further that exactly this practice would falsify the true statistical probabilities of possible trifurcative stemmas and implicitly that the real effective ratio of trifurcations would have to be higher than 1 in 22. A further point of critique is the abstract character of Maas' counting excercize. One conclusion of Timpanaro (2004, p.140) is that without considering additional empirical values, the problem cannot be usefully assessed by probability calculus. Yet another philological argument is roughly that recensionism especially in the Byzantine empire led to a text type which is preserved in younger manuscripts and which the less probably surviving older manuscripts do not exhibit. Thus, the presence or absence of these features could constitute another reason for root bifurcations in stemmas of Greek and Latin classics. Timpanaro (2004, pp.128) lists yet other scenarios, which could explain bifurcativity and root bifurcativity. After consulting Timpanaro, one can get the impression, that it is not clear why exactly the numbers of bifurcating or root bifurcating stemmas are so high, but that a large variety of possible reasons can account for various processes leading to bifurcation or root bifurcation.

Timpanaro (2004, p.157/158) himself proposes to mention instead of one single stemma in the prefaces, more than one single stemma and to be more precise about the probabilities with which a philologist assumes a certain internode or edge. At times, a philologist may also permit himself to not reconstruct certain branches or subbranches if they are simply too complicated or show too few evidence for credible hypotheses. This polimorphous approach would include the possibility to include computer generated stemmas.

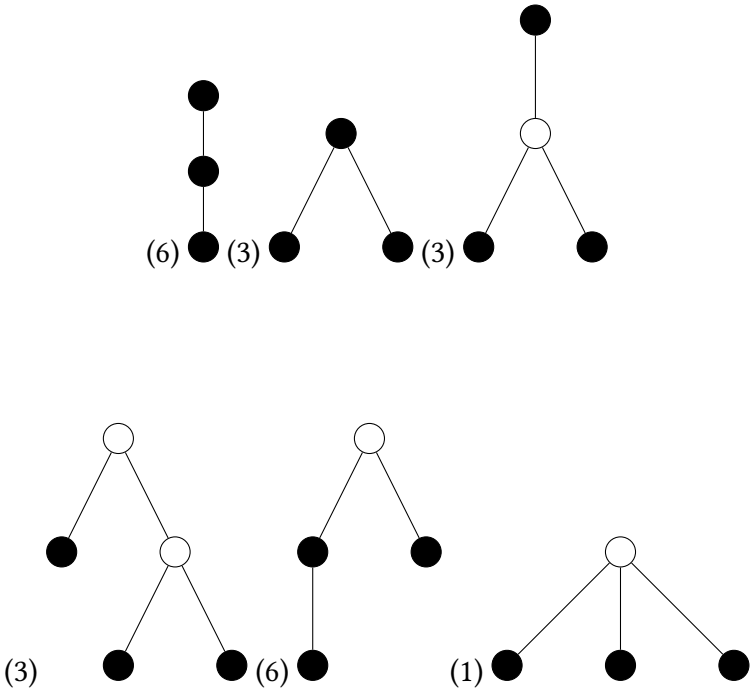


Figure 1.3: The unlabelled rooted (root topmost node) topologies of possible stemmas for three surviving manuscripts as thought of by Maas (1960), also found in Timpanaro (2004); Flight (1990). White nodes symbolize reconstructed, lost manuscripts (unlabelled), whereas black nodes are survivors (to be labelled). The number in brackets refers to the number of possible labelled trees for each topology.

1.5.5 Fourquet

Fourquet (1946) introduces a terminology for the separation of two different trees operating in the stemmatic process: the first being the arbre réel, which is the entire true (historically hidden) tree, which represents the whole genealogy of a tradition. The second relevant tree is the stemma codicum itself, which contains all survivors and roughly all manuscripts on the path from any one of those to the closest common ancestor. This subtree of the arbre réel is then further modified by the process of contraction. Maas (1960); Haugen (2015) explain how in philological practice, *codici interpositi* that is lost manuscripts with indegree one and outdegree one in direct succession should be contracted. This practice together with historical manuscript loss transforms an arbre into a stemma codicum. Fourquet (1946) gives an example taking only the upper portion of a stemma and then generalizes towards an answer to Bédier involving the aforementioned contraction process. According to him, looking at Figures 1.4, 1.5, 1.6, the case of a trifurcation in the stemma would only happen if the two surviving lines would branch off at the same point in the arbre, because of contraction of unifurcation chains of lost manuscripts otherwise being blocked at two different points of the arbre réel.[17]

Bifurcations would thus represent many more historically plausible scenarios (arbres) than trifurcations. In conclusion, Bédiers forest would be no longer strange. Certainly, the argument is interesting and could indeed make bifurcations overall quite prominent, it also entails that postulating a trifurcation in a stemma is quite a strong claim. However, this would presumably not hold for root furcations of orders larger than one, since no matter what would happen below root, if root had been copied n times with $n > 2$, then if manuscripts in more than two branches survived, a multifurcation would have to be reconstructed, no matter the practice of contracting unifurcations or concommitant probabilities. Bédiers strange forest was based more than all else on root bifurcations.

1.5.6 Castellani

Castellani (1957) in his inauguration lecture rejects Bédier's conclusions and the argument presented by Fourquet (1946) on the basis of hypothetical scenarios

[17]Note, only such nodes are to be contracted, which are both hypothetical and have none but one equally hypothetical child.

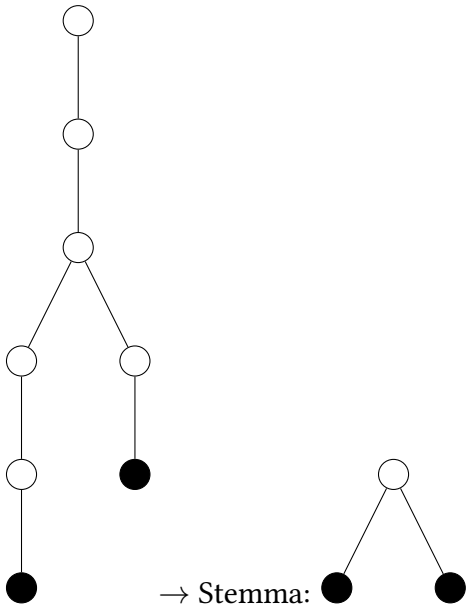


Figure 1.4: This depiction shows an arbre réel from which a root bifurcating stemma would result. A third surviving manuscript could now stem from the same branch or a branch attached to any of the lost nodes.

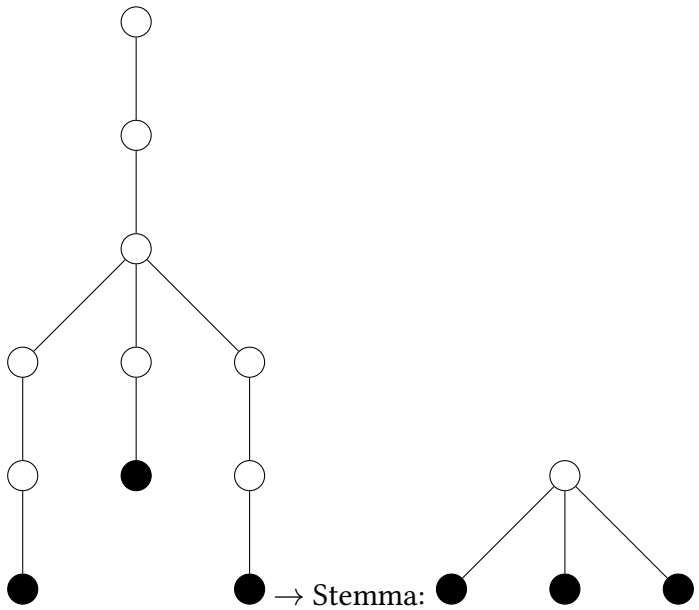


Figure 1.5: Only in this single case the resulting stemma would be trifurcating.

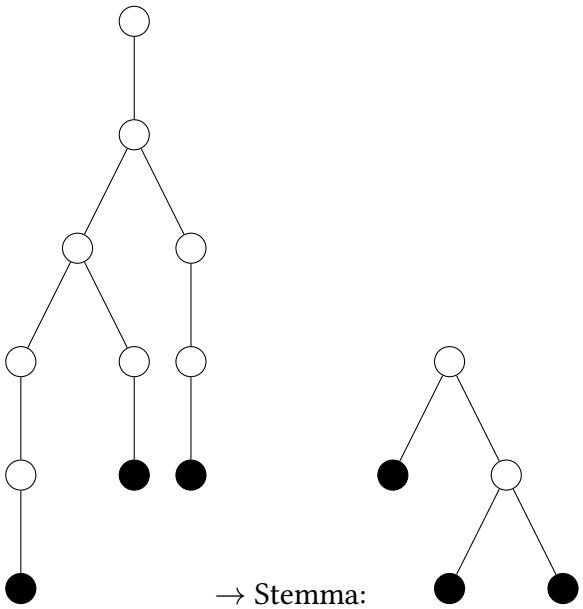


Figure 1.6: In many cases, stemmas would be root bifurcating. The original root unifurcation in the arbre however is the same in all cases.

and calculations. Bédier had stated that the scribes had wanted to be read; Castellani (1957, pp.25) reacts to this stating that instead of wanting to be read, a minstrel (and Bédier's collection had been based mostly on minstrel texts) would want to keep a new work secret and uncopied as long as possible in order to maintain his monopoly on the song. According to Castellani (1957) such a scenario would explain root bifurcativity to a certain degree.
In addition to hardly testable hypothetical scenarios, such as this one, Castellani (1957) did try to collect stemmas in order to count himself the incidence of root bifurcativity. Due under more to method and availability issues, he arrived at a slightly different collection than Bédier and his results were: 94 stemmas (8 from Bédier himself), 71 root bifurcating, 15 multifurcating and 8 uncertain.

1.5.7 Whitehead/Pickford

Whitehead and Pickford (1951) primarily react to Fourquet (1946). They outline counter examples and think, that his argument is based on a false assumption that "every stemma that is theoretically possible is equally likely to occur in practice". Finally, given three original branches[18] Whitehead and Pickford (1951) enumerate probabilities for root bifurcating, trifurcating or indeterminate stemmas. They conclude from their numbers that, given Bédiers observations would spring rather from historical circumstances, these must entail that the number of extant manuscripts be small (which for many of Bédier's stemmas should be so) and that the production rates in the middle ages must also have been low. They introduce yet another argument: a large proportion of two branch to three branch stemmas similar to Bédier's ratio would only be obtainable if the "disproportion between the families becomes very marked indeed" (p.88/89).

Whitehead and Pickford (1973, p.149) argue among other things that with an average copy number of three, in the tenth generation one would have already around $59,000$ manuscripts (3^{10}). With this number being obviously too large and with private copies becoming sterile as what regards their copy out-

[18]Many articles in the debate, for instance Bédier (1928); Maas (1960) simply reduce the question to the opposition of two and three branched stemmas, either without mentioning unifurcations and multifurcations or implicitly assuming that the statements generalize. This bares similarity to the editorial debates on variation, where instead of the entire work especially illustrative lines or word groupings or manuscript groups are the primary object of discussion from which general conclusions are or are not drawn.

put, assuming less copies would make root bifurcating stemmas very likely and unsurprising. They conclude that textual cirticism had meanwhile moved away from the extreme conservatism initiated by Bédier (p.156).

1.5.8 Guidi and Trovato

Trovato and Guidi (2004); Trovato (2014) addressed the influence of heavy manuscript loss on bifurcativity and root bifurcativity in stemmas. Trovato summarizes the debate and investigates early print editions as role models for manuscript survival. Print editions are younger and survival more probable, in fact the archetype often survived (Trovato, 2014, pp.108). Edition numbers may be present and so even if none of the actual exemplars survives, it is possible to know some missing nodes. At least a reasonable approximation of an arbre reel may be possible. Estimates of percentages of loss can be gained from numbers of surviving prints, where one knows how many books have been printed.

Trovato and Guidi (2004) examine the probabilities for the survival of k-root-furcating stemmas given different probabilities of loss (equiprobable loss for all manuscripts) systematically. They take 3 stemmas from printed editions and provide tables with the expected survival probabilities of complete loss, 1, 2 and 3 branched stemmas for different percentages of decimation. All three traditions are asymmetric in that one branch has many, the other two rather few witnesses. However in the third, due to its subbranching pattern, survival of a root bifurcation would be most probable for a high rate of loss. According to Trovato and Guidi (2004, p.45) most of the trees examined, which had not been included as illustrations were similar to this third tree.

For the other two examples, they find that unifurcations are dominant after heavy loss. Whitehead and Pickford (1951); Castellani (1957) had already found that asymmetry in the arbre would increase root bifurcation probabilities for the true stemmas. Ascertaining Greg (1931), Trovato and Guidi (2004, p.45) conclude: “As a result of the peculiarities of the genealogical trees and at high decimation level, a two-branched tradition appears to be dominant.”

1.5.9 Kleinlogel

Kleinlogel (1968) reviews discussions on the “Stemmaproblem”. He calculates the proportion of root-bifurcating to root-trifurcating stemmas for a certain stem-

Survivors	all	uni-	bi-	tri-	quad-	quint-	sext-
1	1	1					
2	3	2	1				
3	22	9	12	1			
4	262	88	151	22	1		
5	4 336	1 310	2 545	445	35	1	
6	91 984	26 016	54 466	10 425	1 025	51	1

Table 1.1: Numbers of k-furcating rooted Greg trees, as in Hering (1967, p.175) (Engel). Note, that the larger the furcation degree the quicker the increase.

matic architecture with same sized branches inspired by Castellani (1957). From the results of the calculation, he concludes that he agrees with Bédier in finding such large amounts of root bifurcations highly improbable. A further calculation is conducted oriented at human genealogies but Kleinlogel (1968, p.73) immediately dismisses it since manuscripts other than humans can remain without children for some generations and then produce offspring.

After such reasoning, he rejects all statistical approaches as yielding too few benefits. The results would only confirm the initial presuppositions while parameters could not be decucted empirically from history or used for evaluation and if they could, there would be no more need for an estimation of those parameters. Grier (1989, p.266) even exaggerates this standpoint and hopes Kleinlogel (1968) has "ended all speculation on statistical grounds by noting that even the most sophisticated mathematical model cannot account for all the variables at every level of transmission."

1.5.10 Hering

Hering (1967) tries to deduct the proportion of bifurcative to multifurcative stemmas for cases where the number of surviving manuscripts is up to 6. He enumerates all possible 1-6 furcating stemmas for up to $n = 6$ manuscripts according to Maas' method. Without looking at unifurcations (the second largest numbers), he computes the ratio of bifurcating to multifurcating stemmas and speculates that this ratio would converge to around 0.33 (p.175), see 1.1.

With this, he is presumably after Maas the first to enumerate the numbers of rooted Greg trees for small n, a name invented by Flight (1990) for this type

of stemmas proposed by Maas. Flight (1990) gives an exact recursive counting procedure also for unrooted Greg trees. Hering (1967) attributes the provision of the numbers to a friend, who was professor of Mathematics in Rostock, Prof. Dr. W. Engel without mentioning any formula or mathematical details. For n larger 6, he deems the calculation mathematically difficult (p.175).

For Latin texts he outlines another scenario how root bifurcativity could be plausible. Many ancient texts would be known to have survived the dark ages just in one copy, whereas a certain number of cases exists however where two survived which would lead to root bifurcation.

1.5.11 Weitzman

In Weitzman (1982), the author implements a computer simulation of a manuscript tradition. He uses parameters estimated from real world manuscript numbers. Kendall (1948) in another context had elaborated a formula for the estimation of final population size involving a birth rate λ, a death rate μ and a survival probability ν, which Weitzman uses as a prequisite for his simulation. He assumes

$$1 + \frac{\lambda\nu}{\lambda - \mu}[e^{1000(\lambda-\mu)} - 1] \simeq 40 \tag{1.1}$$

since typically on average 40 mansucripts survived for Classical texts. From impression, he sets probability for survival to 90% (p.58). In the final simulation, the birth rate is slightly larger than the death rate. Through the generation of two random variables, he simulates a graph. For each living manuscript at each step, the algorithm generates events. Depending on the variables, the event is death (disappearance of the manuscript) or birth (after 1500 just one random variable determines death or survival, since the advent of print is assumed to have interrupted handcopying). Likewise the time interval for the event is determined, for details see Weitzman (1982).

The simulation comprises 46 trials, whereof 15 traditions survived (not all manuscripts were lost). In history it seems plausible, that a large number of traditions was lost. One tradition had one, another only 2 survivors, in which case the stemma is easily solvable. Finally, in ten cases, the resulting stemma was root bifurcating and bifurcations had been most numerous in general.

Weitzman (1987) substantially elaborates the prior simulation and its interpretation. He bases new estimates on birth- and deathrates on historical data

and looks at Latin and Greek traditions separately. He states that "Both in Greek and Latin, real traditions vary in size between 1 and well over 100.", Weitzman (1987, p.292). Based on the elaboration of the model, Weitzman (1987) analyses the probabilities for branching in his setting. He states the conditional probability for two branchedness of an archetype to be 0.77 for Greek and 0.71 for Latin. Generally, younger archetypes had larger probabilities to become multifurcating.

Weitzman (1987, p.303) concludes that Bédier should rather not have been surprised at the large number of root bifurcations.

1.6 Post-crisis stemmatology

While the debate is still ongoing and will be investigated by means of combinatorics in chapter 2, it has divided the field into scholars continuing to use stemmatology and such which prefer not to. The history of stemmatology proper thus continues but its scope is narrowed down to a facultative tool for editors.

Stemmatological descriptions were first published in the prefaces of editions and form still part of the standard content of a preface in many modern editions. Paul Maas' book *Textkritik* issued in 1927 Maas (1927) was presumably the first purely stemmatological textbook (Halonen, 2015, p.580). It methodologically elaborated under more the concepts put forth by Lachmann and others. The question of how to edit a text, that is roughly the question of how to compile/represent one single/main text, which modern readers can all refer to in the same way; this was deeply connected with the reconstruction of a copy history of the extant manuscript's texts. The discussion on how to do this was central in the years following the onset of stemmatology which saw publications of ever new editions of the Bible and of Greek and Latin classics and the descriptions of the approaches the editors took. Important theoretical contributions came from such famous philologists as Erasmus of Rotterdam, John Scaliger, Karl Lachmann, Carl Zumpt, Paul Maas and Joseph Bédier. Some philologists aim at reconstructing an urtext, that is ideally the author's original text, de facto often the text of the latest common ancestor to all extant manuscripts. Stemmatology is a tool for the philologist in reconstructing the urtext, a tool, which can, but which doesn't have to be used. Under certain conditions, such as heavy contamination referring to the practice of multiple vorlagen for one copy, stemmatology is considered un-

applicable and consequently not practised by all, comp. Maas (1937, p.294).[19] In the strict sense, then not a tree, but a network would have to be assumed the appropriate graph theoretical model for representation, however, since most manuscripts presumably have one and only one main parent, the skeleton of such a network could still be a tree.

In philological theory and practice, no single normative approach to stemmatology has been developed, but some appreciated principles like *lectio difficilior potior*[20] or practices such as *eliminatio codicum descriptorum*[21] have become part of the de-facto-standard curriculum for editors. Other approaches have rather divided the community, an example being the question if one should take an emended[22] text as the basis for an edition or the text of an extant best manuscript. (The first approach is termed *neo - Lachmannian*, the second *Bédierist best text* approach by Robinson (2000) and others.) It is under more by these disagreements, that different types of editions persist (critical, diplomatic, etc.).

Unfortunately, most principles have not been quantified, so the editor's arguments remain hypothetical and controversial in nature.[23] Within a wide range of possible approaches, each scholar compiling an edition can locate him or herself wherever he/she wants and argue why for his tradition his particular, say neo-Lachmannian approach compiling a diplomatic edition, is the best approach and why a new edition is necessary.

[19]Maas (1937, p.294) postulates that there is nothing one can do against contamination, original wording: "Gegen die Kontamination ist kein Kraut gewachsen.", a very prominent quote among editors. There is however no consensus about this; Flight (1994) highlights the possible usefulness of stemmatology even in the face of heavy contamination.

[20]The more obscure meaning has a higher chance for being the original.

[21]This denotes the practise of pruning manuscripts considered of no valuable genealogical information, which often applies to very young manuscripts, copied from other extant relatively young manuscripts. Eliminatio codicum descriptorum was and is exercised by many editors, it can considerably reduce work load.

[22]Emendation denotes the process of introducing the variant, deemed more likely ancestral/authorial by the editor at any position of the manuscript. Emendation can entail inserting a reconstructed non attested variant.

[23]Andrews (2014, p.538) warns, especially in case of computational application of such principles, that one must be "extremely careful before adopting any sort of rule-based guideline for the classification of variants [...] It is far too easy to be led blindly into poor results.".

scribe\vorlage	**Var 1 (HF)**	**Var 2(LF)**
Var 1(HF)	*correct*	Error: **2**→**1**
Var 2(LF)	Error:**1**→**2**	*correct*

Table 1.2: Columns point out which variant the vorlage really has, rows the one the scribe decides to write.

1.6.1 Excursus: Lectio Difficilior

To look more closely at the nature of one of the philological principles, here, a short excursus on lectio difficilior is presented. This will give an impression on the inner workings and complexity of the still debated manual methods of classical stemmatology before we will proceed into the digital. A famous philological principle is called *lectio difficilior potior* (LD).[24] It postulates that scribes had a tendency to replace obscure typically low frequency (LF) items by more comprehensible typically high frequency (HF) ones leading to a higher a posteriori probability of LF items to be authorial wordings. There is much controvery about the principle and it is widely agreed upon that not each instance of cases where an LF and an HF variant are aligned can be explained through it. At the same time however it is a famous and widely applied philological principle. For a more in depth review of the principle and its various applications in textual criticism see Tov (1982). A possible explanation by Reynolds and Wilson (2013, p.222) is: "careless copying or a desire to simplify a difficult passage" by the scribes "sometimes consciously, sometimes inadvertently". LD implicitly postulates (rephrasing the principle in terms of game theory) an optimal scibal strategy in always writing the more frequent variant comitting only errors of the type *LF* → *HF*. Errors *HF* → *LF* a posteriori appear to be dispreferred. For illustration, see Table 1.2.

It is assumed that for a sloppily written/hardly recognizeable word, two possible readings surface in the scribes mind: variant 1 and variant 2. The first surfaces because it is the by frequency and context most probable reading or first guess, just as in a *cloze task*, Taylor (1953),[25] the second emerges after closer in-

[24] Also: *difficilior lectio potior* or simply *lectio difficilior*. It's earliest mention in modern philology is given as 1696 by Jean Le Clerc's *Ars critica* in Trovato (2014, p.117).

[25] Cloze task refers to a paradigm in psychological experiments, where the subject is required to fill in a gap in an otherwise explicit sentence. One observation is, that the closer the gap is

spection. Now if both are incongruent, where the more frequent one is always termed variant 1, the scribe has to decide how to proceed. Lectio difficilior expects a non uniform probability distribution underlying the scribal choice. The above table illustrates that the optimal strategy of the scribe to maximize his outcome is to always write the more frequent variant 1. In that case, he will only commit errors of the type 2→1 or less frequent → more frequent. The motivation for him to do so has been explained as tendency to clarify the text (Brotzman, 1994, p.128). West (1973, p.26) notes: "Obscure words and proper names frequently baffle the scribes". Both types of words refered to are typically low frequency.

Philologists have found *lectio difficilior* as an explanation following from their empirical data hundreds of years later. [26] Interestingly, Reynolds and Wilson (2013, p.222) have been pointing out that at least some of the variant substitutions go back to unconscious miscopying of a less frequent variant, resulting in a more frequent counterpart.

LD has a striking parallel in reading research where HF tokens are a priori processed quicker than LF items (for a discussion see Rayner et al. (2012, p.54/55) on lexical decision times). Therefore, it could be that the rather elaborated error dispreference interpretation LD is an *a posteriori* explanation for the observed data but does not explain unconscious cases. For text in holes and gaps in manuscripts, the scribe would have had to take a conscious decision, but for many of the observed errors where a more frequent variant substitutes a less frequent one a simple subconscious misreading is a more economic explanation for the substitutions. This presupposes that misreading is heavily correlated with frequency, compare also Rayner et al. (2012, p.54/55).[27] However, the uni-

located to the end, the more restricted is the choice of possible words to fill in. At the same time the item is easier to fill in. Generally some syntactic positions are more restricted than others.

[26] The opposite case of high frequency items replaced by low frequency ones is not excluded thereby and indeed, lectio difficilior is understood as a rule of thumb rather than as an infallible principle in philology.

[27] The dispreference for one error type must be well motivated. In case a punishment (by divine or worldly instances) for introducing a weirder variant into the manuscript is the reason for the scribes dispreference, one could expect the scribes to deliberately replace low frequency items, even if there is no visual similarity or misreading behind, simply because they would minimize criticism. This impulse could of course be blocked consciously because of a good intention to preserve witness identity and hence become selective, that is it would only apply in subconscious decisions or below a certain frequency threshold. At this point however, the mechanism becomes

versality of the reading process as a human ability with its neurological set-up should not be understood as a concurrent model of explanation for manuscript genealogy interpretation but as an additional source of knowledge for inherently ambiguous historical data. However, the lack of the term "misreading" or of any synonym in the subject index/glossary of the summarizing works "Psychology of Reading", Rayner et al. (2012) and "The Science of Reading - a Handbook", Snowling and Hulme (2014) may be seen as evidence for few research activity being devoted to the topic.

1.7 Stemmatology in the Digital Age

In 1973, West (1973, p.72) had stated about the use of the computer, that it might be a tool that could rather help with the 'unsubtle' tasks. In 1984, Griffith (1984, p.83) in a similar vein states, that "the computer can, fortunately, never do the scholar's job for him: it can only, *dis faventibus*, help to illuminate his path.". Although, given the nature of historical loss entailing inherent unresolvable ambiguity, the aforementioned view remains popular and has been reexpressed partly or entirely in many different and similar wordings, see for instance Bordalejo (2015). However, reliance on the results of the computer rather than on philological intuition has been proposed as a serious alternative lately (Andrews, 2014).

It may be impossible to give one exact date of the beginning of computational stemmatology. A number of works (Robinson and O'Hara, 1996; Howe and Windram, 2011; Robinson, 2015) mention Lee (1989), as the first to apply modern phylogenetic programs to reconstruct manuscript genealogies in 1989. However, the computer had been used earlier in connection with stemmatology or related tasks, comp. for instance Poole (1974), who wrote an Algol60[28]program aiding him in and performing stemmatic analyses. Yet 17 years earlier, Ellison (1957) wrote a doctoral thesis entitled "The use of electronic computers in the study of the Greek New Testament text" at Harvard University, where he developed a program for the automatic comparison of manuscripts determining their distances and establishing groups which he compared with previous scholarly classification. Froger (1968); Zarri (1976) developed stemma-producing programs and algorithms and Hockey (1980) additionally points to two stemmatologically

quite speculative making many empirically untested assumptions.

[28]Algol60 is a programming language.

relevant publications by Dearing (1970) and Shaw (1974), which appeared in the Bulletin of the Association for Literary and Linguistic Computing (ALLC).[29] Together with Griffith (1968) and Platnick and Cameron (1977), the aforementioned near-contemporaries could all be considered as among the earliest possible onsets of the discipline. While thus, the computer was used for stemmatological calculus at least since the 1950ies/1960ies later including stemma producing programs, the widespread use of phylogenetic programs (PHYLIP, PAUP, SplitsTree)[30] for stemma generation presumably started with Lee (1989) and eventually different, at times more rule-based approaches imitating or incorporating philological variant choice based procedures became less numerous, although they continued to exist, see for instance Gjessing and Pierce (1994). Depending on the concrete starting point one assumes, the presumable age of computational stemmatology somewhere between 50 to 60 years, which is only slightly younger than commercial availability of computers.

Reconstructing the genealogical relationships of manuscript texts is very similar to the reconstruction of phylogenetic trees, compare Platnick and Cameron (1977). In fact, stemmatology as well as phylogeny are sciences working with *trees of history* as O'Hara (1996) calls them, which include historical linguistics (linguistics).[31]

While the import of phylogenetic techniques is multi-faceted, see for instance Robinson and O'Hara (1996), it firstly took place on a theoretical level. In 1968, Griffith (1968) applied principles of numerical taxonomy to several manuscripts (Howe and Windram, 2011). Platnick and Cameron (1977) elaborate on the similarities between cladistics, text and language evolution. From the 1990ies onwards, phylogenetic software was widely applied, see Table 1.3. From the beginning, the field had an interdisciplinary character and an important impact. As van Reenen et al. (1996, p.IX) puts it

> The advent of the computer is not only seen as a 'handy tool', [...] rather the implementation of the computer has fundamental theoretical implications. [...] the entire stemmatological process has to

[29] This association was founded in 1973, according to `eadh.org`, and is today known as European Association for Digital Humanities (EADH).

[30] PHYLIP Plotree and Plotgram (1989), PAUP Swofford (1990), SplitsTree Huson (1998).

[31] O'Hara (1996) points to some early non systematic contacts of those disciplines. The most famous example of a *tree of history* is presumably Charles Darwins only figure in *On the origin of species*, a species family tree, Darwin (1859).

> be redesigned, and philologists have to learn to relativize their own decisions.

More recently, Bod (2013, p.349) sees stemmatology as a "normal science" among humanities disciplines, but Andrews (2014, p.538) does not yet see stemmatology at this point.

Many algorithms and a plethora of software have been developed for generating biological trees, where additional techniques such as *bootstrapping*,[32] *consensus trees*,[33] *jackknifing* (Quenouille, 1949; Tukey, 1957) or *shotgun sequencing* (Staden, 1979; Anderson, 1981),[34] were designed to cope with specific problems, which are not all transferable in a straightforward manner into philology. Additionally, financial support for bio-informatic software development is much larger than for philology. Consequently, not all methods used in biology have been transferred yet to philology. Table 1.3 gives an overview over publications, which used phylogenetic methods and algorithms.

The table has been compiled under more but not at all solely on the basis of some recent recapitulations of stemmatology's history: Howe and Windram (2011); O'Hara (1996); Andrews (2014); Bordalejo (2015); Macé et al. (2004) and the works van Reenen et al. (1996, 2004); Andrews and Macé (2014). The publications are such where the authors apply computational methods, especially phylogenetic software and present (software-generated comprehensive) visualisations of their traditions. More works, mentioned thereafter are directly concerned with aspects of stemmatology and apply the computer in that process. Together, these publications cover to the authors best knowledge at least a major part of the qualitative and quantitative publications on applied computational stemmatology to the date of the compilation of this thesis in the Western Sphere.[35]

[32] In bio-informatics, this refers to a statistical sampling technique, which is used to estimate the reliability of branches of a genealogical tree, see Efron (1979).

[33] For instance, in cases, where more than one tree is seen as equally informative, a consensus tree may be generated based on the commonalities of the former trees. This can lead to politomies/multifurcations, where the original trees only had bifurcations.

[34] This technique could be used in aligning fragments or assembling a whole text, where only fragments survive.

[35] This implies primarily the English literature (29 of the publications in the table) and extends partly to French (6), German, Italian (1), Dutch and other European languages, where authors do often choose to publish in English, but where especially in philology large bodies of often untranslated articles considered important in the field prevail. It covers furthermore stemmatological approaches for a wide range of subject languages, for instance Sanskrit, see Phillips-Rodriguez

Publication	Methods/Algorithms	Software	Kinds of trees	Units
Lee (1989)	Parsimony	MacClade, PHYLIP	rooted bifurcating	
Robinson and O'Hara (1992)	Parsimony	PAUP	cladograms	variant readings
Robinson and O'Hara (1996)	Parsimony	PAUP		variant readings
Salemans (1996)	philological preparation, Parsimony	PAUP	cladogram	
Robinson (1996)	Parsimony	PAUP	cladograms	variant readings
Barbrook et al. (1998)	Split decomposition	SplitsTree	unrooted bifurcating	variant readings
Salemans (2000)	philological preparation, Parsimony	PAUP	cladograms	
Spencer and Howe (2001)	Neighbour Joining	PAUP	unrooted bifurcating	variant readings
Spencer et al. (2002)	CBGM, Parsimony, Consensus Tree	Matlab,PAUP	circular, cladogram	variant readings
Macé et al. (2003)	multi dimensional scaling, visual observation of clusters	PHYLIP	unrooted bifurcating	variant readings
Mooney et al. (2003)	Parsimony, split decomposition	PAUP, SplitsTree	unrooted bifurcating	variant readings
Woerther and Khonsari (2003)	Parsimony on triples, Bootstrapping	PHYLIP	cladogram	variant readings
Stolz (2003)		SplitsTree	unrooted bifurcating	
Spencer et al. (2003a)	Neighbour Joining, Bootstrapping, Consensus Tree	PAUP	unrooted bifurcating, multifuracting	variant readings
Spencer et al. (2003b)	Item order distances (IEBP), Parsimony?	PAUP	unrooted bifurcating	level:sequences
Macé et al. (2004)	Parsimony, Neighbour Joining	PAUP	unrooted bifurcating multif clado consensus	variant readings
Lantin et al. (2004)	Parsimony	PAUP	cladogram, consensus	readings
Spencer et al. (2004c)	red. median	SpectroNet	unrooted bifurcating	?
Spencer et al. (2004a)	Parsimony, Neighbour Joining, Bootstrapping, Split Decomposition	PAUP, SplitsTree	unrooted bifurcating, Neighbour Net	readings
Spencer et al. (2004b)	Neighbour Joining, Consensus Tree	PAUP	unrooted	readings
Yorav et al. (2005)	Parsimony	PAUP	unrooted bifurcating	
Eagleton and Spencer (2006)	Split Decomposition	SplitsTree	unrooted bifurcating	variant readings
Pouliquen (2007)	lattices	ConExp	multifurcating cladogram	variant readings
Windram et al. (2008)	Parsimony,Split Decomposition	PAUP,SplitsTree	unrooted bifurcating	variant readings
Roos and Heikkilä (2009)	many	many	unrooted bifurcating	variant readings
Phillips-Rodriguez et al. (2009)	Neighbour Joining, Bootstrapping	PAUP	cladogram, unrooted bifurcating	
Roelli and Bachmann (2010)	Leitfehler	PHYLIP	unrooted bifurcating	variant readings
Le Pouliquen (2010)	Neighbour Joining, Levenshtein distances, Compression Distances		cladogram	
Le Pouliquen and Csernel (2010)			cladogram	
Roos and Zou (2011a)	Expectation Maximisation		graph	
Heikkilä (2014)	RHM, Semstem, Parsimony, Neighbour Joining, Leitfehler, Split Decomposition	PAUP, PHYLIP, SplitsTree	unrooted bifurcating	
Roelli (2014b)	Leitfehler		unrooted bifurcating	
Roelli (2014a)	Philological Calculus		unrooted bifurcating	
Barabucci et al. (2014)	Philological Calculus		rooted multifurcating	variants, subword units, multi word units
Halonen (2015)	Parsimony, Split Decomposition, RHM	PAUP, SplitsTree	unrooted multifurcating	
Robinson (2015)	Parsimony, RHM	PAUP	rooted multifurcating cladogram	words

Table 1.3: Table of publications applying phylogenetic (or related) software or phylogenetic visualisations.

The publications in Table 1.3 have been reviewed to characterize (phylogenetic) stemmatology briefly in words with respect to factors such as methodology and interdisciplinarity. The programs PAUP, PHYLIP and SplitsTree account for the vast majority of software applications.[36] Algorithmically, heuristics to obtain a parsimonious tree (maximum parsimony), Split Decomposition, Bandelt and Dress (1992) as well as Neighbour Joining, Saitou and Nei (1987) represent the most applied theoretical foundations of applied computational stemmatology.

In Table 1.3, 42 different persons compiled 36 publications having in all 89 authors and co-authors, here are 2.5 authors per publication ranging from 1 to 8. Each person occupies 2.1 positions as author or co-author and published 0.86 papers. Some authors are very active in the field, for instance C. Howe (11 publ. from 1998 to 2010), P. Robinson (9 publications from 1993 to 2015) and M. Spencer (9 publications from 2001 to 2006).[37] Other authors frequently publish on stemmatology likewise, meaning that the community as constitutive of the located publications appears to have a core of frequently publishing experts from philology, computer science and bio-informatics and some occassional contributors or co-authors, which can be philological experts on certain traditions or computational experts on certain methods. From 36 authors of publications in the table, information on academic career or affiliation was available on the internet: 15 were rather philologically oriented, 11 biologists or bio-informaticians, 7 computer scientists and one mathematician. This exemplifies the interdisciplinary character of the field, which is further visible through the publication types and platforms. Many of the articles are longer journal articles or articles in proceedings of a conference or in the summarizing works van Reenen et al. (1996) and van Reenen et al. (2004) and monographs. The publication media stretch from biological journals in theoretical and computational biology to computational conferences and philological editions with a large portion being located within

et al. (2009) or Hebrew, see Yorav et al. (2005). Ancient texts in Chinese and other semasiographic writing systems additionally bare the complexity of telling apart variation, which is due to the development of the writing system and variation arisen from other aspects of the transmission. Although philologically constructed pedigrees are present, comp. Simson (2006, p.153, p.221), the author is not aware of an application of computational methods to compile stemmas in such cases.

[36] All three programs exist in a variety of versions.

[37] There are additional publications of all three authors concerning stemmatology in the wider sense, which are not included in the table.

m_1	m_2	m_3	m_4	m_5	m_x	γ	β	α	**Variants**
that	this	ðis	ðis	t'	this	ðis	this	this	A-D
iz	is	is	is	is	'	is	is	is	A-C
one		a	a	an	an	an	one	an	A-D
text	text	text	text	text	tekst	text	text	text	A,B
DCBA	AADA	BACA	BACA	CAAA	ABAB	BAAA	AABA	AAAA	**PseudoDNA**

Table 1.4: An example of how pseudo-DNA can be generated. This approach is practised but ignores word similarities.

digital humanities.

Bootstrapping and compiling consensus trees (from either the bootstrapped trees or from equally parsimonious trees) is relatively common. The level on which computation is based (preprocessing to input of pseudo DNA, for an exemplifications, see Table 1.4, or valid input to bio- informatic programs) is overwhelmingly referred to as readings or variants, but it is not always exactly specified what that concretely entails, especially if one "reading" can comprise more than one word. Moreover, oftentimes preprocessing and normalization are conducted. [38] Experiments with weighting have been conducted (see for instance Spencer et al. (2004b)) but could be described as rather marginal, because as the authors argue, the results are not very different from their non-weighted counterparts (p.238). This is true for groups of leaf nodes but it is not prima facie clear if the statement extends to internode branching.

Unrooted bifurcating trees (cladograms) are the major visual representations used in the publications using computational methods. Classical stemmas in the publications do usually not adhere to the same represenational formats. Exceptions such as Roos and Zou (2011a) do not usually use bio-informatic software for visualisation. An extension to this visualization landscape, which can be easily combined with the output of bio-informatic software will be presented in the next chapter.

Apart from the direct application of phylogenetic software producing visualizations, others have used the computer in connection with stemmatology or laid out theoretical foundations or computed stemmas in different ways (for instance Haigh (1971); Dearing (1970); Weitzman (1987); Roos et al. (2006); Andrews and Macé (2013); Andrews (2014); Andrews et al. (2012); Christopher J. Howe and Windram (2012); van Reenen et al. (1996, 2004); Andrews and Macé (2014); O'Hara (2006); Mink (2004); Wachtel (2004); Merivuori and Roos (2009); Najock

[38]Such a step entails philological judgement on insignificant variation and can thus strictly speaking be seen as external intervention on which the results depend.

and Heyde (1982); Gjessing and Pierce (1994)). Mink (2004); Wachtel (2004) work on the huge tradition of the Bible and consequently target problems of large numbers of manuscripts. Among the earliest tasks, solved by aid of the computer was the alignment or collation of different manuscript texts, which has a very similar and parallel history to that of the production of the final stemma and considering multiple sequence alignments is even undisentangibly intertwined with stemma production, see for instance Robinson (1994); Dekker and Middell (2011). Partly connected with this were approaches on the evaluation of scribal errors (Spencer and Howe, 2002, 2001; Spencer et al., 2006) where parallels to research on mutations in genomes are exploited, and Andrews and Macé (2013) which also provided an annotation platform and made a comparative study on the effect of the elicitation of significant errors. The website `stemmaweb.org` is a related website, which offers the traversal of variant graphs, see Figure 1.7, for some real and artificial traditions as well as the possibility to upload traditions and generate collations, variant graphs and stemmas using a choice of 3 algorithms, Semstem (Roos and Zou, 2011a), RHM (Roos et al., 2006) and NeighbourNet, based on split decomposition. Additionally, a fine grained analysis of variants of the artificial traditions is offered, which allows the user to examine which variants go with and which against the stemma, the particularities described in Andrews (2014).

Another task, for which the computer has been applied successfully was the detection of exemplar shift (den Hollander, 2004). Yet another task was the rooting of an unrooted tree, for which Haigh (1971) provides a method. That part of the literature, concerned with methodological and theoretical questions is being summarized along with the debate on Bédier. The computer stemmatological methodologies have also been applied to written evidence, which is non-textual, for instance in musicology (Windram et al., 2014). Likewise printed traditions have been examined, (Bergel et al., 2015).

Finally, many publications on (digital) editing and editions include (computational) stemmatologically relevant discussions, see for instance Robinson (2000); Greetham (2010); Bordalejo (2015). According to Elena Spadini,[39], computational stemmatological analyses are already used in recent electronic editions, such as an edition of Dante's divina commedia.[40] This edition, according to Spadini uses

[39] `http://ride.i-d-e.de/issues/issue-3/commedia/`, last accessed on 26.07.2016.

[40] `http://sd-editions.com/AnaAdditional/commediaonline/home.`

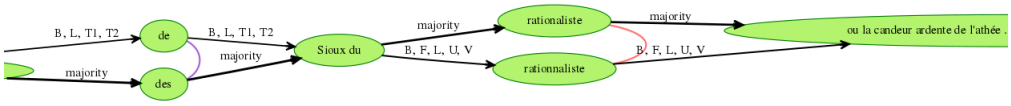


Figure 1.7: A variant graph as described in Andrews (2014).

PAUP and generates a stemma or tree (unrooted and bifurcating), for instance per line.

With an approximated age of around 50 years, it is hard to identify definite trends, but after their introduction by Lee (1989) phylogenetic methods have come to dominate the field and a slow shift from the simple import and use of external/unadapted (mostly phylogenetic) software and an interpretational adaptation of the results towards the technical adaptation of algorithms and a more manuscript-centered approach with or without various combinations of philological and computational calculus, comp. for instance Roos et al. (2006); Roos and Zou (2011a); Roelli and Bachmann (2010); Andrews and Macé (2013) may be considered as such a trend. The next section will describe another important trend or aspect resulting from the import of phylogenetic software.

1.8 Characteristics of the Mainstream Visualisation

Since visualization is central to stemmatology, the history of visualizations used in stemmatology to display manuscript genealogy will be briefly outlined and summarized with a look to the characteristics of these visualizations. Philological stemmas are similar in their form to genealogical trees of aristocratic families,[41] representing themselves probably the first visualisations of genealogical relations, (Timpanaro, 2004; Lima, 2014). Here, the nodes do not have a border but only free floating text and the edges are always angular, see Figure 1.8. Other

`html`, last accessed on 26.07.2016.

[41]The metaphorical real-world *tree* (plant) has seemingly not been used in stemmatology, but only in historical linguistics and phylogeny.

stemmas feature encircled nodes and mostly abandon rectangular shapes. There is no explicit standard for the visual representation of stemmas and some more creative approaches use colors underlying groups of manuscripts and time lines, see Figure 1.9. A tree with straight edges, where unbordered nodes carry the name of the extant manuscript, Greek letters for lost manuscripts at internode positions,[42] and α or ω for the archetypes, a dotted line for marginal contamination or corrections and a continuous line for genuine contamination, allowing multifurcations, multiple roots unusual, can be described as the dominant framework of contemporary philological stemmatic representation, see Figure 1.10.

This format is not the same for the output of computational applications inheriting from phylogenetics, which mostly adopt the manifold phylogenetic representations such as seen in Figure 1.11. Figure 1.12 shows another typical graphical output of bio-informatic software, which is not adapted to stemmatology in representing multifurcations. Almost all publications in Table 1.3 use either one or both of the typical phylogenetic visualisations, the plain output of the programs used. A special case is a Neighbour Net, see Figure 1.13. This visualisation displays distances between subgroups and single manuscripts by mapping them onto rectangular structures in the interior of the net and is therefore more information-rich than the others, but at the same time it must be learned, how to read it.

Apart from stemmatic visualisations, philological literature has used other schematisations, such as depictions of the binding of pages (Kelemen, 2009). Variant graphs (Schmidt and Colomb, 2009) are an example of a new interactive philological visualisation developed in the digital medium applied to stemmatology (Robinson, 1994; Andrews, 2014).Variant graphs are more than just a visualisation and may function also as annotation tool. Since stemmas can be obtained by clustering and based on distance matrices, theoretically, all possible cluster visualisation techniques are available, but have been scarcely used in the traditional or newer literature as a means of representing manuscript distance matrices. Griffith (1984) is an exception, he used a technique with an output similar to that of multidimensional scaling and principle component analyses in 1984 and plotted manuscript's vectorial localisations into a 2 dimensional coordinate system. Given large amounts of contamination corrupting genealogical affiliation same scale bivariate cluster plots may be an option to meaningfully vi-

[42]Lost manuscripts at leaf positions are rarely visualized.

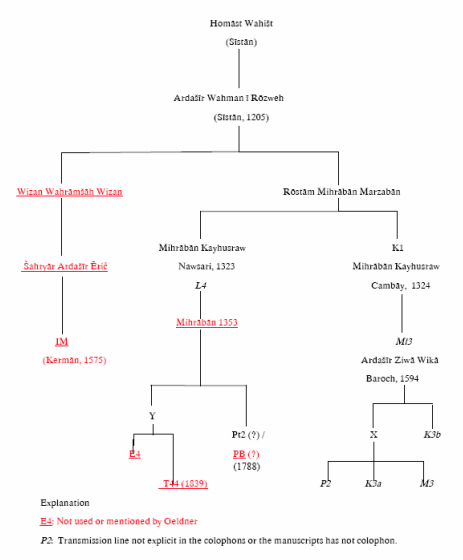


Figure 1.8: A stemma of some Avestan manuscripts resembling classical family genealogies with scribes names and additional information added, from $http://ada.usal.es/videvdad/manuscripts.htm$, last access on September 18^{th} 2015.

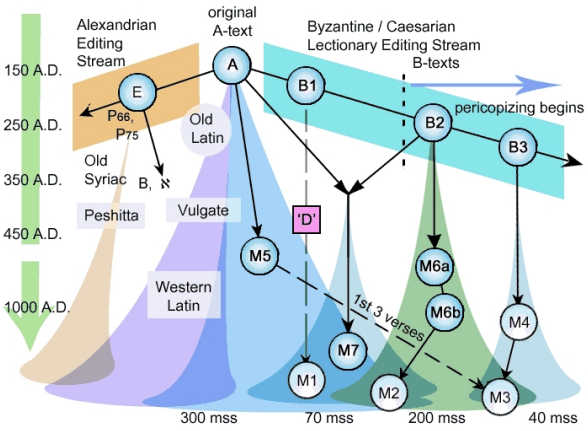


Figure 1.9: A stemma using additional devices from $nttextualcriticism.blogspot.de/2010/12/variations-in-genealogical-stemma.html$, last access on September 18^{th} 2015.

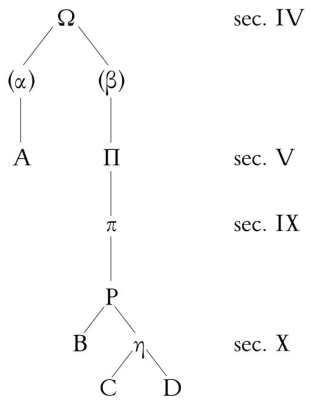


Figure 1.10: A typical stemma, used for exemplification on Wikipedia, $it.wikipedia.org/wiki/Stemma\_codicum$, last access on September 18^{th} 2015.

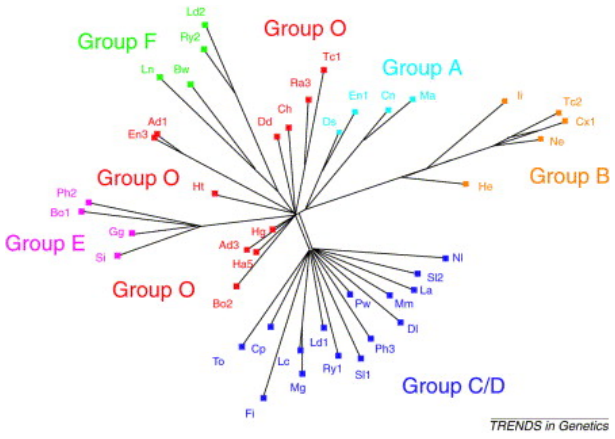


Figure 1.11: A stemma generated automatically by phylogenetic software, from Howe et al. (2001).

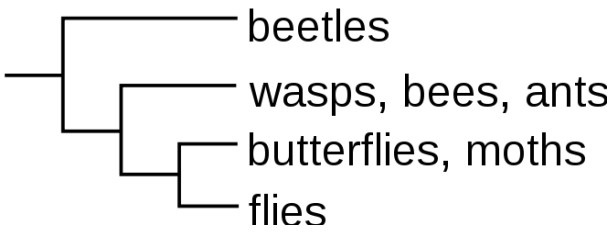


Figure 1.12: A cladogram, typically entirely bifurcating from English Wikipedia "Cladogram", last accessed on 26.07.2016.

sualize the manuscript space without running the danger of assuming too many doubtful genealogical relationships. It is thus not true, that large amounts of contamination would imply that the use of the computer is of little additional value to the field. Another technique applied in phylogeny is the comparison of trees for instance through overlays (Munzner et al., 2003; Hillis et al., 2005). Especially in the comparison of philologically obtained manual stemmas and computer generated ones, this could enrich the visualisation landscape.

In summary, concerning visualisations, stemmatology still has many visualisation techniques apparently unexplored. Representing the output of bioinformatic software in a more traditional way has not yet been an initiative of computational stemmatology. It may involve additional manual or computational editing, such as rooting a tree, but generally it should increase intuitive readability of a visual representation of a stemma, since it is congruent not with the biological needs and premisses (such as evolutionary splits into exactly two) but with the traditional stemmatological ones, which have also brought forth exactly that kind of graphical representation. Moreover, in the digital age, for the first time, dynamic stemmas, which display the evolution of a stemma in a similar way as graph evolution for instance in social network simulations can be produced easily.

1.9 Towards a broader visualization landscape

Visual representations of stemmas had changed with the introduction of bioinformatic software. From the depictions tailored for the focusses of biologists (which are leafs and not internodes (Cameron, 1987), and bifurcations, compare Hoelzer and Meinick (1994) for the dominance of bifurcativity) being output by

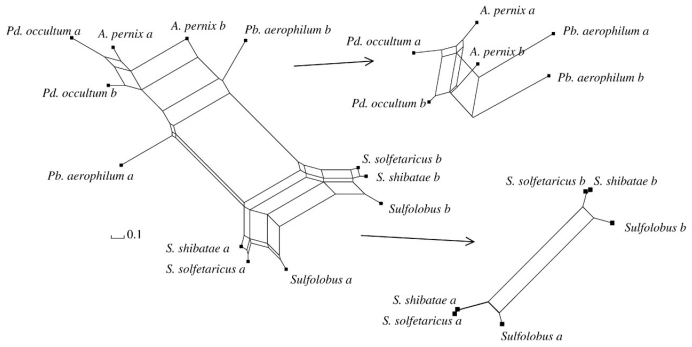


Figure 1.13: Neighbour nets from Bryant and Moulton (2004), where the rectangles display information on distance of subgroupings from each other, see also Bandelt and Dress (1992).

bio-informatic software, unrooted bifurcating trees focussing on the leafs (see Figure 1.6, page 23) have been extensively used in stemmatology (comp. Table 1.1 on page 16).

On the other hand, philologists had produced stemmas, which were rooted and would allow a clear view of the internodes, see for instance the philological stemmas in Roelli and Bachmann (2010); West (1973). "Zoologists are most interested in the end-points of their trees, that is, the individual taxa. [...] The textual critic is not really interested in the endpoints of the tree, that is, the specific manuscripts.", Cameron (1987, p.239). Don Cameron elaborates on that concluding that for the zoologist the endpoints represent positive values, survival and adaptation, whereas for the stemmatologist, they represent negative properties, error and decay. Although the new biological depictions were thus unusual and in some way counterintuitive for stemmatology, stemmatologists accepted them as new viewpoints and customized them coloring groups (Barbrook et al. (1998)), but also discussing the implications for instance of unrootedness for interpretation (see for instance Bordalejo (2015)). In terms of graph theory, the issue is that trees used in biology mostly do have labels only for leafs, whereas in traditional stemmas all nodes are labelled. This makes stemmatology most similar not to zoology in general, but to paleontology, which shares many more prequisites with stemmatology than other biological subdisciplines, such as dealing with historical entities.[43] Most importantly, in paleontology, internodes are often labelled,

[43]Working with historical materials, paleontology shares many traits and methodological approaches with traditional philology, such as defining principles for the formulation of the build-

see Harper Jr (1976). Consequently looking to paleontology, new concepts, for instance in but not limited to, visualization can be looked for.

Roos and Zou (2011a) do not use bio-informatic programs for the computation of the stemmatic hypothesis, but they are (among) the first who produce and include in their publication multifurcating depictions presumably resulting directly from computer software. Thereby, they not only explore the algorithmic space for more manuscriptology adapted solutions, they also create visualizations, which are more adapted to the needs of stemmatology. However, most publications in computational stemmatology use bio-informatic programs and their visual output.

Although most bio-informatic software outputs entirely bifurcating trees, consensus trees (for instance either of bootstrapped or equally parsimonious trees) do have (typically relatively few) politomies/mutlifurcations. Thus the depiction of the stemmatic structure in a more traditional philological way, directly from the output of bio-informatic software depends only on a technological link between the bio-informatic output and the final depiction. If internodes and their structure can be displayed in a more traditional way, this could enable observations and motivate discussions of the internal stemmatic structures produced by different algorithms and their congruence with traditional research. Therefore, a stemma generator was programmed, converting a bio- informatically encoded tree (Newick format, `http://evolution.genetics.washington.edu/phylip/newicktree.html`:26.07.2016)[44] into a pdf Figure (it assumes a rooted tree).

1.10 Dynamicity, Slide shows

In the digital age, visualisations become feasible, which were hardly possible in print. Especially, dynamic depictions such as graph evolutionary ones combining a graph's configuration at different timesteps into a film have become possible. By adding the "fourth dimension" - time,[45] more information can be conveyed. Especially, manuscript loss or stemmatic cycles through correction

up of a phylogeny, see Harper Jr (1976).

[44]The converter uses the Newick format with all node names labelled, as for instance in "(A,B,(C,D)E)F;".

[45]Actually, in many cases it might be only the second or third dimension of an actual representation.

(Andrews and Macé, 2013) can be displayed in a dynamic representation without the need for an additional visual parameter (such as crossing out for lost manuscripts). Dynamicity reduces thus the danger of visual overload, a term mentioned for instance in Mazza (2009). The stemma of Schlyter, Figure 1.1 displays the dimension time however as an underlying grid. Here, the temporal succession of the copy processes coincides with the depth of the node in the tree as measured by the underlying grid. This grid in the author's view is a more attractive depiction of the temporal dimension, since it does not "slip away".

However, a dynamic version does have particular advantages. Especially the potential to reduce visual overcrowding may make it useful for very large traditions. The produced converter explicitly offers the possibility to generate a dynamic stemma.

1.11 Additional Visualizations

Additionally, it is possible to use different tree visualizations. Lima (2014) lists many different possible ways to visualize a tree (for instance Icicle trees). A hypothetical example shall illustrate these possibilities. If a researcher would want to emphasize manuscript provenance and effect, he/she could use a circular treemap and underly a map leading to a new kind of stemmatic visualization, see Figure 1.14. The interface of maps and stemmas could be a promising area in stemmatological visualization. In the example, the larger circles represent older manuscripts, the exact location of which is marked with a cross in the same color. The extension of the circle (which could be adapted to other more convenient and manuscript-individual shapes) may approximate the reach of the influence of the manuscript's textual version and different colors could be used for different manuscript or text groups. For a depiction of this kind, the dynamicity induced by a slideshow is probably one of few ways to bring in the dimension time. Another possible rendering of a stemma could be a cubic, 3-dimensional rendering, opening the possibility for the cube's facets to be connected to different kinds of additional information, such as geographical, temporal or contextual information. Many more factors intermingled with stemmatology, such as manuscript material, writing system, etc. can be mapped to visual parameters in order to produce stemmas, which for each tradition intertwine and highlight the interplay of

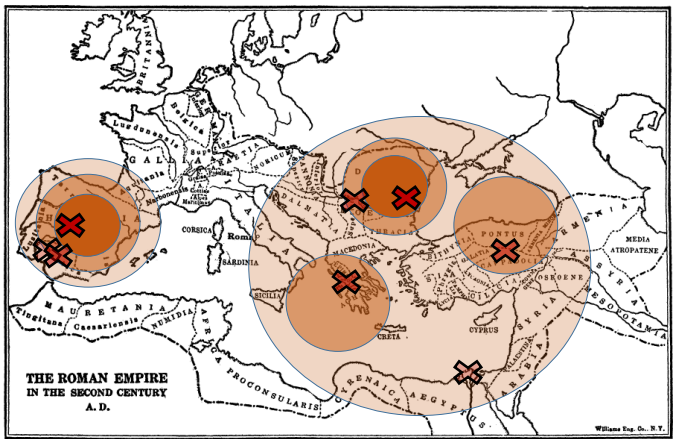


Figure 1.14: Circular Tree Map as stemma combined with a map.

the important factors playing a role in the traditions copy history's narrative.[46] As stated before, other visualizations for clustering, such as multi dimensional scaling and principle component analyses, can also be employed. Summarizing, in visualization, the whole array of possible visual depictions for stemmas has seemingly not yet been fully exploited and the digital medium offers manifold new possibilities of illustration.

1.12 DynStemGen

The dynamic stemma generator (dynStemGen) is a java class which converts trees encoded in the Newick format (with node names given) which can be obtained as output from phylogenetic software (for instance from R and its package "ape" (Paradis et al., 2004; Paradis, 2011)). The information is converted into LaTeX code, which then has to be compiled by a LaTeX compiler. Input as a rooted structure is presupposed, a manual reconfiguration or automatic rooting can be an additional step. The input configuration however is not the duty of the visualization. For the definition of the single timesteps, the user can specify the sequence in which the nodes should appear. Additional edges depicting contamination can be inserted through specification in the user dialogue. See Figure 1.15 for an example of the program's usage.

[46]Narrative textual criticism to which this term refers as a new holistic approach, trying to focus on the whole copy history is described in further detail for instance in Lin (2016).

```
^Choenen@hoenen-ESPRIMO-P9900:~/Documents/ResourcesInternal/dynStem$ java -jar dynamicStemmaImFIN.jar
#####CAPTION INPUT#####
Please input caption and continue with 'Return':
My dynamic stemma
#####TREE INPUT#####
Please enter your stemma backbone tree in Newick format naming all nodes [example "(((m1,m2,m3)Ce,(m4,m5)m6)Beta)Alpha;"]:
(((m1,m2,m3)Ce,(m4,m5)m6)Beta)Alpha;
#####CONTAMINATION INPUT#####
If you wish to generate any additional edges indicating contamination,
please give a list of edges and attributes in the same way as below, otherwise input "0":
Sourcenode-Targetnode,edgetype,pointed
edgetype must be one of: normal|dotted|dashed
pointed must be one of: arrow|noarrow
separator for multiple contamination edges is ";"
example: Beta-m2,dotted,arrow
Beta-m2,dotted,arrow
#####EDGE-SEQUENCE INPUT#####
Please enter your edge sequences including eventual contamination edges or "0" for a static stemma,
Sourcenode-Targetnode OR Sourcenode-(Targetnode1,Targetnode2,...,TargetnodeN)
Multiple Edges or Edgegroups at one time step can be combined with "&"
Separator for different time steps is ";"
example: "Alpha-Beta;Beta-Ce&Ce-(m1,m2);Ce-m3;Beta-m6;m6-(m4,m5)"
Alpha-Beta;Beta-Ce&Ce-(m1,m2);Ce-m3;Beta-m6;m6-(m4,m5)
#####OUTPUT STATEMENT#####
Thank you for using the program, the files should now have been written.
For compilation you will need LaTeX with the packages 'forest' and 'animate' installed, as well as their dependencies such as 'tikz'/pgf.
If you are using Windows please execute (as administrator if necessary) the file run.bat
If you are using Linux please execute (as administrator if necessary) the file run.sh
Your result should be written to "dynStem.pdf".

If you are using the program for a publication, please cite:
Armin Hoenen, "Das erste dynamische Stemma, Pionier des digitalen Zeitalters?",
 in DHd 2016, Modellierung - Vernetzung - Visualisierung - Die Digital Humanities als fächerübergreifendes Forschungsparadigma, Konferenzabstracts. 2016.
http://dhd2016.de/boa-large.pdf
hoenen@hoenen-ESPRIMO-P9900:~/Documents/ResourcesInternal/dynStem$
```

Figure 1.15: The user dialogue for creating a dynamic stemma.

Figure 1.16: Dynamic Stemma first time step.

Figure 1.17: Dynamic Stemma second time step.

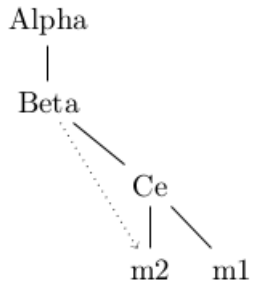


Figure 1.18: Dynamic Stemma third time step.

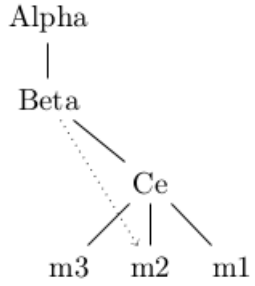


Figure 1.19: Dynamic Stemma fourth time step.

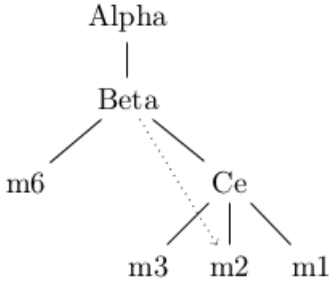


Figure 1.20: Dynamic Stemma fifth time step.

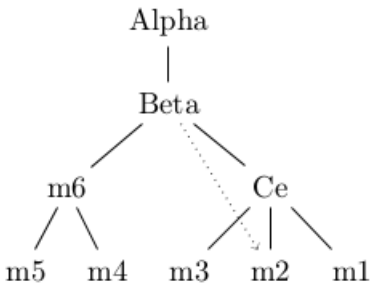


Figure 1.21: Dynamic Stemma sixth time step.

The output is a dynamic pdf, which means that via the LaTeX library "animate",[47] the different timesteps of the stemma are presented for a specified time one after another as slides. The slideshow stops with the ultimate slide, from where it can be repeated. The last slide is congruent with the traditional non dynamic stemma. This is important since although dynamic depictions display more information, they bare an inherent danger. Any of the intermediary slides can slip attention, whereas a static stemma is always in view, meaning that a slip of attention is easier to cope with, not requiring an entire replay. Since the user does not have to specify any edge sequences, the software can be used to produce traditional static depictions as well. The combination of software was used since the output is close to traditional depictions and allows the philological user to manipulate (enrich, change) the LaTeX output for editorial purposes if needed. Figures 3 to 8 show the output of the program for a hypothetical stemma and edge sequence. Since a dynamic image can not be printed, the Figures depict the succession of the single slides.

1.13 Literature Summary

The history of stemmatology proper started after the onset of the European Print age although it has roots in antiquity. With the first stemma appearing in the preface of an edition in 1827, the discipline quickly developed an apparatus of rules and heuristics to interpret and devise stemmata before the advent of the computer. Since the historical truth is hidden, naturally, some criticism at the objectivity of the method can appear and this happened, especially through J.

[47] `https://www.ctan.org/pkg/animate`

Bédiers criticism of editorial 'fraud and fallacy'. His claim divided the field into those who would use stemmatology (Neo-Lachmannians) and those who would abandon it (Bedierists) in favor of their own approaches (for instance editing a non stemmatically determined 'best manuscript'). We made an excursus in which we focussed on the principle of lectio difficilior to understand the complexity and interdisciplinarity of the methods of classical stemmatology. Stemmatology proper immediately after the advent of the computer sees its use in a wide range of tasks among which clustering, rooting, visualization, alignment, exemplar shift detection, stemma simulation and others. In the early 1990ies phylogenetic methods get introduced into the field and come to excert a profound influence on it, where many scholars use phylogenetic programs such as PHYLIP for their data thanks to an easy conversion of aligned textual data into some DNA like sequence. Additionally, neighbor nets as conceptually very different visualizations appear. Some peculiar features of phylogenetic trees such as having only bifurcations and manuscripts at leaf node positions, as well as visualisations which do not reveal possible roots easily may have helped more stemmatologically adapted solutions to appear. The introduction of gold standard datasets made evaluation possible and may have attracted more computer scientific engagement in the task. Since the end of the 2010s some new methods begin to appear which are more stemmatologically adapted. At the same time more and more scholars start using phylogenetic methods for instance in digital editions such as Dantes Commedia edition and to use their outputs as alternative steps (or working hypotheses) in the stemmatic process. Meanwhile, entirely manual stemmatology remains the most important means of stemma generation in classical philology.

Besides narrating the history of the discipline, the chapter has introduced a very simple dynamic visualization tool. Finally, it must be said, that the object of study is difficult in that there is very few gold standard data, in that there are other unresolved problems such as how to use metadata and other caveats which prohibit the use of machine learning methods or entail a grave danger of overfitting. With such an object, a breadth first approach may be a good way to use results from one subtask in order to better conduct the others, which is why this thesis will try to encircle and circumscribe broadly stemmatic methods engaging in different subtasks as far as reason can carry the author in each single endeavour on the testbeds available.

Chapter 2

Theoretical Stemmatology - the Debate on Bifurcativity

2.1 Introduction

In this chapter, the question of the general reliability of manually produced stemmas, ultimately raised by Bédier (1928) is assessed by means of a simulation of stemmas and some statistical considerations. Especially, the question of the underlying distribution for manuscript copying is surveyed under some simplifying assumptions. Arguments will be mathematically assessed. Finally, a simulation of stemma generation including 3 putative loss scenarios are being elaborated and their results with different parameter settings analyzed. This chapter represents work which in slightly different form went into two publications: Hoenen et al. (2017) and Hoenen (2016b). The first of these contains additional content primarily going back to the coauthors and especially the content of 2.4.1 but also a simpler formula for formula 2.3 and an analysis of the convergence of rooted labeled trees can be read in that publication, while here the original approach is being outlined including some additional text, tables, a proof (Appendix) and other details, which would not have all fitted into the paper. 2.5 through 2.8 represent a substantially elaborated version of the second publication.

2.1.1 Haugen

O.E. Haugen has been one author extensively engaging in the assessment of the theory put forth by J. Bédier. Haugen (2002, 2010, 2015) analyses and provides

Series	uni-	bi-	tri-	quad-	quint-	sext-
Bibliotheca A.	56	180	30	7	3	0
Editiones A.	55	165	27	4	3	1

Table 2.1: Numbers of furcations (of any node) in conclusive stemmas of two series of editions of various Old Norse texts after Haugen (2015).

Collection	root bifurcation	root three or multifurcations
Bédier (1928)	95.5%	4.5%
Castellani (1957)	82.5%	17.5%
Haugen (2015) Bibliotheca A.	85.5%	14.5%
Haugen (2015) Editiones A.	80.5%	19.5%

Table 2.2: Percentages of root bifurcative stemmas in four collections, Haugen (2015).

statistics on furcations in an Old Norse tradition. He summarizes the main contributions to the debate and philological practice. He is one of the few who assess the observation included in Bédiers essay that stemmas of recensors tended to be more root multifurcating than those of editors, for which Haugen (2015) found no direct proof in his data. In Old Norse traditions, although the stemmas were 'Lachmannian', the editions were 'Bédierist' based on a best manuscript. Thus, Haugen (2002) concludes, that wanting to choose freely among two variants from two reconstructed texts below the root of the stemma (fraud) could not apply. The force of dichotomy he found realistic (which will naturally also influence root bifurcations, not for fraud, but for fallacy) looking at the distribution of furcations in his data.

He concludes that if stemmas on the continent and in the North both show very large numbers of bifurcations despite being copied in very different modes (more professional and in larger numbers on the continent) then the prevalence of bifurcations must be rather methodological fallacy than historical circumstance. See Table 2.1 for the observed numbers of furcations and Table 2.2 for percentages in different collections enumerated by Haugen (2015).

2.1.2 Other contributions

Greg (1931) in a letter mentions the possibility that heavy decimation could have

led to dominance of bifurcativity. Among other additional reactions to the debate are for instance Irigoin (1954); Erbse (1959); Alberti (1979).

2.1.3 Summarizing remarks

The debate is still not concluded and has seen much activity since 1928. In mathematical terms combinatorics (of tree constitution) and probability calculus have been consulted to assess how expectable bifurcatity and root bifurcativity is apriori. The underlying scenarios often included attributing some unknown historical variable to random. On the other hand philologists have supplied many arguments for and against a methodologically or historically induced high incidence of bifurcations and root bifurcations.

In the debate, terminological and functional categories have cristallized helping to tackle the overall problem. Especially two terms have been shown to be useful in the discussion: The term arbre réel as denoting the complete and true historical tree irrespective of (not before) loss and without any methodological simplification (omission of codici interpositi, the contraction of unifurcation chains) and the term stemma codicum as the tree reconstructed from the surviving manuscripts.

In the next section, the number of possible unrooted labelled trees for n nodes is enumerated and the implications for the debate discussed before a simulation investigates the implications of possible underlying distributions of outdegrees in an arbre réel paired with different simulations of loss.

2.2 Previous works in counting manuscript trees

Although a number of the above enumerated studies count portions of stemmatic trees or stemmatic trees under certain conditions, Flight (1990) is apparently the first one to find a generalized definition for those stemmas possible for three nodes as enumerated by Maas (1937), see Figure 1.3. He invents the notion of *Greg tree* and defines a Greg tree (which he names after the textual critic W. W. Greg) as a tree of m labelled vertices standing for surviving manuscripts and n unlabelled vertices symbolizing hypothetical manuscripts, the latter must have a degree of at least three; for a rooted Greg tree (directed), the distinguished root

vertex is if unlabelled allowed to be of degree 2.[1] This is the formal translation of the way in which Maas (1937) constructs permissible stemmas: an internode is not allowed to have an outdegree of 1. This corresponds to the aforementioned practice to contract unifurcations of lost manuscripts (if one starts from an arbre). In rooted Greg trees, there can be no chains of unlabelled nodes (hypothetical manuscripts) with outdegree one. A rooted Greg tree symbolizes a stemma codicum not an arbre réel. With this definition, the numbers of possible trees for 3 surviving manuscripts as postulated by Maas (1960) are recovered. Root can be labelled or unlabelled.

Flight (1990) gives a recursive formula for the enumeration of (rooted and unrooted) Greg trees, building on all possibilities to add a labelled vertex including a possibility at internode position or root positions.

$$g^*(m,n) = (m+n-2) * g^*(m-1,n-1) \\ + (2m+2n-2) * g^*(m-1,n) * (n+1) * g^*(m-1,n+1)$$

g^* is a rooted Greg tree with m being the surviving manuscripts (=labelled nodes) and n the hypothetical/reconstructed ancestors (= unlabelled nodes). The counting procedure bares some similarity with that presented in Felsenstein (1978) which treats a similar problem. Flight (1990) tabulates all possible Greg trees for up to 12 labelled nodes generalizing from, extending values mentioned by and correcting Maas' numbers, he does not assess root furcativity. The tabulated numbers for rooted Greg trees can be seen from Table 2.3. From the 22 rooted Greg trees for 3 survivors, there are 12 root bifurcating ones, compare again Figure 1.3.

If one wants to determine the number of root bifurcating Greg trees recursively, looking at Flight's condition (2) of how to add a new labelled node stating that this can be done attaching it to any existing one, the problem that arises is that upon addition without knowing the degree of the old node, one cannot know the new degree. Flight (1990) does not give a proportion of root bifurcations.

2.3 Percentage of root bifurcations in arbres reels

Flight (1990, p.122) aims at solving the question, he attributes to Maas (1958),

[1] In what follows, the trees of interest are all implicitly assumed to be directed if not otherwise stated.

m	$g^*(m)$
1	1
2	3
3	22
4	262
5	4 336
6	91 984
7	2 381 408
8	72 800 928
9	2 566 606 784
10	102 515 201 984
11	4 575 271 116 032
12	225 649 908 491 264
13	12 187 240 730 230 528
14	715 392 567 595 403 520
15	45 349 581 052 869 924 352
16	3 087 516 727 770 990 992 896
17	224 691 760 916 830 871 873 536
18	17 406 010 163 637 184 225 490 944
19	1 430 047 520 046 948 331 002 540 032
20	124 200 350 766 670 456 501 164 404 736
21	11 369 658 029 287 586 374 482 405 343 232
22	1 094 163 229 267 119 931 524 656 040 820 736
23	110 431 651 791 290 984 684 732 682 274 340 864
24	11 663 858 850 495 260 368 338 450 544 310 288 384
25	1 286 679 332 089 295 797 268 733 166 710 207 741 952

Table 2.3: Numbers of rooted Greg trees for m surviving manuscripts, (1..12 as in Flight (1990)), 13..25 supplemented).

"for some given number of surviving manuscripts, how many different stemmata may exist?". As we have seen, in order to solve the question, he invents a generating function for and counts Greg trees.[2] Greg trees most congruently describe stemmas, or possible stemmas enumerated for m survivors. They project to different possible arbres as reconstructible truths and comprise a set of arbres of different sizes. However, to tackle the problem from the other side, since there is ultimately only one historically faithful arbre, one could start from this arbre. Consider Kleinlogel (1968, p.68):

> Castellani [meant is Castellani (1957)] geht von der richtigen Überlegung aus, daßman die Frequenz bzw. die Wahrscheinlichkeit zwei- und mehrspaltiger Stemmata im Grunde nur dann zu überblicken vermag, wenn man hypothetisch den ursprünglichen Gesamtbestand der Tradition mit allen Verästelungen zugrunde legt und daran untersucht, wie großeffektiv die Aussichten waren, daßsich eine feste Zahl erhaltener Handschriften auf zwei oder mehr Substem-

[2]According to Josuat-Vergès (2015), a similar problem in phylogeny has been described and tackled by Felsenstein (1978) as recognized by Knuth (2005). Najock and Heyde (1982) counts expectable numbers of leafs for trees of size n.

mata verteilen konnte.[3]

Underlying arbres réels for a number of n nodes and their root bifurcativity patterns have, to the author's best knowledge not been investigated in the way described here. We ask, how large the percentage of root bifurcations is among all possible trees, in the case of stemmatology rooted labelled trees, for a given number of nodes n.

As Harper Jr (1976) finds, building on Cayley (1889), in a labelled rooted tree, $T = (V, E, \nu)$ the number of different rooted labelled trees for n nodes is:

$$n^{n-1} \tag{2.1}$$

In order to determine the proportion of bifurcations directly below the root, their number is determined generating/enumerating all possible root bifurcating trees over the same set of labelled nodes V as present in T. Throughout the section large numbers will be displayed not everywhere in scientific notation or on logarithmic scales, so as to provide exact numbers and allow exact visual inspection and impression.

1. Each of the possible trees must be rooted and has thus one out of n possible roots.
2. Below each possible root, there is to be a bifurcation. This entails that cutting the two outgoing edges below root, a partition of the tree into the root and two rooted subtrees is achieved. The sum of the number of nodes of the two subtrees is $n-1$ or $|V \setminus \nu|$. Since neither the empty set (definition of set partition) nor the complete set of possible nodes (in that case, the second set would have to be empty, which is again prohibited) is permitted as any subset of a binary partition, the size of the subtrees (number of nodes) must range from 1 to $n-2$.
3. Each possible combination of size of subset a and size of subset b must satisfy $|a| + |b| = n - 1$. In order to not count possible subtrees twice, only instances, where $|a| \geq |b|$ have to be considered. This is equal to considering only unordered sets of the size pairs of the partitions (and ultimately trees). Each possible partition can be obtained by incrementing a counter

[3]Castellani departs from the correct assumption that the frequency resp. probability of bi- and multifid stemmas can actually only be assessed, if one lays out the entire hypothetical whole of a tradition with all branching patterns and then investigates how large the probabilities effectively were for a fixed number of surviving manuscripts to be distributed onto two or more substemmas.

k for all possible subset sizes in the range 1 to $n-2$ and determining the size of the other subset as complement to k, hence $n-1-k$. However, for subsets of the same size, the number of counted combinations must be divided by 2, since for each combination of i chosen and j discarded elements, the complement of the same j chosen and the same i discarded elements would induce the same sets a second time.

4. Now, for each possible bipartite partition, the number of all possible combinations of labelled nodes into these partitions can be determined by the binomial coefficient, since the binomial coefficient induces for each combination a partition into the chosen and the discarded elements (definition binomial coefficient). This number can then be computed as $\binom{n-1}{|a|}$.
5. For each so obtained possible combination of nodes in the subsets, one can apply formula (1) again to generate/count the sum of possible rooted labelled trees induced by these subsets, for instance $|a|^{|a|-1}$. With each tree in set a, each tree in set b can be combined, leading to the product of those for each partition equalling the number of possible pairs of trees attachable to root.

What follows from 1. to 5. is one possible formula for the calculation of the proportion of root bifurcating labelled trees among all labelled rooted trees for n nodes:

$$\frac{n * \sum_{k=1}^{n-2} \frac{\binom{n-1}{k}}{s} * (k^{k-1} * (n-1-k)^{n-2-k})}{n^{n-1}} \tag{2.2}$$

$$k \geq n-1-k, n > 2, \begin{cases} s = 2, & \text{for } k = n-k-1 \\ s = 1, & \text{else} \end{cases}$$

This is equal to

$$\frac{\sum_{k=1}^{n-2} \binom{n-1}{k} * (k^{k-1} * (n-1-k)^{n-2-k})}{2n^{n-2}}; n > 2 \tag{2.3}$$

since all possible pairs are being generated exactly twice. A proof and examples can be found in the Appendices (IV.).

Vertices	Trees (rb)	All Trees	Proportion
1	0	1	0
2	0	2	0
3	3	9	0.667
4	24	64	0.563
5	240	625	0.512
6	3 000	7 776	0.482
7	45 360	117 649	0.463
8	806 736	2 097 152	0.449
9	16 515 072	43 046 721	0.438
10	382 637 520	1 000 000 000	0.430
11	9 900 000 000	25 937 424 601	0.424
12	282 953 722 920	743 008 370 688	0.419
13	8 854 183 084 032	23 298 085 122 481	0.415
14	301 082 946 198 216	793 714 773 254 144	0.411
15	11 055 312 913 182 720	29 192 926 025 390 625	0.408
16	435 947 695 312 500 000	1 152 921 504 606 846 976	0.405
17	18 374 686 479 671 623 680	48 661 191 875 666 868 481	0.403
18	824 377 838 834 826 948 384	2 185 911 559 738 696 531 968	0.401
19	39 224 968 544 199 943 323 648	104 127 350 297 911 241 532 841	0.399
20	1 972 939 268 802 528 786 938 040	5 242 880 000 000 000 000 000 000	0.397
21	104 595 456 000 000 000 000 000 000	278 218 429 446 951 548 637 196 401	0.396
22	5 829 338 521 745 651 495 255 543 640	15 519 448 971 100 888 972 574 851 072	0.394
23	340 722 447 865 533 153 352 438 775 808	907 846 434 775 996 175 406 740 561 329	0.393
24	20 840 996 415 727 216 548 467 783 320 944	55 572 324 035 428 505 185 378 394 701 824	0.392
25	1 331 420 263 348 807 936 733 024 039 731 200	3 552 713 678 800 500 929 355 621 337 890 625	0.391

Table 2.4: Numbers of root bifurcatig (rb) trees among all labelled rooted trees. Proportion for $n = 100$ is 0.373, there are roughly $3.61 * 10^{197}$ trees; for $n = 1000$ it's 0.368 with roughly $3.68 * 10^{2996}$ trees.

2.3.1 Resulting numbers

As for root unifurcations, there are $n * (n-1)^{(n-2)}$ planted trees, since there are $(n-1)^{(n-2)}$ labelled trees rooted by the node below root, see Table 2.5.

In summary, the proportion of possible root bifurcating arbres is large at least for the range of historically observed tradition sizes (according to Weitzman (1987, p.292) 1 to well over 100). The proportion would be even larger, if the root unifurcations would disappear into another context of n because of the philological practice of collapsing hypothetical unifurcation chains. In that case (as an upper limit provided root is always lost), the proportion of root bifurcating trees at $3, 10, 50, 100, 1000$ manuscripts would be $1, 0.672, 0.598, 0.59, 0.583$ respectively.

2.4 Root bifurcating Greg trees

Now, we can apply a similar counting procedure as the one for rooted labelled trees to count root bifurcating Greg trees, which we should expect apriori for m survivors. For each number m of surviving manuscripts, there are $n = 0..m -$

Vertices	Trees (ru)	All Trees	Proportion
1	0	1	0
2	2	2	1
3	6	9	0.667
4	36	64	0.563
5	320	625	0.512
6	3 750	7 776	0.482
7	54 432	117 649	0.463
8	941 192	2 097 152	0.449
9	18 874 368	43 046 721	0.438
10	430 467 210	1 000 000 000	0.430
11	11 000 000 000	25 937 424 601	0.424
12	311 249 095 212	743 008 370 688	0.419
13	9 659 108 818 944	23 298 085 122 481	0.415
14	326 173 191 714 734	793 714 773 254 144	0.411
15	11 905 721 598 812 160	29 192 926 025 390 625	0.408
16	467 086 816 406 250 000	1 152 921 504 606 846 976	0.405
17	19 599 665 578 316 398 592	48 661 191 875 666 868 481	0.403
18	875 901 453 762 003 632 658	2 185 911 559 738 696 531 968	0.401
19	41 532 319 635 035 234 107 392	104 127 350 297 911 241 532 841	0.399
20	2 082 547 005 958 224 830 656 820	5 242 880 000 000 000 000 000 000	0.397
21	110 100 480 000 000 000 000 000 000	278 218 429 446 951 548 637 196 401	0.396
22	6 120 805 447 832 934 070 018 320 822	15 519 448 971 100 888 972 574 851 072	0.394
23	356 947 326 335 320 446 369 221 574 656	907 846 434 775 996 175 406 740 561 329	0.393
24	21 788 314 434 623 908 209 761 773 471 896	55 572 324 035 428 505 185 378 394 701 824	0.392
25	1 389 308 100 885 712 629 634 459 867 545 600	3 552 713 678 800 500 929 355 621 337 890 625	0.391

Table 2.5: Numbers and proportions of planted labelled trees (root unifurcating trees, ru) among all labelled rooted trees. Proportion for $n = 100$ is 0.373, there are roughly $3.73 * 10^{197}$ trees; for $n = 1000$ it's 0.368 with roughly $3.68 * 10^{2996}$ trees.

1 internodes possibly added in a Greg tree. The root can be either labelled or unlabelled. For each combination of m with n, depending on the quality of root, there are m labelled and $n - 1$ unlabelled nodes in the subset below root, which must be partitioned into two, if we want a root bifurcation; or there are $m-1$ and n nodes. For each setting, we can partition the unlabelled nodes into two sets and at the same time, we produce every possible partition of the labelled nodes into two sets and then we combine the sets in each possible unordered way. Now, for each combination of unlabelled and labelled nodes, there are $\binom{m}{l}$ permutations if root is unlabelled and $\binom{m-1}{l}$ permutations if root is labelled; l being the size of the subset of labelled nodes. Counting permutations for the unlabelled nodes is unnecessary. Further on, the number of rooted Greg trees for each such subset combination can be obtained from the recursively precomputed tables of rooted Greg trees for m survivors and n internodes according to Flight (1990), (numbers computationally expanded), see Table 2.6.

However, since an internode must have degree three, the number of labelled nodes from which there are Greg trees if the number of unlabelled nodes is exactly one, cannot be smaller than two (in other words, the empty fields above

	$n = 0$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$m = 0$											
1	1										
2	2	1									
3	9	10	3								
4	64	113	70	15							
5	625	1526	1450	630	105						
6	7 776	24 337	31 346	20 650	6 930	945					
7	117 649	450 066	733 845	650 188	329 175	90 090	10 395				
8	2 097 152	9 492 289	18 760 302	20 925 065	14 194 180	5 845 455	1 351 350	135 135			
9	43 046 721	225 159 022	523 411 836	704 731 170	600 063 310	330 420 090	114 774 660	22 972 950	2 027 025		
10	1 000 000 000	5 937 424 601	15 880 844 122	25 071 331 516	25 710 520 550	17 741 233 510	8 229 931 710	2 472 970 500	436 486 050	34 459 425	

Table 2.6: Numbers of rooted Greg trees for m labelled and n unlabelled nodes as tabulated up to $n = 5$ by Flight (1990). Note, that for combinations, where $n >= m$ (diagonale and above) there are implicitly no trees.

m	$G^*_{rb}(m)$	$\frac{G^*_{rb}}{G^*(m)}$
1	0	0
2	1	0.333
3	12	0.545
4	151	0.576
5	2 545	0.587
6	54 466	0.592
7	1 417 318	0.595
8	43 472 780	0.597
9	1 536 228 588	0.599
10	61 466 251 616	0.6
11	2 746 907 348 768	0.6
12	135 619 260 805 568	0.601
13	7 331 022 129 923 648	0.602
14	430 638 151 053 316 480	0.602
15	27 315 015 477 709 844 352	0.602
16	1 860 627 613 021 322 933 248	0.603
17	135 465 573 609 158 928 964 096	0.603
18	10 498 038 569 346 091 127 451 136	0.603
19	862 792 664 850 194 915 870 874 112	0.603
20	74 956 476 321 749 641 725 226 812 416	0.604
21	6 863 570 707 505 269 884 254 448 731 136	0.604
22	660 677 227 364 107 011 845 607 225 147 392	0.604
23	66 695 290 463 729 869 207 893 188 983 046 144	0.604
24	7 045 786 917 185 412 308 910 365 013 952 397 312	0.604
25	777 384 096 762 179 732 141 765 186 486 469 263 360	0.604

Table 2.7: Numbers and proportions of root bifurcating Greg trees (G^*_{rb}) for m surviving manuscripts. Note that the first numbers agree with Hering (1967).

the diagonal of the table of rooted Greg trees, which Flight (1990) gives must be 0). Now since l and k are complementary and since all combinations of labelled and unlabelled nodes have been summed twice in differing sequence, we have produced each tree four times. We must thus divide by 4. Combining all subsets we arrive at the number of all possible root bifurcating Greg trees for a given number of survivors and conversely at the proportion of them. For a fixed m:

$$\begin{aligned}&\sum_{n=0}^{m-1}(\frac{m}{4}*\sum_{k=0}^{n}\sum_{l=0}^{m-1}\binom{m-1}{l}*((g^*(l,k)*g^*(m-1-l,n-k))+(g^*(l,n-k)*g^*(m-1-l,k)))\\&+\frac{1}{4}*\sum_{k=0}^{n-1}\sum_{l=0}^{m}\binom{m}{l}*((g^*(l,k)*g^*(m-l,n-1-k))+(g^*(l,n-1-k)*g^*(m-l,k))))\end{aligned} \tag{2.4}$$

See Table 2.7 for numbers and proportions of root bifurcating Greg trees. Further combinatorial clarifications and examples are to be found in the Appendices (IV.). Root unifurcating Greg trees are easily computed. The root can only be labelled, since an unlabelled node as root must have degree at least 2. Then, the number of possible root unifurcating Greg trees simply corresponds to $m * g^*(m-1)$. See Table 2.8.

m	$G^*_{ru}(m)$	$\frac{G^*_{ru}}{G^*(m)}$
1	0	0
2	2	0.667
3	9	0.409
4	88	0.336
5	1 310	0.302
6	26 016	0.283
7	643 888	0.270
8	19 051 264	0.262
9	655 208 352	0.255
10	25 666 067 840	0.250
11	1 127 667 221 824	0.246
12	54 903 253 392 384	0.243
13	2 933 448 810 386 432	0.241
14	170 621 370 223 227 392	0.239
15	10 730 888 513 931 052 800	0.237
16	725 593 296 845 918 789 632	0.235
17	52 487 784 372 106 846 879 232	0.234
18	4 044 451 696 502 955 693 723 648	0.232
19	330 714 193 109 106 500 284 327 936	0.231
20	28 600 950 400 938 966 620 050 800 640	0.230
21	2 608 207 366 100 079 586 524 452 499 456	0.229
22	250 132 476 644 326 900 238 612 917 551 104	0.229
23	25 165 754 273 143 758 425 067 088 938 876 928	0.228
24	2 650 359 642 990 983 632 433 584 374 584 180 736	0.227
25	291 596 471 262 381 509 208 461 263 607 757 209 600	0.227

Table 2.8: Numbers and proportions of root unifurcating Greg trees (G^*_{ru}) for m surviving manuscripts. Note that the first numbers agree with Hering (1967).

2.4.1 Generalisation

The content of this section primarily goes back to the work of S. Eger, who has provided a mathematically more compressed and generalising way of representing Greg Tree root furcation formulas by selecting all tuples of labelled nodes for the subtrees directly from the binomial coefficients. The main Java implementation however goes back to my own work and is almost congruent with this formalisation (but counts trees for all k at once, see below). For this reason and for completeness and ease of formal representation, the main ideas shall be briefly repeated here. For details, please consider and cite Hoenen et al. (2017).

In order to generalise, instead of partitioning the number of labelled nodes respectively into two non-empty subsets one must partition them into k such sets.[4] There are $\binom{n}{s_1,\ldots,s_k}$ possibilities to do so. Counting these sets representing subtrees in this way, we overcount since we count each tuple $(s_1, \ldots, s_k)$ as distinct, while the same sets of subtrees make up the same tree regardless of the order of their elements. Thus, it must be divided by $k!$. For a fixed $(s_1, \ldots, s_k)$,

[4]Instead of extending (2.4) to ever more subtrees with ever more combinations of unlabelled and labelled nodes and an adjustment of the denominators, the general formula composes tuples in the iteration below the sum.

although we have distributed the numbers of labelled nodes among the subtrees, each subtree of s_i labelled nodes can have $0 \dots s_i - 1$ or $a_i < s_i, a_i \in \mathbb{N}$ additional unlabelled nodes (see rooted Greg tree definition). The number of possible trees for each combination (s_i, a_i) is always given by the precomputed/tabulated numbers of (m,n)-trees. Thus, for each subset s_i, there are a_i possible (m,n)-tuples (or (s_i, a_i)-tuples) consituting possible rooted Greg trees for the corresponding branch. For each s_i we can combine any one of such tuples in each possible way with any one of the other subbranches (s_i) with different i (exactly only one per subbranch) leading to a product of k factors of (m,n)-trees. The sum of these products is then to be taken as basis for the permutations of the labelled nodes and finally, it has to be summed over all different $(s_1, \dots, s_k)$. One must however split the counting between such trees, which have one of m possible labelled nodes and $m - 1$ additional labelled nodes to be distributed among the subtrees and such trees, which have an unlabelled node and m labelled nodes for the subtrees. Formally, this can be expressed as:

$$\begin{aligned} &\frac{m}{k!} * \sum_{(s_1,\dots,s_k)\in\mathcal{C}(m-1,k)} \binom{m-1}{s_1,\dots,s_k} \sum_{p\in P_{s_1,\dots,s_k}} g_k^*(p) \\ &+ \frac{1}{k!} * \sum_{(s_1,\dots,s_k)\in\mathcal{C}(m,k)} \binom{m}{s_1,\dots,s_k} \sum_{p\in P_{s_1,\dots,s_k}} g_k^*(p) \end{aligned} \tag{2.5}$$

Here, p is a tuple of two natural numbers (including 0), where the first stands for the number of labelled and the second for the number of unlabelled nodes. All combinations of possible tuples are defined by

$$P_{s_1,\dots,s_k} = \{((s_1, a_1), \dots, (s_k, a_k))|a_i \in \mathbb{N}, a_i < s_i, i = 1, \dots, k\}$$

. We define g_k^* as the product of the numbers of rooted Greg trees obtained by the k subsets in a subset p of $P_{s_1,\dots,s_k}$:

$$g_k^*((s_1, a_1), \dots, (s_k, a_k)) = \prod_{i=1}^{k} g^*(s_i, a_i)$$

As an example, if $k = 3$ and $s_1 \dots, s_k = \{2, 2, 1\}$, then the inner sum sums the corresponding products of the numbers of $g^*(2,0)g^*(2,0)g^*(1,0)$, $g^*(2,0)g^*(2,1)g^*(1,0)$, $g^*(2,1)g^*(2,0)g^*(1,0)$ and $g^*(2,1)g^*(2,1)g^*(1,0)$. (2.5)

can also be regarded as a generalisation of rooted labelled trees (where simply the number of unlabelled nodes is always 0) simplifying the formula further. For details, see Hoenen et al. (2017).

Tables 2.9 and 2.10 show the growth of furcations until 25 (with exact numbers until $m = 10$ and for (m-1)-trees and numbers in scientific notation above). Athough it is true, that the proportion of multifurcations against bifurcations increases, the same is true for the ratio of bi- to unifurcations, resulting in the fact, that the proportion of bifurcations among all trees continues to grow, leaving root bifurcations the largest expectable root furcation pattern at any observed point.

Note, that following from their topology, the root k-furcating Greg trees for $k = 1$ coincide with the above mentioned numbers and the root (m-1)-furcating trees for all $m \neq 2$ with the pentagonal numbers (sequence $A000326$ in the OEIS[5]);[6] for a root m-furcation, there is always only 1 tree. These equalities can be used to slightly speed up computing.

2.5 Intermediary Conclusion and Philological Debate

The proportions of root bifurcating trees presented in connection with stemmatology and the Lachmann-Bédier debate need some interpretative intervention

[5] `https://oeis.org/`

[6] This is so [authors finding], because there are only 3 possible types of (m-1) root furcating trees:

1. the case in which there are no unlabelled nodes, which can have m different roots, therefore there are m such trees
2. the case in which there is one unlabelled node (root), for which there are $m * (m - 1)$ trees, since in one branch there must be an additional labelled node attached, for which there are that number of possible combinations of labels
3. the case in which one unlabelled node is child of root (unlabelled) and has 2 children: this case has $\binom{m}{2}$ combinations of labels

Further trees with more unlabelled nodes cannot exist, because of the restriction for unlabelled nodes to be of degree 2 (root) or 3 (any other). The sum of the three enumerated topologies is equal to the formula for pentagonal numbers, $\frac{3n^2-n}{2}$, which one immediately sees doing some basic arithmetics, but in the case of $m = 2$, root cannot be unlabelled, nor can any unlabelled node be inserted so that the result diverges from the pentagonal series as addends 2 and 3 above are absent.

m	1	2	3	4	5	6	7	8	9	10	11	12
1	0	-	-	-	-	-	-	-	-	-	-	-
2	2	1	-	-	-	-	-	-	-	-	-	-
3	9	12	1	-	-	-	-	-	-	-	-	-
4	88	151	22	1	-	-	-	-	-	-	-	-
5	1 310	2 545	445	35	1	-	-	-	-	-	-	-
6	26 016	54 466	10 425	1 025	51	1	-	-	-	-	-	-
7	643 888	1 417 318	286 321	31 780	2 030	70	1	-	-	-	-	-
8	19 051 264	43 472 780	9 102 604	1 090 201	80 360	3 626	92	1	-	-	-	-
9	655 208 352	1 536 228 588	329 980 456	41 636 973	3 368 001	178 290	6 006	117	1	-	-	-
10	25 666 067 840	61 466 251 616	13 457 494 060	1 763 775 280	152 280 345	8 964 417	358 890	9 390	145	1	-	-
11	$1.13 * 10^{12}$	$2.75 * 10^{12}$	$6.1 * 10^{11}$	$8.23 * 10^{10}$	$7.46 * 10^{9}$	$4.74 * 10^{8}$	$2.13 * 10^{7}$	$6.61 * 10^{5}$	$1.4 * 10^{4}$	176	1	-
12	$5.49 * 10^{13}$	$1.36 * 10^{14}$	$3.05 * 10^{13}$	$4.21 * 10^{12}$	$3.96 * 10^{11}$	$2.66 * 10^{10}$	$1.3 * 10^{9}$	$4.64 * 10^{7}$	$1.18 * 10^{6}$	$2.02 * 10^{4}$	210	1
13	$2.93 * 10^{15}$	$7.33 * 10^{15}$	$1.66 * 10^{15}$	$2.34 * 10^{14}$	$2.27 * 10^{13}$	$1.59 * 10^{12}$	$8.31 * 10^{10}$	$3.25 * 10^{9}$	$9.41 * 10^{7}$	$1.97 * 10^{6}$	$2.82 * 10^{4}$	247
14	$1.71 * 10^{17}$	$4.31 * 10^{17}$	$9.86 * 10^{16}$	$1.41 * 10^{16}$	$1.39 * 10^{15}$	$1.02 * 10^{14}$	$5.58 * 10^{12}$	$2.34 * 10^{11}$	$7.47 * 10^{9}$	$1.71 * 10^{8}$	$3.16 * 10^{6}$	$3.83 * 10^{4}$
15	$1.07 * 10^{19}$	$2.73 * 10^{19}$	$6.29 * 10^{18}$	$9.01 * 10^{17}$	$9.2 * 10^{16}$	$6.89 * 10^{15}$	$3.94 * 10^{14}$	$1.75 * 10^{13}$	$6.03 * 10^{11}$	$1.61 * 10^{10}$	$3.27 * 10^{8}$	$4.89 * 10^{6}$
16	$7.26 * 10^{20}$	$1.86 * 10^{21}$	$4.31 * 10^{20}$	$6.21 * 10^{19}$	$6.48 * 10^{18}$	$4.97 * 10^{17}$	$2.94 * 10^{16}$	$1.36 * 10^{15}$	$4.91 * 10^{13}$	$1.45 * 10^{12}$	$3.27 * 10^{10}$	$5.69 * 10^{8}$
17	$5.25 * 10^{22}$	$1.35 * 10^{23}$	$3.16 * 10^{22}$	$4.65 * 10^{21}$	$4.86 * 10^{20}$	$3.8 * 10^{19}$	$2.31 * 10^{18}$	$1.11 * 10^{17}$	$4.28 * 10^{15}$	$1.32 * 10^{14}$	$3.26 * 10^{12}$	$6.34 * 10^{10}$
18	$4.04 * 10^{24}$	$1.05 * 10^{25}$	$2.46 * 10^{24}$	$3.65 * 10^{23}$	$3.86 * 10^{22}$	$3.07 * 10^{21}$	$1.91 * 10^{20}$	$9.51 * 10^{18}$	$3.81 * 10^{17}$	$1.24 * 10^{16}$	$3.27 * 10^{14}$	$6.96 * 10^{12}$
19	$3.31 * 10^{26}$	$8.63 * 10^{26}$	$2.03 * 10^{26}$	$3.03 * 10^{25}$	$3.24 * 10^{24}$	$2.62 * 10^{23}$	$1.67 * 10^{22}$	$8.41 * 10^{20}$	$3.53 * 10^{19}$	$1.11 * 10^{18}$	$3.35 * 10^{16}$	$7.65 * 10^{14}$
20	$2.86 * 10^{28}$	$7.41 * 10^{28}$	$1.77 * 10^{28}$	$2.66 * 10^{27}$	$2.87 * 10^{26}$	$2.35 * 10^{25}$	$1.52 * 10^{24}$	$7.94 * 10^{22}$	$3.39 * 10^{21}$	$1.11 * 10^{20}$	$3.51 * 10^{18}$	$8.41 * 10^{16}$
21	$2.61 * 10^{30}$	$6.86 * 10^{30}$	$1.62 * 10^{30}$	$2.46 * 10^{29}$	$2.67 * 10^{28}$	$2.22 * 10^{27}$	$1.46 * 10^{26}$	$7.76 * 10^{24}$	$3.4 * 10^{23}$	$1.24 * 10^{22}$	$3.77 * 10^{20}$	$9.62 * 10^{18}$
22	$2.5 * 10^{32}$	$6.61 * 10^{32}$	$1.57 * 10^{32}$	$2.38 * 10^{31}$	$2.61 * 10^{30}$	$2.19 * 10^{29}$	$1.46 * 10^{28}$	$7.91 * 10^{26}$	$3.54 * 10^{25}$	$1.33 * 10^{24}$	$4.19 * 10^{22}$	$1.11 * 10^{21}$
23	$2.52 * 10^{34}$	$6.67 * 10^{34}$	$1.59 * 10^{34}$	$2.42 * 10^{33}$	$2.67 * 10^{32}$	$2.26 * 10^{31}$	$1.53 * 10^{30}$	$8.31 * 10^{28}$	$3.84 * 10^{27}$	$1.47 * 10^{26}$	$4.79 * 10^{24}$	$1.32 * 10^{23}$
24	$2.65 * 10^{36}$	$7.05 * 10^{36}$	$1.68 * 10^{36}$	$2.58 * 10^{35}$	$2.86 * 10^{34}$	$2.44 * 10^{33}$	$1.66 * 10^{32}$	$9.29 * 10^{30}$	$4.32 * 10^{29}$	$1.61 * 10^{28}$	$5.65 * 10^{26}$	$1.61 * 10^{25}$
25	$2.92 * 10^{38}$	$7.77 * 10^{38}$	$1.86 * 10^{38}$	$2.86 * 10^{37}$	$3.19 * 10^{36}$	$2.74 * 10^{35}$	$1.89 * 10^{34}$	$1.07 * 10^{33}$	$5.05 * 10^{31}$	$2.02 * 10^{30}$	$6.89 * 10^{28}$	$2.02 * 10^{27}$

Table 2.9: Numbers of k-furcations for m-Greg trees. Note, that the first numbers until $m = 6$ occur in Hering (1967) (apart from where $m = 1$, where there is one Greg Tree, which doesn't have any furcation) and that all (m-1)-trees with $m \neq 2$ equal the respective pentagonal numbers.

m	13	14	15	16	17	18	19	20	21	22	23	24	25
13	1	-	-	-	-	-	-	-	-	-	-	-	-
14	287	1	-	-	-	-	-	-	-	-	-	-	-
15	$5.01 * 10^4$	330	1	-	-	-	-	-	-	-	-	-	-
16	$7.35 * 10^6$	$6.65 * 10^4$	376	1	-	-	-	-	-	-	-	-	-
17	$9.56 * 10^8$	$1.08 * 10^7$	$8.53 * 10^4$	425	1	-	-	-	-	-	-	-	-
18	$1.18 * 10^{11}$	$1.55 * 10^9$	$1.54 * 10^7$	$1.08 * 10^5$	477	1	-	-	-	-	-	-	-
19	$1.42 * 10^{13}$	$2.11 * 10^{11}$	$2.46 * 10^9$	$2.16 * 10^7$	$1.35 * 10^5$	532	1	-	-	-	-	-	-
20	$1.7 * 10^{15}$	$2.78 * 10^{13}$	$3.66 * 10^{11}$	$3.79 * 10^9$	$2.97 * 10^7$	$1.66 * 10^5$	590	1	-	-	-	-	-
21	$2.05 * 10^{17}$	$3.61 * 10^{15}$	$5.25 * 10^{13}$	$6.16 * 10^{11}$	$5.71 * 10^9$	$4.02 * 10^7$	$2.03 * 10^5$	651	1	-	-	-	-
22	$2.49 * 10^{19}$	$4.61 * 10^{17}$	$7.38 * 10^{15}$	$9.58 * 10^{13}$	$1.01 * 10^{12}$	$8.43 * 10^9$	$5.37 * 10^7$	$2.45 * 10^5$	715	1	-	-	-
23	$3.09 * 10^{21}$	$6.15 * 10^{19}$	$1.03 * 10^{18}$	$1.46 * 10^{16}$	$1.61 * 10^{14}$	$1.62 * 10^{12}$	$1.22 * 10^{10}$	$7.07 * 10^7$	$2.94 * 10^5$	782	1	-	-
24	$3.92 * 10^{23}$	$8.16 * 10^{21}$	$1.45 * 10^{20}$	$2.19 * 10^{18}$	$2.78 * 10^{16}$	$2.94 * 10^{14}$	$2.54 * 10^{12}$	$1.74 * 10^{10}$	$9.2 * 10^7$	$3.41 * 10^5$	852	1	-
25	$5.07 * 10^{25}$	$1.1 * 10^{24}$	$2.05 * 10^{22}$	$3.28 * 10^{20}$	$4.47 * 10^{18}$	$5.15 * 10^{16}$	$4.95 * 10^{14}$	$3.81 * 10^{12}$	$2.45 * 10^{10}$	$1.18 * 10^8$	$4.13 * 10^5$	925	1

Table 2.10: Numbers of k-furcations for m-Greg trees. All (m-1)-trees equal the respective pentagonal numbers.

at this point. The above generated rooted labelled trees could in theory either be interpreted as arbres réels or as stemmata codici, or even as an intermediary between both, filtered for loss but not contracted.[7] If taken as the arbres réels, if all of them were equally likely which is the best assumption in absence of better definitions, the apriori probability of a root bifurcation tree would be quite large. However, different topologies may show very different patterns. What implications this bares for the root bifurcativity of stemmata codici needs further examination. It would relate to the probability of the root bifurcation pattern of an arbre réel to be reflected in the stemma. Trovato and Guidi (2004) have shown, and Weitzman (1982, 1987) comes to a similar conclusion, that historically realistic scenarios of loss would imply a large quantity of bifurcations and root bifurcations in the selected stemmas based on arbres réels. Those do exceed the observed proportions. A thinkable effect of historical loss is to increase the percentage of root bifurcations in which case the observed proportions would rather operate as a lower bound. However, as Trovato and Guidi (2004) have remarked, the topology of the tree (that is in their case assymmetry) may have a profound influence on the resulting pattern.

As for the number of root bifurcating Greg trees, the proportions are large. At a number of 100 survivors (the range now quasi covers historically relevant traditions) the proportion has still been 0.606. From the table for root k-furcating rooted Greg trees it can be seen, that the lower the furcation, the slower the growth. Within historically probable tradition sizes however, root bifurcations outweigh all other furcation patterns by far. This is an important result of the restriction for outdegrees of unlabelled nodes, or in philological terms the practice (and partly theory) not to postulate codici interpositi.

In summary, a number of studies using combinatorics or probability calculus with various degrees of historical adaptation has shown that there is a large incidence of expectable root bifurcations, consider for instance Weitzman (1987); Trovato and Guidi (2004); Hering (1967). While the agreement between many of the studies implies that Bédier's observations were at least in part due not to fraud and fallacy, but to the mathematical nature of the problem, at the same time fraud and fallacy are not disqualified thereby as possibilities. It appears,

[7] For enumerating the number of possible contracted stemmas for n nodes, it suffices to subtract all planted trees from the respective results. That number could be basis for interpretation as intermediary step contracted but without loss.

that some fraction of the observed root bifurcativity could indeed follow Bédier's suspections. In order to assess the human tendency for overseparation and to a limited degree that of editorial root bifurcation fraud, possibly the best way would be a psychological experiment, where philologists would be given surviving manuscripts of an artificial tradition with known (but to them hidden) genealogy and then draw a stemma from which in comparison to the arbre réel, the degree of overseparation could be measured.[8] Baret et al. (2004) present two manually constructed stemmas for an artificial tradition: one has a root bifurcation, where in the arbre there was none and one has none. Already from this tiny example one can easily imagine that the degree of bias may well depend on the person or on such extrinsic factors as experience, skill, time pressure, payment or personality. The study of Andrews (2014) is assessing the reliability of reconstructions in a similar vein.

2.6 Simulation

What has not been assessed by means of a simulation to the best knowledge of the author is the underlying distribution of outdegrees in arbres réels and their putative influence on furcativity in stemmata.[9] A knowledge of the distributions of outdegrees could help assessing the probabilities of certain topologies of arbres.

2.6.1 Theoretical prequisites

How often did scribes copy manuscripts in the chirographic age? The question will not be answerable with any fair degree of certainty considering its historical nature, but since it has consequences for the ways in which we reconstruct genealogies, a look into possible scenarios may be considered worthwhile. This said, a simulation can be used to test hypotheses. Various authors have imagined scenarios connected with this question, especially the question of how often an archetype or an original has been copied (Bédier debate). For instance, Langosch

[8] Mathematics cannot ultimately be used to proof or disproof Bédier's suspections. It can warn of the intuition of humans failing the true underlying proportionalities.

[9] Weitzman (1982) has defined birth and death rates of stemmas but gives no information about the distributions of outdegrees in his stemmas or abres réels.

et al. (1964) describe how a scribal practice could have led to a very large number of outdegrees, which are 2 or a multiple of 2.[10] Castellani (1957) outlines the so-called production maximum scenario,[11] where at least in the upper portion of an arbre, the outdegree would coincide with the number of successive generations a manuscript had survived. Bédier himself believed that the authors had wanted their oevres to be read and thus multiple copies of an original would be expectable while Castellani (1957) outlines the opposite. Historical sources such as librarian and convent catalogues, collophones and eye witness reports could additionally contain information on probable numbers of copies. Haugen (2015, p.605) discusses such evidence on the circumstances of copying in Norway and Iceland and concludes that in Iceland "the number of copies of each manuscript may have been rather low" (in opposition to the continental mainland). At least this would exclude larger outdegrees. Different such scenarios might have been at work simultaneously.

Generally, the outdegrees in an arbre can be enumerated as a set of integers. This set can be ordered and viewed as a distribution and it is not unlikely, that it fits some parametricized variant of a known distribution. Whether the historical scenarios are presupposing a rather uniform way of copy distribution or different types of distributions for different types of scenarios (one may think of Maas (1960) and his much read vs. barely read traditions or of the distinction between continental and Nordic traditions Haugen (2015)) is not clearly answerable, but can be assessed in a simulation by adapting the simulations to such parameters.

2.6.2 Distributions

Two basic distributions come to mind most quickly as historically easily interpretable: a normal distribution and an exponential distribution. In considering the outdegrees to be normally distributed, we assume that there is one certain number of copies which is most probable, peaking the others; the more the outdegree diverges negatively or positively the fewer manuscripts with this outdegree will be found. This could translate into a hypothetical historical projection where many manuscripts of a tradition were copied and wore off at similar rates. The

[10] In this case, the scribes copy halves of codices and then interchange the halves in order to minimize the time they would otherwise have to wait for the other to finish copying.

[11] At each generation, all extant manuscripts in the tree are being copied. The number of manuscripts per generation would grow rapidly.

simulation would profit from an investigation of small and larger estimations of the mean corresponding to smaller and larger traditions or such read by few and such read by many.

Assuming an exponential distribution, things become more hierarchical. For instance, a powerful organization could declare some of the manuscripts authoritative which would lead to those prestigeous manuscripts being copied many times more than others.

As a special case of a power-law distribution, fitting a similar historical projection, we simulate a Zipfian distribution (Zipf, 1949) which has been observed fitting for many natural processes. In that case manuscripts would be ranked (prestige) and the numbers of copies made would exhibit the typical rank frequency relation, where frequency translates (interpretationally) to outdegree.

Other distributions exist, which are in principle interpretable, some very generally applicable. The geometric distribution could be interpreted underlying copy distributions with a look to the number of copies made until manuscript death. Here, each copy would be the "unsuccessful" draw from an urn which has n white (survival) and m black (death) balls. Each time a white ball is drawn, a new copy is made, until the first black ball is drawn and the manuscript vanishes or is in such desolate condition, that it is stored away elsewhere.

Similarly, a hypergeometric and binomial distribution could be interpreted readily, although the interpretation might seem a little less straightforward. There could be a probability for each attempt to copy a manuscript (each time a scribe or patron thinks about /plans to make a copy) to be successful. For both cases, this can entail an initial urn with n successes and m failures. In the hypergeometric case, the probability would change (be conditional), if a manuscript has been copied already or if an unsuccessful attempt has been made (since a once drawn ball is not replaced); consequently the number of attempts is exhaustive, more than $n + m$ draws are not possible. For the binomial case, the probability is the same for each draw and theoretically infinitely many draws can be taken.

The scenarios underlying different distribution vary and can get very complex making many assumptions; for certain types of distributions it is extremely hard to envision an interpretative scenario, which does not preclude them from occurrence. The following simulation tests the more readily interpretable distributions mentioned in this paragraph. Interpretability is taken as an estimate

Distribution	n. of outdeg.	Parameters	Example distribution generated
normal	5	mean = 4, sd=2	2,2,3,3,5
normal	5	mean = 5, sd=1	3,4,5,5,5
normal	20	mean = 4, sd=2	0,2,2,2,2,2,3,4,4,4,4,5,5,5,5,5,6,7,7,8
normal	20	mean = 5, sd=1	3,3,3,4,4,4,5,5,5,5,5,5,5,6,6,6,6,6,7
exponential	5	rate = 0.5	1,1,1,3,9
exponential	5	rate = 0.25	1,1,1,5,9
exponential	20	rate = 0.5	0,0,0,1,1,1,1,1,1,2,2,2,2,3,3,4,4,6,8
exponential	20	rate = 0.25	0,0,1,1,1,2,2,2,3,3,3,3,5,5,5,6,7,7,9,24
zipfian	5	N=5, s=1	1,1,1,2,3
zipfian	5	N=5, s=4	1,1,1,1,2
zipfian	20	N=20, s=1	1,1,1,1,1,2,2,2,2,2,2,3,3,5,5,6,12,13,13,18
zipfian	20	N=20, s=4	1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2
geometric	5	p=0.1	1,8,13,17,19
geometric	5	p=0.5	0,1,1,2,4
geometric	5	p=0.9	0,0,0,0,1
geometric	20	p=0.1	0,1,1,2,2,2,3,5,5,6,8,8,8,9,10,10,11,13,16,21
geometric	20	p=0.5	0,0,0,0,0,0,0,0,1,1,1,1,1,2,2,3,3,3,7
geometric	20	p=0.9	0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,2
hypergeometric	5	m=1,n=9,k=5	0,0,1,1,1
hypergeometric	5	m=5,n=5,k=5	3,3,3,3,3
hypergeometric	5	m=9,n=1,k=5	4,4,4,4,4
hypergeometric	20	m=1,n=9,k=5	0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1
hypergeometric	20	m=5,n=5,k=5	1,1,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,4,4
hypergeometric	20	m=9,n=1,k=5	4,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,5,5
binomial	5	size=10,p=0.1	0,0,1,1,3
binomial	5	size=10,p=0.5	1,5,5,6,6
binomial	5	size=10,p=0.9	8,9,9,10,10
binomial	20	size=10,p=0.1	0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,2,2,3,5
binomial	20	size=10,p=0.5	1,2,3,3,3,3,4,4,4,4,5,5,5,6,6,6,6,7,7,9
binomial	20	size=10,p=0.9	8,8,9,9,9,9,9,9,9,10,10,10,10,10,10,10,10,10,10,10
random	5	max = 14	0,7,8,11,13
random	5	max = 14	0,4,10,10,13
random	20	max = 14	0,1,1,2,3,4,4,5,5,7,7,9,9,9,11,12,12,13,13,14
random	20	max = 14	1,1,1,2,3,5,6,6,6,6,7,7,9,10,10,11,11,11,12,13

Table 2.11: Example random distributions produced for smaller traditions with 5 internodes and larger traditions with 20 internodes.

of larger probabilities to be congruent with historical processes. In order to assess the various scenarios, where different circumstances lead to different conditions governing different copy distributions for different traditions, a simulation can be conducted mixing traditions with various types of distributions. Finally, the distributions themselves can be forwarded to a randomizer, so as to generate purely random distributions of outdegrees in a realistic range of numbers of copies.

For examples of generated distributions, see 2.11.

2.6.3 Monte Carlo Simulation

In order to simulate these distributions, we need large numbers of arbres from which loss can be simulated and an appropriate model of manuscript loss. Then, we can simulate a large number of stemmas (1000) and manuscript loss, whereafter we look at the tree topology and compare it with actual numbers. We can

additionally count the number of bifurcations and root bifurcations with a look to the Bédier-Maas debate. For the simulations, smaller traditions have been approximated, setting the number of non leaf nodes to 5, for larger traditions, this number has been set to 20 and for the simulation of the mixed scenario, a randomizer chose n for every of the 1000 simulated turns. Finally, a simulation has been undertaken, where one distribution of outdegrees was randomly chosen for each turn (for n=5, n=20 and n determined randomly). The parameters for the distributions are approximations of relevant ranges:

1. normal: mean 1 to 9, standard deviation from 1 to 4
2. exponential: slope parameter from 0.1 to 1 in steps of 0.1
3. zipf: slope parameter from 1 to 4
4. geometric: probability of death from 0.1 to 1 in steps of 0.1
5. hypergeometric: 1 to 9 white and black balls, so that their sum equals 10 and 1 to 10 draws
6. binomial: probability from 0.1 to 1 in steps of 0.1
7. random: number of copies from 0 to 14

2.6.4 Loss scenarios

Historical loss does not affect all manuscipts evenly. Canfora (2002, p.92/93) with reference to Strasburger (1977) found private exemplars to be less affected whereas since libraries tend to be burnt in wars, public exemplars suffer loss more easily. Many other factors (climate, authoritativity, etc.) excert influence on manuscript loss most of which have never been made subject to generalizable quantifications and it is questionable if this can ever be done. Herein, we elaborate a simple model of loss using only two basic assumptions. We simulate loss of 73-100%, which is realistic according to Trovato (2014, p.107/108). We assign each node a probability related to its age (the older the more probably lost) and its outdegree (the more copied, the more probably kept in good conditions, the less probably lost). Since aging is considered to be stronger than preservational effort, we square the age dependent parameter. The probability of loss for each node is thus determined through the calculation of

$$l^2 * slow(i) \tag{2.6}$$

where l is the height of the current node i incremented by 1 and slow(i) is the outdegree slowdown function. The outdegree slowdown function is

$$\begin{cases} \frac{1}{o(i)} \text{if } o(i) > 0 \\ 1 \text{else} \end{cases} \tag{2.7}$$

where o(i) is the outdegree of node i. Note, that there is no distinction for nodes with an outdegree of one and leafs. The so-obtained values v are summed for all nodes and then each nodes v divided by that sum is its probability.[12] First we determine a percentage between 73 and 100 percent to be lost by a randomizer, then compute into how many lost nodes that translates using rounding where necessary.

This model of loss produced rather desirable loss probability distributions as is exemplified in Figure 2.1. However, we also use pure randomization of loss with equal probabilities, as has been done before according to Weitzman (1987), and loss applying the aging factor alone without squaring and preservational slow-down.

For the simulation, we use R and Java. First, for simulating normally distributed copying, we generate distributions randmonly drawn from a normal distribution using the R function rnorm.[13]

This distribution is now our distribution of outdegrees in the to be simulated stemma, each value represents one outdegree. Starting with root, we randomly choose an outdegree and add as many children to the actual node as this outdegree. For each of the so-generated children, we draw another outdegree and add as many children, recording them in the next generation. We iterate the process until all outdegrees are applied exactly once. This results in a differing number of leafs and a differing size of the tradition for each simulation. Since medieval traditions were probably not equal in size, the effect of this sampling is not controlled for further. The leafs, which are generated at the end of this process are not counted as zeros for the distribution since we assume, that the coming of the

[12]For the same approach as purely age dependent loss in another context, see Mehler (2011); Mehler et al. (2011).

[13]Since rnorm produces real numbers and negative numbers, we round all values and leave negative values aside. Since this may lead to a distortion of the so-sampled distribution deviating from a normal distribution, we test for the desired shape by a Shapiro-Wilk normality test in R and only keep distributions which have a p-value above 0.05.

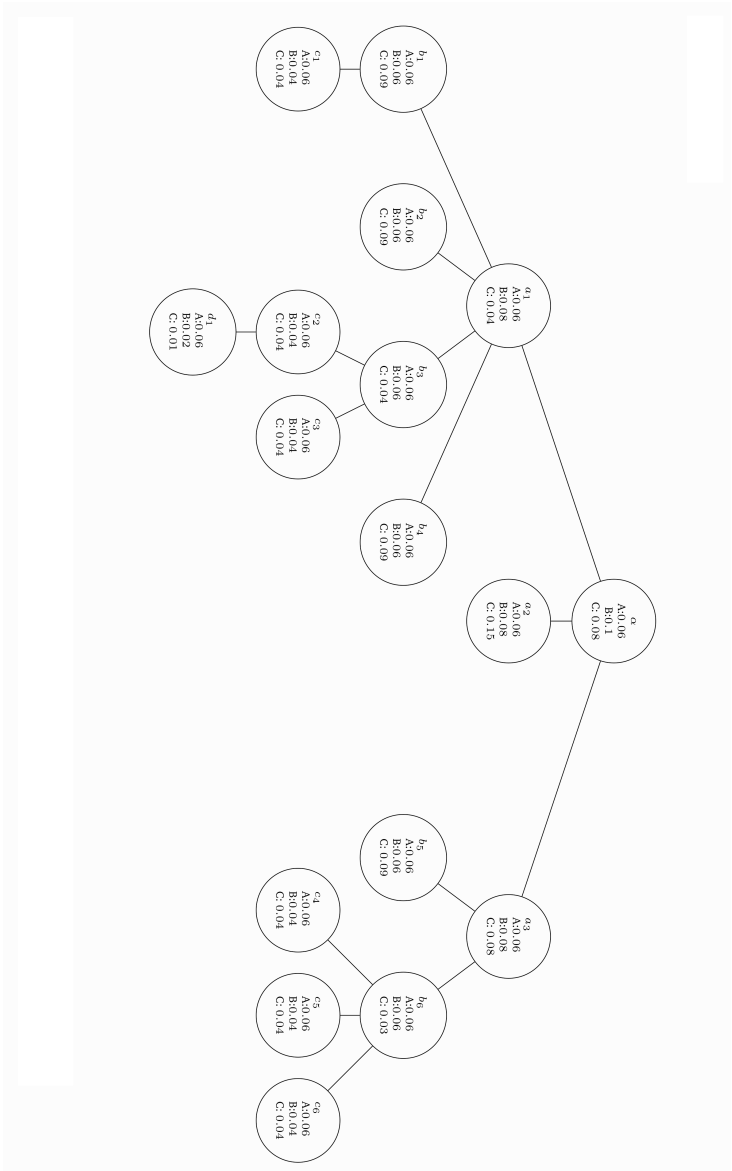


Figure 2.1: Probabilities of loss among nodes of a hypothetical arbre réel (complete manuscript genealogy), A is equiprobable loss, B age dependent (depth dependent) and C age dependent but slowed down loss.

print age more or less apruptly stopped manuscript copying and thus those leafs would have been copied.

For the exponential distribution, we do the same and vary the slope.[14] Likewise we proceed for the other distribution types. We simulate two types of traditions, one family of larger traditions ($n = 20$) and a family of smaller ones ($n = 5$). In the random instance of the simulation, n is decided upon by a randomizer at each of the 1000 turns.

With each tradition, loss is simulated in the three above described ways. We keep all nodes, which are on the path from root to any survivor but delete all other lost nodes. In correspondence to the practice of not positing codici interpositi, we collapse all unifurcation chains of lost nodes (for root this leads to arriving at the archetype). What remains is the true reconstructible stemma (TRS) or 'stemma reale', as called by Timpanaro (2004) in parallel to arbre reel. Since philologists do not have sufficient information (and probably time) for the reconstruction of entirely lost branches, which would be too spurious an endeavour anyway, the stemma we simulated is the maximally faithful reconstruction given our simulated ground truth.

2.7 Results

Illustrative cases have been visualized. Figure 2.2 shows the distributions of furcations of degree 1 to 10 before and after loss in elicited paramtericizations.

In almost all cases regardless of distribution, loss condition and manuscript size, the numbers of furcations after loss showed the same pattern (ignoring leafs): unifurcations became most numerous, followed by bifurcations which were more numerous than trifurcations and so forth. Actually, throughout all distributions in roughly 6% of the trials and only roughly 2% of neighbouring furcations, it happened that not before quadfurcations and generally rather towards the end of a long tail higher order furcations were slightly more numerous than previous ones (such as a 12-furcations having 7, 11-furcations 5 occurrences).

[14] Again, since rexp produces real numbers and negative numbers, we round all values and leave negative values and zeros aside (leafs are determined differently). Since this may lead to a distortion of the so-sampled distribution deviating from an exponential distribution, we test for the desired shape by a Kolmogorov-Smirnoff test in R and only keep distributions which have a p-value above 0.05.

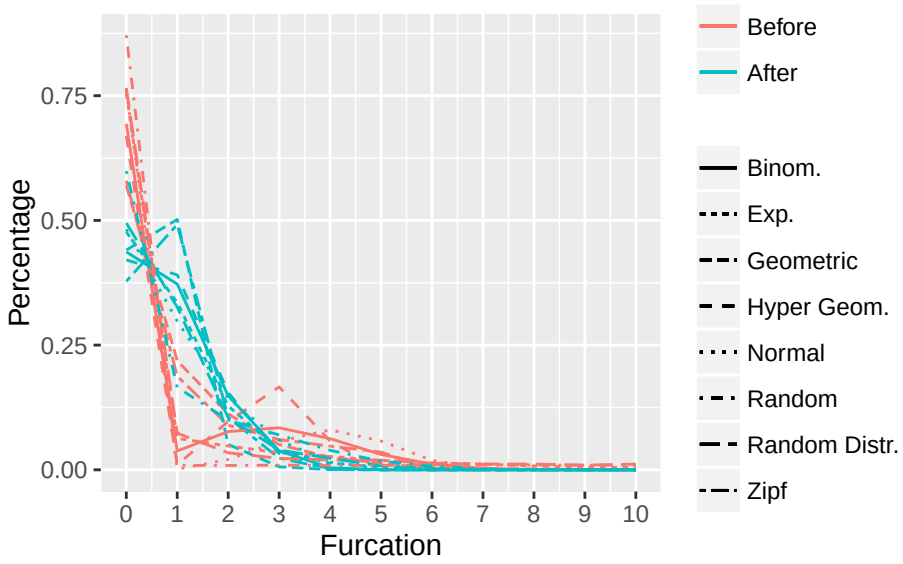


Figure 2.2: The Percentages of k-furcations (x-axis) for distributions, Zipf at $s = 1$, normal with mean 4 and standard deviation 1, geometric with a probability of success of 0.5, binomial with a probability of 0.3 with 10 draws, exponential with a slope parameter of 3, hyper geometric with 4 draws from 7 successes and 3 failures. Tradition sizes determined by number of internodes $n = 5$, age dependent loss.

In 94 of 125 cases, the difference in count was 1 (72 cases) or 2 (22 cases), but in 9 cases the difference was larger or equal to 100 with a maximum of 398. In that last case, the simulation had been a binomial one with random (uniformly distributed) loss, with the chance of success of almost 1. This produced $181,000$ leafs (summed over all 1000 trials) and exactly $20,000$ decafurcations since the outdegrees were generated as the result of 10 draws with the given successrate. There were no other furcations. After loss, the number of nonafurcations was quite large, larger than that of octofurcations for instance though not larger than that of quintfurcations. All stemmas in all trials must have looked the same, a historically unappealing projection. The same may be true for some pther distributions and parameter settings (especially the Hypergeometric case with larger success rates).

For the percentages of root uni- and root bifurcations among all root furcations as well as for bifurcations among all furcations see Figures 2.3,2.4. For the

normal distribution, the overall tree architecture is that of wide trees growing in horizontal rather than vertical direction, the peak occurs at 3 for root bifurcations, meaning, that loss affected frequently either one complete original branch or additional subbranches in case of an archetype. Geometric and zipfian topologies can produce original branches, which are more imbalanced as what concerns their number of children. Even more than in Trovato and Guidi (2004), this may lead to an overwhelming proportion of root unifurcations. The steeper the slope is, the longer the tail of low furcations, which translates into long unifurcation chains in the arbres, which if those occur in the upper portion would lead to many root unifurcations in the stemmas. Meanwhile for the normal distribution a number smaller than but depending on the standard deviation relatively close to the mean should be the expectable root furcation pattern. The binomial and hypergeometric distributions also produce large furcations the larger the probability of success; these will in turn be reduced only to a certain degree by loss leading to ever larger chances of survival for root multifurcations. The geometric case aligns with exponential and zipfian models. A larger probability of manuscript death just as a steeper slope produces slimmer trees with more low outdegrees leading to unifurcations dominating clearly after loss.

The percentages of bifurcations after loss ranged from 0.0 to 0.39 with a mean of 0.2, those for root unifurcations from 0.0 to 0.97 with an average of 0.51 and those of root bifurcations from 0.0 to 0.83 with an average of 0.21. Within the range of tested topologies, there are thus as outlined before such, where root bifurcations are not numerous or even missing. Those cases came from the Hyper-Geometric distribution in more than 70% of the cases, where the fixed number of 5 trials with an initial successrate of 1 to 9 often led to uniform distributions with all outdegrees being 5. Obviously, parameter choices are decisive. One must however not forget, that there is no claim that the simulation covers only realistic scenarios at the same level of being realistic and that due to overlap between the topologies produced by different distributions, average values may become misleading. How realistic which topologies are must be determined by historians and editors working with the traditions. Results would thus show that the possibility exists for large and low amounts of root bi- and unifurcations.

One must take into account, that the number of to be reconstructed nodes can be quite large depending under more on the depth of a tree. Looking at Figure 2.5, while for smaller traditions at a mean of 1 there is supposedly hardly ever

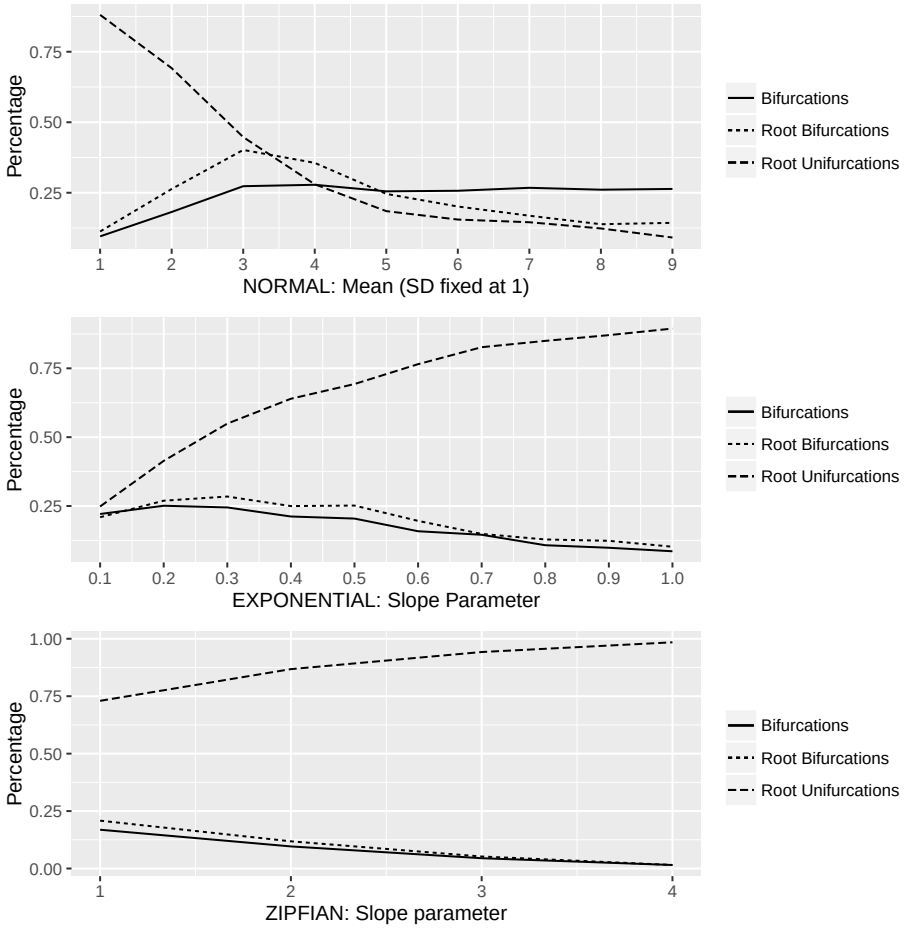


Figure 2.3: The Percentages of root uni- and bifurcations and of bifurcations overall for the normal, exponential and zipfian distribution. Tradition sizes determined by number of internodes $n = 5$, age dependent loss.

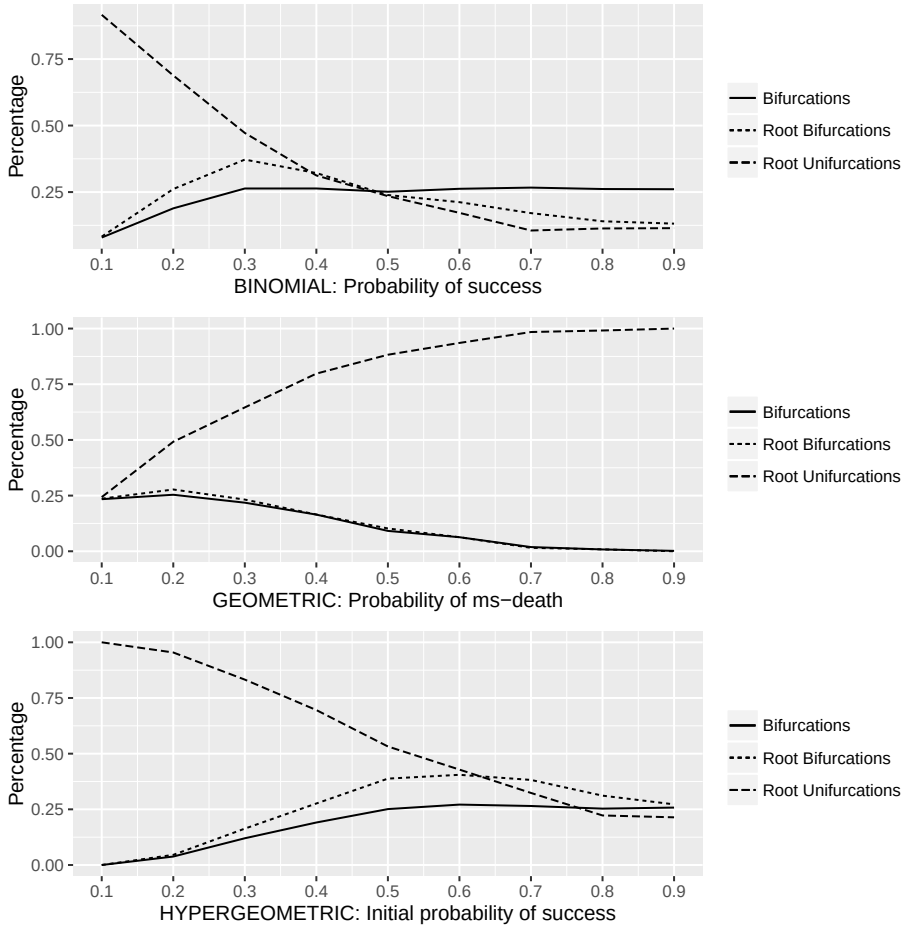


Figure 2.4: The Percentages of root uni- and bifurcations and of bifurcations overall for the binomial, geometric and hypergeometric distribution. Tradition sizes determined by number of internodes $n = 5$, age dependent loss.

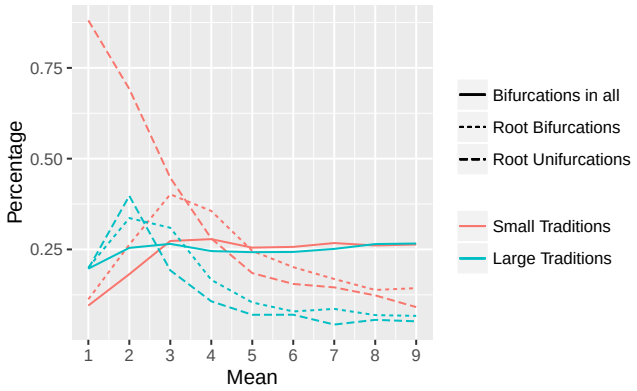


Figure 2.5: The Percentages of root uni- and bifurcations and of bifurcations from normally distributed ground truth for small and large traditions.

any second branch in the tree, for larger traditions, such branching occurs and is preserved, since there are surviving nodes in each of the subbranches. Generally here, for larger traditions original branching is preserved less. This might be correlated with the larger probabilities for imbalance of the original branches.

Confronting the loss conditions, exemplarily consider Figure 2.6. In the random condition, while the overall amount of bifurcations is only slightly differing from that in the other simulations, the root furcations surviving loss are shifted to the right. While for the other conditions for a mean of 3, there are most root bifurcations, for the random condition their peak is with a mean of 4. Thus, in the random loss condition, root furcation patterns are more affected probably through the more probable loss of small branches, where even the leafs have a larger probability of being lost.

2.8 Discussion

While many of the observations must remain speculative not only in the face of lacking historical parameter estimates, one very clear pattern arose. The original distribution was obscured by loss favouring lower order furcations, specifically with unifurcations being most numerous (not looking at leafs). This is not the case for root furcations though. Trovato and Guidi (2004) find large amounts of root unifurcations after loss, especially in cases of heavy decimation. The counts of furcations by Haugen (2015) show the same pattern as found here, except for

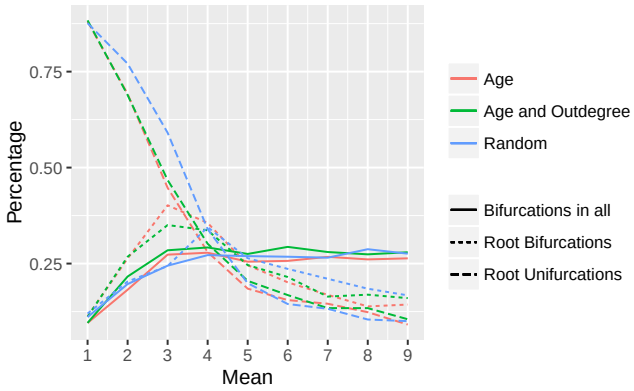


Figure 2.6: The Percentages of root uni- and bifurcations and of bifurcations overall for the normal distribution with the three loss conditions confronted.

a low number of unifurcations. This lower number could be caused by methodological considerations as well as historical factors (such as assymmetry in the original trees). Manuscript texts could tend to be analysed as siblings rather than as in a remote ancestral relationship. This would entail, that the depth of the trees could be underestimated, which is complicated to assess given the practice of contraction of hypothetical nodes. Looking at some stemmas such as the one for *De nuptiis Philologiae et Mercurii* by Martianus Capella proposed in Shanzer (1986), see Figure 2.7, which may well mirror the true historical relationships, the extant witnesses (Latin letters) all occupy leaf positions. While this may often correspond to the historical reality of survival of the youngest manuscripts, another possibility could be methodology inherent. One of the early steps of philological work is in grouping together manuscripts. Not being able to locate that one of them is a remote ancestor (because variation between remote nodes in unifurcation chains may be quite similar to the variation between siblings) could lead to the elimination of many of the original unifurcations (even more so for contaminated traditions).

In one binomial case outlined above, loss led to a burst of a single (deca-) into different furcation patterns. It is not enough to simply think of loss as preserving larger outdegrees less. The topology of a tree paired with stemmatic recontruction practice entails that the higher a furcation ranges (the lower the distance to root), the larger the probabilities that although any one of the siblings is lost it will have to be reconstructed. At the same time, the higher in the tree we are (if

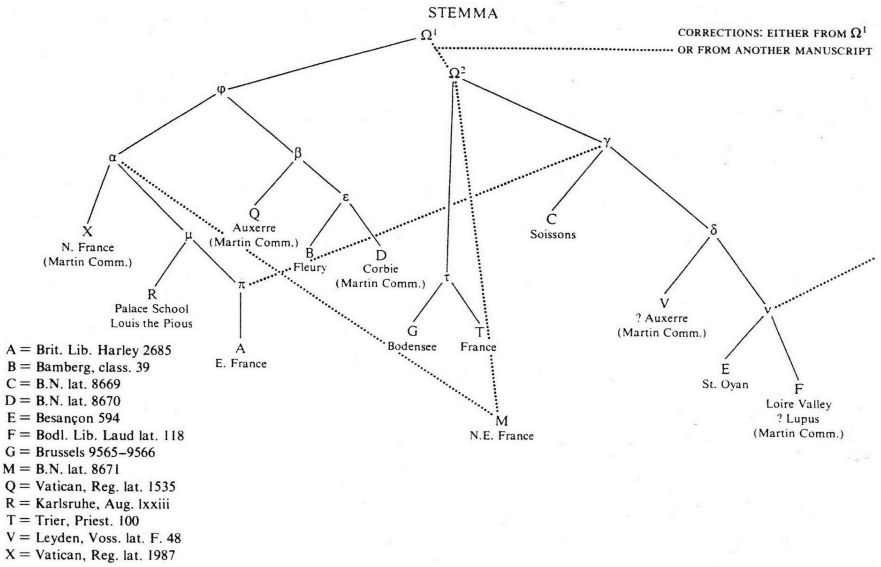


Figure 2.7: A stemma by Shanzer (1986) with extant manuscripts at leaf node positions.

the whole tree is not entirely assymmetric) the fewer furcations exist. For the furcations immediately above leafs this entails, that there is no chance of regaining size through reconstruction once leafs have been lost. Also, unifurcation chains in an arbre ending in a leaf, stand a slim chance to survive since they have few members. Generally, furcations above leafs are affected more directly the more leafs they produce. Thus a manuscript, which has been copied rather for private usages (although this entails larger survival rates) may leave less traces than one which had been copied as often, but for other purposes. Estimates of the proportions of private copies versus monastic ones, may thus be a useful parameter for topology reconstruction. In sum, heavy loss leaving will most likely produce numerous uni- and bifurcations at the lowest level. In the next higher generation, which is less numerous, the effect of loss is slightly more moderate, thus shifted slightly from surviving uni to bifurcations and so forth. As the simulations suggest, the numbers of furcations decrease more rapidly with generation than the number of reconstructed nodes can compensate for.

Finding the best descriptive distribution for the distributions of furcations in the stemmas Haugen (2015) presented, we find, that the exponential distribution fitted the stemmatic outcomes better than the normally distributed variants, see

Distribution	Bibliotheca series	Editiones series
Normality	2	2
Exponentiality	11	9
Both	7	7
None	14	15
Stemmas	34	33

Table 2.12: Tests of distributions of furcations in the collections investigated in Haugen (2015). Tests at significance 0.05, Shapiro-Wilk normality test, Shapiro and Wilk (1965) and the exponentiality test by Kolmogorov-Smirnov computed in R. Applying other tests for additional distributions, the weibull and log-normal distributions fitted the data best in terms of likelihood, suggesting a similar scenario.

Table 2.12.

Haugen (2015) gives the numbers of root bifurcating and root multifurcating stemmas in his two collections (additionally separated into conclusive preliminary stemmata). Numbers are 57 bifurcating to 10 trifurcating ones in one series and 61 root bifurcating, 15 trifurcating one quintfurcating stemma. Thus root bifurcations are clearly very dominating. In the normal, binomial and hypergeometric projections, root unifurcations dominate almost exclusively in the beginning, decrease continuously and are overtaken by root bifurcations at some later point as more frequent root k-furcating pattern. At later points, other root k-furcation patterns must prevail. At the same time, taking the geometric, exponential and zipfian projections, larger root k-furcations dominate smaller ones ever more the steeper the slope or the more probable manuscript death. For smaller traditions, the exponential distribution with a slope parameter of 0.5 (gradient is 0.5) in the age weighted loss (Bibliotheca series) resp. random loss (Editiones series) condition was the model best explaining Haugen (2015)'s data if ignoring unifurcations and leafs (measured by Euclidian distance of the vectors of numbers of furcations).

2.9 Conclusion

It has been shown that independent of other factors the effects of heavy loss overshadowed the original distribution of copies predominantly with the same outcome: the number of unifurcations exceeded that of bifurcations, which exceeded that of trifurcations and so forth. The original copy distributions should therefore rather be deducted from external evidence such as librarian catalogues or eye witness reports. The root furcation pattern may depend crucially on the topology of the arbe. The preponderance of unifurcations and root unifurcations in many settings is, similar as in Trovato and Guidi (2004) an element, which needs further explanation.

Part II

Attempts at Reconstruction

In this part, first the benchmark data sets with known ground truth, which are used for evaluation in this thesis are presented in chapter 3, then chapter 4 presents attempts of stemma reconstruction employing external training data (psycholinguistic letter confusion matrices) and minimum spanning trees.

Chapter 3

Data Sets

3.1 Simulations of traditions

Weitzman (1982, 1987); Flight (1992) all have simulated historical transmission of manuscripts. These simulations are the logical follow-up of scenarios drawn in philological discourse in order to elaborate, visualize and exemplify theoretical and tradition-bound arguments. Like for a theoretical philological argument, it is not necessary to simulate the textual constitution of each and every manuscript in detail to asses certain facts, which would be quite complicated.[1] Instead, for some arguments it is enough to represent agreements of manuscripts, for example as pseudo-DNA, compare also Flight (1992, 1994). In other cases, the simulation of a graph may suffice to asses certain arguments, compare Weitzman (1982, 1987); Hoenen (2016b). In the remainder of the thesis, however, such datasets will be used, which allow for testing algorithms, that could then be applied directly to digitizations of historical texts. Therefore another type of 'simulated data' was used. This type has been created by volunteers copying manuscripts.

To date scholars have produced such publicly available artificial traditions.[2] Artificial means here, that a text has been given to volunteers recently to be copied by them in a randomized setting, while their true stemma has been

[1] Hoenen (2014a) has presented some preliminary experiments on the simulation of textual copying inspired by his supervisor A. Mehler, where he takes a text and simulates a copy process looking at the exchange of letters within words according to under more psycholinguistically determined confusion probabilities.

[2] `https://www.cs.helsinki.fi/u/ttonteri/casc/data.html` and `https://hucompute.org/applications/corpora/`.

recorded. The generation of these is described in Baret et al. (2004), Spencer et al. (2004a), Roos and Heikkilä (2009) and Hoenen (2015a). The artificial traditions share a common trait, that their archetype texts are texts taken somewhere out of a vibrant copy history. Subchapter 3.1.5 represents an elaborated and revised version of the publication Hoenen (2015a).

3.1.1 Artificial Traditions

The working set used in this thesis consists of four artificial traditions. Artificial traditions are such corpora of digitized manual copies going back to an archetype, which have been produced in a scientific experiment. To date three artificial traditions have been created and continue to be used in stemmatological research as "artificial benchmark data sets", Roos and Heikkilä (2009). In the course of this thesis, a fourth one has been compiled, the first in another writing system: the TASCFE corpus, publicly available. There is a fifth artificial data set, *Julies Caesar*[3] which has been used scarcely in the literature, compare Robinson (2015), but the correct stemma is not available online in the required format and furthermore, there was to date of publication no publication evaluating it. The four corpora are:

1. Parzival (English)
2. Notre besoin (French)
3. Heinrichi (Finnish)
4. TASCFE (Farsi).

The data was gratefully distributed by the organizers of the "computer-assisted stemmatology challenge" 2009, the results of which have been published alongside further analyses by Roos and Heikkilä (2009) available online.[4] Parzival has 21 manuscripts and the alignment has 855 lines, Notre Besoin features 13 manuscripts of 1035 lines and Heinrichi 64 (37 after simulation of loss) manuscripts of 1208 words. Since however, for the Parzival and Heinrichi traditions, the complete set of manuscripts is not directly available (but the challenge sets includes only a subset of the manuscripts, which was done to simulate a situation of historical loss), the author received the datasets from personal communication (with Prof. Dr. T. Roos and associates). These datasets have slightly

[3] https://phylomemetic.wordpress.com/2015/02/12/artificial-textual-tradition-julies-caesar/

[4] http://www.cs.helsinki.fi/u/ttonteri/casc/index.html

different numbers due to tokenization differences, where each punctuation character had received its own row. The TASCFE has 54 manuscripts of 137 words.[5] Particular information on generation of the traditions and the data contained as well as the *vorlage* texts can be found in Spencer et al. (2004a), Baret et al. (2004), Roos and Heikkilä (2009) and Hoenen (2015a). The authors provide fully aligned tables of the data.

3.1.2 The Artificial Traditions in Detail

The following description is based on the original publications as well as Andrews (2014); Andrews and Macé (2013). The first data set, a French translation of a Swedish work, is entitled *Notre besoin de consolation est impossible à rassassier.*. The archetype had been dictated to a Dutch-native speaker, whilst the language of the tradition is French. This archetype has then been corrected by a French native, who did not use other materials alongside the manuscript. These givens are similar to ancient authors first producing a text, apart from the variable language, see Reynolds and Wilson (2013). Quoting from Andrews (2014, pp.526)

> This tradition was created for the comparison of several different methods for computational stemmatology (Baret et al., 2004); this experiment is the only one to date for which the results of 'classical', non-computational methods of stemma creation were included alongside the computational versions. In the published experiment, one of the two non-computational methods came closest to reproducing the true stemma, although the computational methods (none of which are able to infer the sort of contamination that was present in the true stemma) were assessed on the basis of the raw output of the algorithm, without any interpretative intervention. The authors note that 'most philologists' were easily able to observe the shift of exemplar from the collationa lone, which suggests that, had the computational methods been subject to interpretation, the outcome may well have been different.
>
> The second artificial tradition is an English translation of a portion

[5] Due to insertions and deletions the single manuscripts must not have this exact number of words, outlined below.

> of the medieval German epic poem *Parzival*. This text is 834 words long, copied by an unknown number of volunteer scribes, [in personal communication with one of the authors, I was told the number should equal the number of manuscripts] and is available in 16 versions. [all 21 have been made available to me] [...] Although the text is a little shorter than *Notre besoin*, the somewhat archaic language gave rise to more frequent variation within copies. [which had been intended] [...] The third artificial tradition is a text in Old Finnish, *Piispa Henrikin Surmavirsi*. This text, also known as the Heinrichi tradition, is roughly 1200 words long and was copied by 17 volunteer scribes. Sixty-seven copies were made, of which 47 were made available for analysis. [to the author again all 67 were available] The creators of this tradition wished to simulate medieval copying conditions as far as possible in the modern era; in service to that goal they chose a text in an archaic language that was only imperfectly known to most of their scribes (speakers of the modern language), they produced a far larger set of manuscript texts, they had some of the volunteers make two or three copies from different exemplars, and several of the copies were mutilated after the volunteer work of copying had finished to simulate damage to manuscripts [...]

3.1.3 Studies on the Artificial Traditions

The artificial traditions have been published alongside corresponding studies. For Notre Besoin, Parzival and Heinrichi, the studies involved an evaluation of methods of stemma reconstruction. For Parzival, the authors looked at the distribution of error rates among locations and found a gamma distribution to fit best. They state (p. 509) that "There were no changes at most locations, but the readings at a few locations changed many times".

Computationally, different types of algorithms have been applied to reconstruct the copy history of the artificial traditions and evaluated. Not all of the results are necessarily directly comparable due to a number of reasons. In some cases, the algorithms have been applied to the reduced data sets (that is only for a subsample of manuscripts the methods have been applied, while the rest of manuscripts have been held back as historically lost), compare Roos and Heikkilä

(2009) against Spencer et al. (2004a). The second reason for comparability issues that can arise is preprocessing. Similar to philological judgement, some features of the original texts such as punctuation may be excluded from the dataset before stemmatological algorithms are being applied consistent with philological practice. Finally, different ways of evaluation may be used, Spencer et al. (2004a) used partition distance, Penny and Hendy (1985) as used in phylogenetics, while Roos and Heikkilä (2009) define and use Average Sign Distance (ASD). Finally, some methods restrict themselves to produce bifurcating trees, typically with the extant manuscripts as leaf nodes, see RHM in Roos and Heikkilä (2009). While a large number of applied algorithms is phylogenetic (especially Parsimony based approaches, Split Decomposition and Neighbour Joining) other approaches have been tested, see for instance Roos and Heikkilä (2009). Especially Compression based methods have been popular, compare Roos et al. (2006); Roos and Heikkilä (2009); Merivuori and Roos (2009); Lai and O'Sullivan (2010); Lai et al. (2010); Lai (2012), Roos and Zou (2011a) use expectation maximization.

Andrews (2014) tested human (philological) judgement on the genealogical significance of textual variation in the Parzival, Notre Besoin and Heinrichi traditions finding that human judgement can lead into poor results.

Andrews and Macé (2013) conduct a qualitative anaylsis, where they define 8 classes of variation and look at the distribution of genealogical variation in the three traditions Parzival, Notre Besoin and Heinrichi. They find different profiles, which they partly connect to language. Punctuation is being identified as potentially non genealogical. An overall analysis of the three traditions leads to a further characterisation. For Parzival, there is a (p.513) "relatively high incidence of 'lexical' variation" caused by misreadings of letter shapes or word shapes. For Notre besoin primarily spelling and punctuation differed and again letter and word shape similarity lead to lexical variation. Notre besoin, according to Andrews and Macé (2013, p.514) contains "far too little variation to reflect accurately the features of a medieval tradition." For Heinrichi "goes perhaps too far the other way. [...] the vast majority of variation conflicts with the stemma somehow." For the distribution, see Figure 3.1.

3.1.4 Distributions and Qualitative Error Analyses

As Andrews and Macé (2013, p.513) have mentioned, their "classification system would have benefitted from a catergory to indicate similarity of word or

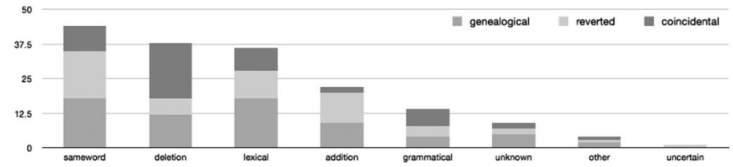


Fig. 10 Breakdown of variation by relationship type in Parzival.

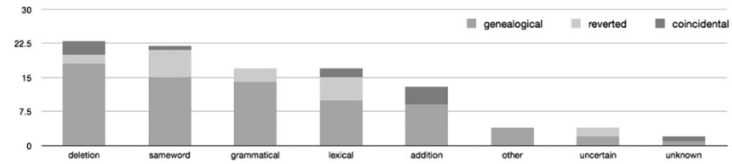


Fig. 11 Breakdown of variation by relationship type in Notre besoin.

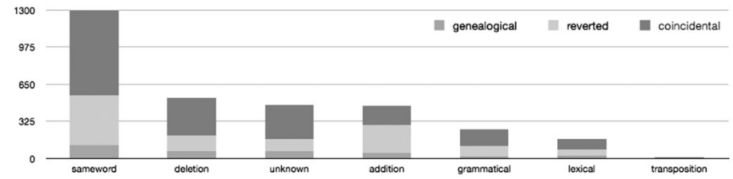


Fig. 12 Breakdown of variation by relationship type in Heinrichi.

Figure 3.1: From Andrews and Macé (2013): variation types in the artificial traditions.

Tradition	lexeme	grammar	grapheme	geminate and doubling	case	unconcentrated	overall
Parzival	101	49	55	4	20	17	217
Notre Besoin	10	11	21	3	3	0	40

Table 3.1: Errors by class in two artificial traditions.

letter shape." In order to assess this, a short analysis of letter shapes has been undertaken along an alternative classification approach. Looking at the letter pairs differing in two of the traditions (Parzival, Notre Besoin),[6] qualitative error classes have been defined and their incidence counted. It is assumed that neither of the set-ups of the artificial traditions had involved any instruction to purposeful changes. For each category, all variants are counted, ignoring positions where the archetype (the root manuscript) had been extended or where one of the two variants was a blank. This assumes that each variant has arisen only once and ignores whole word insertions and deletions, which is of course a simplification. Also, word separation errors were hereby discarded. This was done with the focus on exact letter unit transitions. Punctuation has been excluded from the analysis. The errors were then classified manually and if applicable corresponding confusion pairs were extracted. The classes were

- errors leading to a lexeme alternation, such as <honour> to <horror>
- errors with grammatical alternation, such as <woman's> to <women's>
- graphemic variants, such as <vicissitudes> to <viciscitudes> [7]
- geminate errors and vowel doubling, such as <dispossession> to <dispossesion>
- upper case - lower case alternations, such as <Hell> to <hell>
- errors most probably caused by unconcentratedness with non words as output, such as <crnquering> from <conquering>[8]

The results are summarized in Table 3.1. Sometimes the cases were "borderline" cases and if so were counted once for all possible classes. Especially geminate errors always counted as graphemic errors as well.

[6]Unfortunately, due to a lack of competence in Finnish the author was not able to perform the analysis for Finnish. For Farsi the results are few in number and discussed in the following section.

[7]For instance, in <scene> , <sc> represents /s/.

[8]Digitization errors were not considered, since the thorough compilation of the artificial traditions, their limited size and the numerous studies using them did all not refer to any digitization errors either.

As one can see, *Parzival* and *Notre Besoin* differ in the distribution of the error classes and so do the respective languages and orthographies for English and French. The copyists of French made relatively fewer mistakes overall, which would be consistent with the finding that the English orthography is deeper than the French one. [9] Whilst the class of errors where the graphemic representation of the same phoneme changes and produces variant pairs such as <their> and <there>, which in this case is problematic because it also means a lexeme change, is only the second most frequent error class for English, in French it is the largest group of errors. An example would be <connait> and <connaît> . Grammatical errors were of a slightly different quality in both languages. In French, the many silent letters at the end of words, which for instance encode *person* were sometimes confused as in <ressens> and <ressent>[10] , whereas in English, the final s, regardless of the grammatical function as *3rd person marker*, *genitive marker* or *plural marker* contributed a major part to those variants. [11] Geminates and case variants play a marginal but constant role and occur in both traditions at low frequencies. The most striking feature of the *Parzival* tradition is the high number of lexical variants, which was also found by Andrews and Macé (2013). More importantly, these do also occur in French and they can in the majority of cases be traced back to visual or motoric similarities most probably conditioned by context. It has to be stressed, that the data are very few for far reaching conclusions. Nevertheless, the error class of lexical errors triggered by visual similarities is seemingly a major ingredient of scribal variation in the present data.

[9]The concept of orthographic depth (Katz and Frost, 1992) refers, for instance to the degree to which a writing system uses grapheme-to-phoneme or phoneme-to-grapheme conversion rules that are not based on a one to one representation. Additionally consistency of representation is a factor for orthographic depth. For instance, the representation of /ʃ/ by <sh> employs two letters for one phoneme, this representation is not absolutely consistent, since different orthographic units, such as <ch> as in <chef> can also represent the same phoneme. In this sense, English is a rather deep orthography.

[10]Many of these errors had been introduced during the only dication copy by a French to a Dutch native.

[11]Approximately as many additions of <s> as deletions occurred. This might be interesting or deserve further investigations as it raises the question whether this could be caused by the grammatical development of English towards an even more isolating type, where the speakers would have more difficulty in determining the correct application of <s> , as in the more progressive spoken language, the <s> as a marker has already gotten out of use and confronted with a demanding task such as copying the missing habituation surfaces.

Individual units of corresponding letter tuples have been focussed. For convenience and consistency, only lines of the alignment were considered, which had two variants, the biggest class. Also, a more detailed analysis with respect to established error categories such as *haplography*[12] or *dittography*[13] was omitted to maintain a clear focus on the functional consequences of the transitions. Transposed pairs such as <motley>:<motlely> have been excluded, since in this example it was not clear, which of the <l>s had been inserted. This yielded 54 pairs for the *Parzival* and 10 pairs for *Notre Besoin*. Herein, a search for potentially visually triggered variants was conducted. 41 of 54 for the Parzival were estimated to be caused by visual confusion/unclear writing, where a miscopying of one or more letters from one or more different letters took place. Of the 54 wordpairs 47 were "licensed by the context", that is they did not change the grammatical and semantic validity of the sentence. 8 out of the contextually licensed words were not potentially visually triggered in the *Parzival*. For illustration, the examples from Table 3.2 have been provided. In conclusion, such errors and their genesis are a vital part of the deviations and deserve wider attention if this shows to be a more generalizable distribution for English manuscripts in general. The French data was too sparse to report statistical information and in addition the first copy after the archetype had been a dictation to a non-native, where silent letters might play a crucial role.[14] However, potentially visually motivated confusions like <au> and <du> were present that were contextually licensed ones.[15]

[12]This refers to the accidental omission of one instance of a repeated segment, see Roelli and Macé (2015).

[13]This refers to the accidental repetition of a segment, see Roelli and Macé (2015).

[14]Considering Senoner (1981), this reflects ancient practice. However, the text itself in the artificial tradition is already one with a transmission history.

[15]A particular subhypothesis arising from visual inspection is, that errors involving the visual confusion of letters are supported by the scribe and the copyist using two different variants of the same letter. That is an <a> might be written with an upper gap as it appears in print font or just like an <o> with a bar attached to the right. The sub hypothesis claims that if a copyist who himself uses the o-variant copies from some scribe who used the gap-variant, then the probability for erroneous copying or misreading is higher, helped possibly additonally by slightly unusual letter shapes resp. those a's which in the continuum of the scribes a's are most deviant from the average. In the case of <au> and <du> this points to those a's of the scribe(writer) which have the most unusual proportion of the bar length to the o-body. At the same time an adjacent <d> which is unusually a-like will through visual priming have a similar effect.

GU pair(s)	tradition	word pair	context
or:au; n:0	Parzival	ordinance:audiance	what an outlandish ordinance!
n:r	Parzival	outlandish:outlardish	what an outlandish ordinance!
iseas:esir(?)	Parzival	disease:desire	sweet balm to women's eyes, yet women's hearts disease!
m:w	Parzival	make:wake	But such a man may yet make merry
tr:k	Parzival	astray:askay	too long or goes astray and
d:l(?) dd:ld(?)	Parzival	odd:old	in one odd corner
t:h(?) to:ho(?)	Parzival	torrid:horrid	How lasting is thin ice in August's torrid sun?
a:d	Notre Besoin	au:du	étaient défendables, doit du moins avoir

Table 3.2: Examples of misreadings in context.

As for consistency in misreadings, one scribe read <c> and <l> as <d> twice from the same *vorlage* in <close> to <does> and <clash> to <dash> . On the other hand, <heat> was produced from <heart> in different manuscripts from different *vorlages*. Once again, there is too few data to conclude anything from this but any one counterexample disqualifies the exclusivity claim of the opposite proposition.

West (1973, pp.23) enumerates some constantly confused letters for Latin as do Reynolds and Wilson (2013, pp.223). In the minuscule script <c> and <l> to <d> occur in both references.

In one case a transition of <where> to <with> made sense in a smaller syntactical frame, but when reading the whole sentence not. *Shame and honour clash where the courage of a steadfast man is motley like the magpie.* Possible scenario for this error: Copying proceeds linearily, for the phrase *Shame and honour clash*, the most probable follow up item might be <with> . In any case of unclear handwriting of the next word that really followed, <where> , starting in the same letter, the scribe might interpret it, as an unclearly written <with> . He might check this assumption, in his mind still being occupied, by reading the next few words. He sees *Shame and honour clash XXX the courage of a steadfast man* and is satisfied with his view, he writes <with> . Over copying the next few words other thoughts may come and go and he forgets about the <with> -question. When he copies <is> which is from where the former interpretation would get ungrammatical, he is not aware anymore (of the sentence as a whole, since he copies chunks) and proceeds, leaving behind an ungrammatical sentence, that

a copyist of his manuscript in turn, if not in possession of another exemplar would most probably try to correct. *Shame and honour clash with the courage of a steadfast man is motley like the magpie.* Driven again by visual misreading, he might assume <is> to be originally an <as> , being one of the least invasive interpretations. *Shame and honour clash with the courage of a steadfast man as motley like the magpie* would result. In reality, although in the *Parzival* tradition the <where> was once replaced by <with> , the subsequent two copies simply kept the ungrammatical structure and left the sentence unchanged. The scenario points to chunking being another important factor in copying.

Generally, in Parzival things got partly corrected, especially examples clearly perceived as errors: duplications of words (*the the* and *and and*). Duplications are probably the one case which gets corrected consistently. Different strategies for illegible and apparently inconsistent variants violating linguistic correctness appeared to be at work. One scribe left out the violatory item instead of reconstructing. Some items were partly corrected, partly left unchanged, even if violatory.

All in all this quantitative assessment points to a large influence of visual confusion of single letters, but also to some more complex processes at work.

3.1.5 TASCFE

The Tehran-Artificial-Shahname-Corpus-with-Frankfurt-Extension (TASCFE) is an artificial corpus in the Farsi language of Persia. For this thesis, this artificial corpus has been created, digitized and analyzed in order to find script-independent properties of the copy process. The TASCFE represents the only artificial corpus in a non-Latin script until publication time, see Hoenen (2015a). The true stemmatic relationships have been recorded during creation. The corpus consist of two parts created on different dates in different locations (Tehran and Frankfurt).

The Production of the Tehran Artificial Shahname Corpus (TASC)

Copying in Tehran was conducted during the first quarter of an hour at the beginning of a lecture on 2 subsequent days (24th and 25th of February 2014). The in sum 23 (3 only copied on one day) students were given the original sheet and were then asked to copy it. It was stated that no further questions will be answered during copying. On the first day, the students were asked to copy in two from the same sheet, which was always a printout and to mark on which side of the sheet they had sat. The versions were provided equally through a random number generator using the java Math.random() function. On the second day again equally distributed, the students were given either the same sheet as the day before in print or another version than the day before in print or a hand-copied exemplar from the day before (which had never been their own, as distributed by the random function, although it was not excluded from happening).

The Production of the Frankfurt Extension (FE)

The Frankfurt Extension of the corpus (FE) was produced on the 30th of May 2014 in Frankfurt am Main, Germany. Participants were 9 native Persians who were all fluent in German. They copied each one handwritten manuscript. The manuscripts that were given to them were chosen at random manually, but with a preference for such manuscripts that had been a copy of a handwritten manuscript. The participants were instructed to copy the manuscripts and not to exchange information among each other.

Provenience and Characterisics

The text is an excerpt from the Persian national epos Shâhnâme (*Book of Kings*). This text was created/written down around the year 1000, the authorship is generally attributed to *Abu l-Qasim Ferdoussi*, who gave it its authoritative lyric shape. Throughout Persia, narrative versions of the text emerged and continue to be performed to date, many of them basing their work on or relating it to Ferdoussis edition, Oliaei (2010). However, as Yamamoto (2003) and Rubanovich (2011) discuss, the text has at least been influenced by oral literature, that is oral transmission. Ultimately the discussion on the sources of Ferdoussi is unresolved.

The language of the used excerpt is *New Persian* but contains some grammatical archaisms, archaic lexical items and archaic orthography. This makes it in some respect similar to the other artificial traditions, for instance, the Parzival, where the authors wanted to induce errors through similar properties of the language of the text.

The text is written in stanzas. Participants copied one stanza of 11 verse lines, each consisting of two half lines, which rhyme. The copied excerpt tells the story of the evil sorcerer Sahhak's reign. The first half of the stanza describes the desolate state of the world under the reign of Sahhak, the second introduces some two pure young sisters, who are brought to his palace where they learn only evil, such as killing.

The text was taken from the 1966/67 edition of *Djalal Khaleghi-Motlagh* of the Shâhnâme, (Ferdoussi, 1967, p.55), where 4 versions were produced corresponding roughly to the data in the critical apparatus.

Alphabet

Since a general discussion on the writing system would be out of place here, a short summary of the main and relevant characteristics is given. Persian uses the Arabic alphabet with a number of letters added to represent sounds only present

in the Persian language *Farsi* such as گ (/g/). The Arabic writing system is a so-called *Abjad* which means short vowels are generally not written, but can be indicated by the so-called vocalization, which is a facultative system of diacritic marks indicating the subsequent short vowel for a letter. The direction of writing is from right to left.

Grapheme Inventory: In Persian there are 32 letters. This is the number of abstract letters, as a concept known for instance from Rayner et al. (1980) in psycholinguistics. It is similar to the dichotomy phon and phoneme, uniting the upper case and lower case variant of a letter (descending from distinct alphabets) into one grapheme. The number of actual basic graphical elments including diarcitic marks however is bigger. Whilst in the Latin alphabet one lower case letter together with one upper case letter constitutes the aforementioned *abstract letter*, in Persian Arabic Script *abstract letters* unite four forms. Simplified those are: the word-initial letter shape, the word-internal letter shape, the word-final letter shape and the letter shape of the letter when isolated. Those shapes can all be different for a particular letter or they can converge in various patterns. There are letters, which the subsequent letter cannot be joined to, even if occurring word internally. [16] This leads to spacing in the middle of words. Groups of letters share basic units, which are then punctuated, that is a *sickle* with a *dot* below is a /b/: ب, a *sickle* with *two dots* above a /t/: ت and so forth. Genuine diacritic marks are e.g. the vocalization or gemination marks indicating the short vowels or gemination.

Errors without strict parallel in the Latin alphabet: From the script inherent characteristics, some types of variation emerge, which are not congruent with those in the Latin alphabet. They have to be considered carefully.

[16] Those are called sun letters in the writing systems terminology.

- Dot misreadings[17]:
  - reading the wrong number of dots: ت instead of ث
  - reading the wrong vertical position of the dots: به بت instead of نه بت
  - attributing the dots to the wrong letter (subsequent or previous or from the line above/below): زحت instead of رخت
  - misreading dots for vocalization or vice versa: ثر instead of تَر
  - misreading dots for a character (the three dot ligature in handwriting for middle miim
  - misreading dots as punctuation
- misreading the base form (similar to Latin alphabet, but also misinterpretation of base form as vocalization, dots or punctuation)
- vocalization misreadings (relating to the wrong letter, misreading the wrong vowel, etc.)
- misreading of ligatures
- misreading word/morpheme boundaries due to spacing inside the word

All in all, the writing system is more "layered" and thus the confusion of these layers is possible for all layers with all others. As such layers the base form strokes, base form inherent dots, punctuation and vocalization marks are meant. The four letter forms complete this set of parameters of confusion.

Purpose

The TASCFE is peculiar not only in being written in another writing system, there is also a transition from print to handwriting in many of the copies. [18] The

[17] There are base forms carrying either one, two or three non- diacritic dots as elements of the basic letter grapheme.

[18] Although this transition has been observed (consider Reynolds and Wilson (2013) or Drogin (1983)), the quantity here is not illustrative.

stemma itself can be analysed as four independent trees of a depth no larger than three generations. [19] Copying from print is a dimension of comparison, which can shed light on the question, whether participants copy differently when copying from print font than from handwriting and how these differences manifest.

The corpus has been manually aligned. The 4 different versions, which constitute different independent stemmatic trees stem however from the same critical edition, where it is not clear if there is a true stemmatic relationship between them if they all go back to a common ancestral manuscript or if they can be considered independent because they go back to different dictation events of the same story by the same or different bards (which seems not overly likely in this case). A collated version with some gaps can be produced and consequently a stemma for the whole corpus is easily assumed.

However, if variation in one of the versions is of an oral type, then there must not necessarily be any edge between the manuscripts adhering to that version and the others. In version 2, the title is a complete sentence, whereas the other versions had only a fraction of a sentence. Usually, one would assume the sentence to have been there a priori and to have been shortened afterwards since more versions carry the shortened version. However, considering Oliaei (2010) who finds that modern oral performances of the Shahname are rather prose based, it seems that the sentence is a trace of an emerging transition in the Persian oral tradition of the Shahname from verse to prose. Consequently, the more poetical headlines which are no sentences would be older. At the same time the pattern seems to be a probable oral or performance related type of variation.

Oral versions can be dictated at different places and given the four print versions, each could reflect a different dictation event, for instance by the same bard. Then the conglomerative stemma would represent an entity, which would be wrong in the sense that the four independent stemmas were never interconnected. In the best case, the stemma would assume four hypearchetypes going back to one archetype. In this sense, the TASCFE corpus can serve as an esti-

[19] The time shortage upon corpus creation stipulated an entry from print, since with every generation, the number of manuscripts starting from one handwritten exemplar or in this case from four, multiplies maximally by a factor of two since only two copyists can simultaneously copy a manuscript. In this way, in the whole corpus $\frac{1}{3}$ of the copies were made from handwriting, which was on the other hand due to a time limit indirectly imposed by politeness in exploiting the voluntary work of course participants.

mate for how an established stemma generating algorithm would behave with oral variation. As shall be seen, in case there is enough difference between the versions and in case, the versions do not converge, the least corrupted common tree has one additional node, joining the four hypearchetypes, which are in fact multiple roots of independent trees. In the worst case however, the stemma will mix up manuscripts belonging to different trees. In the TASCFE, some of the scribes, who were copying, were copying two different versions on the subsequent days. They had a minimal exposure to a type of variation, which a medieval scribe could have encountered throughout his/her life. Medieval copying was however different from the TASCFE situation in many respects.

1. The education of the writers is different. Writers are learning to read and write in modern schools with a very developed didactic system.
2. The spread of literacy is much wider, that is while in medieval and ancient societies the percentage of literates was partly very low, nowadays, it is a skill mastered by the major part of the population, although there might be notable differences between cities and the countryside. Not all participants were originally from Tehran.
3. The phase of literacy is two steps ahead of a purely chirographic age. We have passed through the print age and now entered into the digital age with a so-called secondary orality (Ong (2012)). As Ong and others show, each stage of medial transmission has its own dynamics. In societies with large oral residues, there is a concurrency between performers who memorize from script and bards who acquire an inventory that they exploit dynamically. This concurrency in modern societies has faded or given way to the concurrency of the adherents of digital methods and those of established print age methods, but had profound influence on some practices and attitudes (as to which voluntary changes were understood as permitted) concerning chirographic copying.
4. The body posture, writing utensils, writing environment and instruction or purpose do not match the ones of medieval scribes.
5. Finally, the value of a written product and the authoritativity are profoundly different in modern times.

Nevertheless, a certain universal character, especially in the visual processing of handwriting can be presupposed. This is the basis for interpretation allowing the data to serve as an approximation of a historical corpus, in order to evaluate

Version Number	Number of changed lines	Number of Changes	Number of Words	Note
1 (base)	0	0	107	text of edition, typos such as جز, last line
2	2	ca. 20	122	title changed
3	5	6	107	many lexical substitutions
4	3	3	107	subset of substitutions from version 3

Table 3.3: Size of the four versions.

automatic stemma generation.

Other Characteristics

The text has four different versions according to the critical edition. The size is given for each of the four versions, see Table 3.3.

The changes that took place can be observed in detail on the following pages and contain changes in word order, lexical items, the title and morphemes involved. The changes are oriented at representing the manuscript variation as displayed by the critical edition. However, the amount of difference is rather limited but especially for the title in version 2, awareness of versions is not improbable.

The corpus comprises 50 digitized manuscripts (43 from Tehran, 7 from the Frankfurt extension). In the digitized portion of the corpus there are roughly 50.000 tokens.

The correct stemmatic relationships displayed rather flat structures, the stemma of all print versions one can be seen in Figure 3.2.

Qualitative Evaluation, Error Analysis

Aligned, the four versions have some minor and one major difference. The major difference is the title. There are three different title texts. In one case, it is rather a sentence and in another the wording is different. Qualitative analyses of the artificial corpora have previously focussed on classes of variation, especially Andrews and Macé (2013) study extensively the different classes of variation and their relation to such factors as language or writing system. Since they found the distribution of variation among the classes to vary with for instance language, to minimize this effect, the enumeration is confined to very clear basic classes. Yet still the borders are fuzzy.

1. oral variation
2. lexical

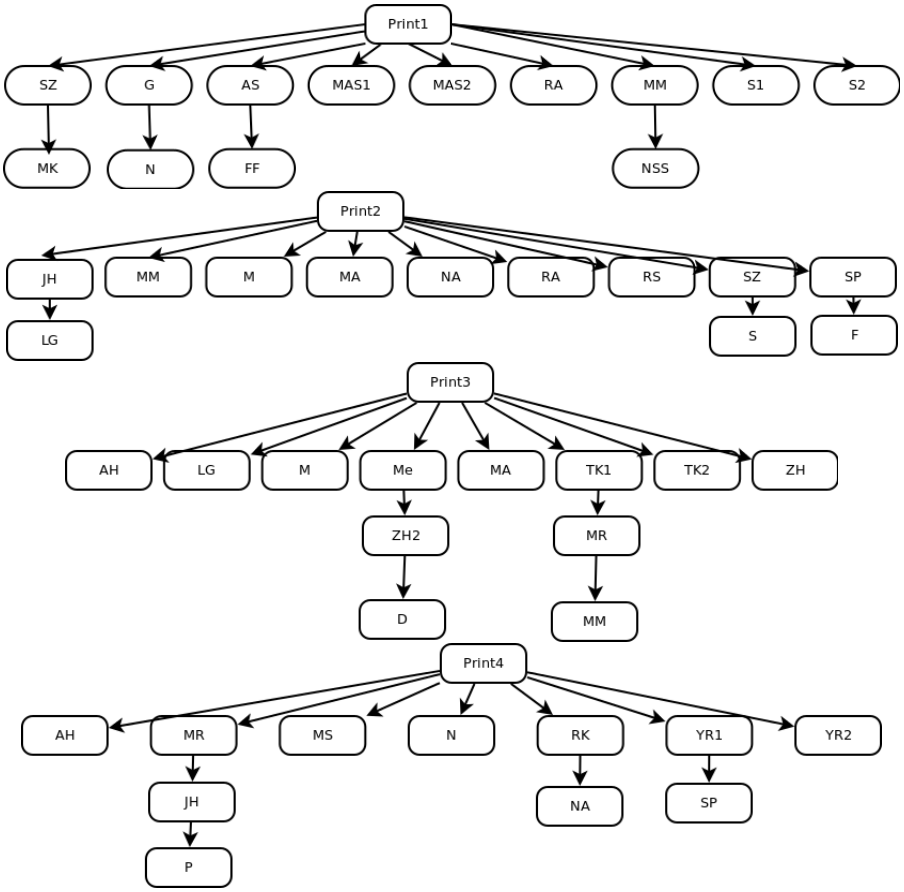


Figure 3.2: Correct stemmas of the versions. Node names have letters which are methodologically non-meaningful labels, they stand for the copyists and refer to some letters in their names.

Version	1st Occ.	2nd Occ.	3rd Occ.
1	Y-N	Y-Y	Y-Y
2	N-Y	Y-Y	Y-Y
3	N-Y	N-Y	Y-Y
4	Y-N	N-Y	Y-Y

Table 3.4: Versions of the name Sahhaak (Tashdiid - Arabic K), occurrence of tashdiid and Arabic k. If a version has both, configuration is Y-Y.

3. morphological
4. transposition or insertion.

Lastly, there were some digitization mistakes, which produced either a non word or unusual/obsolete vocalization marks.

No single copy did not alter at least details of the vorlage. Some scribes were extremely accurate and only differed from vorlage in letter shape variant and punctation.

Versions On the second day, some of the participants were given the same version as on the day before, others were given a different version. Scribes in old times may have encountered different versions of the same story, either as two oral performances dictated or in reading different versions or a mix of both. Since the title in one of the versions was significantly different from the others, the participants must have realized this.

In one case, there was an addition of a *tashdiid*, indicating gemination as in پّ in opposition to پ, which could have been due to the encounter of an unusual low frequency word (name) on the previous day.

The name of the evil sorcerer *Sahhaak* appears 3 times in the text. The first two occurrences are in relative vicinity, the first in the title and the second in the first halfline. The third occurrence is much further down the text.

The *tashdiid* and the arabic k (ك) are distributed as seen in Table 3.4.

Note that only version three was consistent in the first two instances. The patterning was counter intuitive in that usually the first occurrence is vocalized and subsequent ones are then left unvocalized. Here, the opposite was the case in version 2 and 3. Version 1 was constantly giving the tashdiid and version 4 gave it again after a longer distance. Focussing on the tashdiid, one participant

who had seen version 1 the previous day, which had the tashdiid in the title, added the tashdiid on the second day to her copy of the now unvocalized name of print version 2 she saw the day after. Since nobody else either in version 2 or 3 did that or reestablished any tashdiid in the title from the text of the first line, it could be that the inspiration to do this does not stem from the second occurrence of the name in this version but rather from the encounter on the first day. Although this is but a single instance, it points to a possible and probable influence of different versions onto each other independent of transmission. Even more importantly, this deviation is not restricted to corpora including oral variation. This phenomenon is one, which in retrospect could seem very much like contamination, one would have to ask oneself how much of the actual contamination, which can be observed stems from such priming related phenomena and if the vast amount of contamination, that has been detected for some traditions is then rather an indicator of very busy scribes with a tight schedule and mass production rather than massive usage of more than one vorlage. [20]

Copy from print: There were 6.4 deviations per copy process in copies from hand and 7.14 in copies from print. Thus the absolute shape identity of the print letters did not seem to help avoiding copy errors.

Due to the characteristics of the Persian writing systems, the errors which can occur upon copying by hand differ from those made within the Latin alphabet as explained above.

The vocalization patterns in the vorlage were mixed. For unusual words such as اژدَهافَش (dragonking), they were given for the whole word (omitted only those obsolete ones which preceeded a long vowel). They were copied to varying degrees, sometimes, the whole vocalization was left out, sometimes only some of the vocalization signs. The tashdiid was given only upon the first encounter with a word in the text (e.g. on the word جمّشید) or if recurring after a longer time. It was added by some scribes also to the occurrences, where it was not

[20] Theoretically the participant could have been familiar with the name and thus written it this way. Although this points to another source of variation, namely external habit. The same manuscript had the name written in another way later rendering this interpretation unlikely.

present in the vorlage. The first occurrence of the evil sorcerers name in the title in version two didn't have it, while the first in the text did have it.

Similar to vocalization, punctuation was an element that could be inserted for better prosodic clarity, but it didn't have to. The edition featured two commas and one semicolon in the stanzas. There was a slight tendency to leave them out. The only punctuation marks that were inserted have been full stops for the title of the second version, which was a complete sentence.

The final Kaaf ك in our sample of modern Persian handwriting was occurring in the middle/begin form such as in كو, instead of the original end form. Our vorlage had in all titles apart from version two the middle Kaaf at the end of Sahhaks name, while in the first line, the Arabic final Kaaf appeared. Almost every Persian writer substituted the final Arabic kaaf by the middle kaaf, there was only one exception. This implicit change shows, that paleographic practises concerning font shape might not be indicative of manuscript genealogy. Instead, they can be seen as orthographic adaptation.

The particle *ra*, nowadays a definiteness marker/subject marker is present in two chirographic variants, one ligature, one as two strokes. Some writers seem to use both forms. In one specific case, a writer who preferred the separate strokes copied from a writer preferring the ligature. When reading an unusually written ra-ligature (still in the spectrum of principally recognizeable ra-ligatures, presumably more to those, who use the ligature than to those who don't), in a position where it was an archaic oblique marker, the writer hesitated to copy it and instead left a blanc space, relatively larger than the other spaces. Maybe he/she had decided to come back later to this point in text to take a decision and left the larger blank in order for her to visually retrieve the position but then never did so.

As one class of unusual patterns, the participants encountered digitization errors. There were four of them: sister, خواهر, was written with the dot below instead of above, producing the word جواهر, meaning jewel which was a valid replacement in terms of context, since it referred to two virgin girls. However it

was corrected by 4 scribes in 5 instances. Once the token خواهر copied from جواهر had been underlined, a kind of extra notation, not dissimilar to a comment. This indicates the conscious correction.

The next error was an obsolete vocalization of a short a at a high frequency token: دگَر. Especially high frequency tokens are usually not vocalized at all. 30 times the vocalization of the vorlage was omitted regardless of whether the vorlage was print or handwriting.

Participants knew that the presenter in the course was a foreigner and could have attributed digitization errors to this circumstance. There was no explicit instruction to alter or not alter anything, the instructions had been to copy the text. The copyists apparently felt more at ease to correct very obvious mistakes/deviations than to alter unfamiliar tokens. A second case of vocalization was بدند, which occurred twice at the end of each halfline of a line. The second occurrence featured the short vowel vocalization u, بُدند, while the first did not. Nobody added this disambiguating vocaliztation (between بَدند (were bad) and بُدند (were)) and the short u vocalization was left out when copying in 14 cases.

The last token was that of خز, which means 'fur', and was mistyped instead of جز, meaning 'lest'. Participants had a harder time to correct the item. In contrast to جُواهر, the token did not fit into the context. There was an intermediary strategy between plainly correcting the item and leaving it unchanged. This was leaving out one dot. The subsequent scribe would thus find a less likely such as خر, meaning 'donkey'. Not only is the item more unlikely, in retrospective of a subsequent scribe, the probability that a dot has been forgotten seems much higher, than that of a wrong dot being left for doubt. Thus the scribe who changed خز to the version with one dot, which happened three times independently, made a correction in a subsequent copy more likely. It is questionable if this is the objective of a copyist acting this way or if simply violating the context resulted in an inhibition of production of the connected dot. The reason for a harder time of correcting this item was probably the poetic genre paired with the archaic language: the copyists were scanning their internal lexica for an additional ancient meaning of the word خز because they would not want information loss in their copy. The transition from a content word to a preposition

would mean such a loss, whereas this is not the case for خُواهر and جُواهر, which were both content words.

All in all, the digitization mistakes were very informative. Especially for errors in medieval manuscripts, the processes observed herein show, that correction is not correction. Furthermore, similar decisions lead to a certain convergence, which looks like contamination, but which has nothing to do with it.

Some words "attracted" changes more than others because their form was more ancient. Many (orthographical) modernisations could be found. As example may serve: پراگنده and پراکنده The insertion of the possessive marker *ezafe* was likewise optional.

Many of the substitutions had a potentially grammatical character, such as اوریدند and its variants. Another possible interpretation for a subset of these deviations was a "slip of the eye" as is responsible for many types of deviations such as word skips, line skips and page skips. One lineskip was witnessed from after the first word of the second line, which was a preposition (during) to the second word of the third line, which fit after the word. Additionally there were two word skips of a different kind. First, the last word of the first line was substituted by the very similar looking last word of the second line both semantically context appropriate. The second was the second word of the second line being substituted by the second word of the third line, coinciding with the fourth word of the second line. The resulting sentence was repetitive but gramatically correct. In one case, the name شهرناز was substituted by the noun شهریار, meaning king. Although as the last word of the half line it occurred in, the rhyme was alienated by شهریار, the preceeding halfline mentions a crown. Since there is semantic priming, since furthermore, the name Shahrinaz only occurs once in the text whereas Shehryar was already mentioned in the first line, it seems more plausible to postulate a "slip of the mind" or a premature end of the reading process (reading Shehr, the priming activation was already enough to excite Shehryar so much, that reading was interrupted and copying began), than a line skip of 7

lines. The copy was done from print.

Finally, there were some mistakes, which could have easily resulted from a lack of concentration, such as writing a ک for a گ in a place where it produced a non word. Generally the number of deviations was very small.

As explained above, a space character frequently occurs within words. In the print used and in printed books in general, the spaces between words are sometimes larger than the ones within the words. This must not be the case for handwritten material. Some of the participants consistently marked word boundaries by using larger spacing, others did not. Consequently word separation errors should be more probable, especially given the greater number of homographs considering unvocalized material. In our data, a word/morpheme separation error براز read as به راز was found coinciding with an orthographic modernisation.

A revealing/appealing example: Looking at the exmple in Figure 3.3, the following hypothesis seems to be consistent with the data point:

> *If a token's chirographic gestalt allows for misreading of a token which fits better into the context (which may or may not correlate with that token having a higher general frequency in the language) it will be misread.*

Additionally, if a writer copies a variant of a letterform/ligature which he is not familliar with, misreading is more probable.

In the example, both افسر and its substitute اختر fit into the context and ultimately the unusual form of the letter ف led to a confusion. This entails that to some degree the distance of two tokens depends on their surrounding context, *dock* and *dog* may be more confusable in the immediate contextual setting of a *harbour* -to speak metaphorically- than they would be in an *city*. Such and similar hypothesis are closely connected to a body of research looking at the

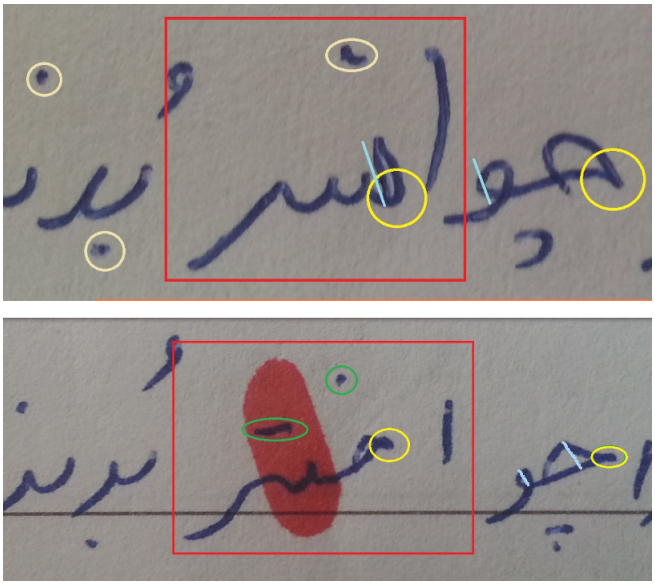


Figure 3.3: The unusual lettershape miscopying of افسر to اختر in red rectangles. The miscopying was caused by certain features marked. The ف in افسر above is not as rounded or circular as usual and larger than expected (blue stroke). Additionally the single dot of ف is accidentally wider than usual (skin-colored). Yet, there is clearly only one dot in the red rectangle above, whereas two (green) below. Furthermore the س has clearly two hooks, whereas the corresponding letter in the copy marked with the red pen has only one resulting hook. Finally, the lower right of the ف is rounded, whereas the same writer clearly shows the typical edgyness in the preceding tokens first letter (yellow). Although the miscopying was induced by the unusual ف shape, including base form and dots, the transitional item was most importantly licensed by the context.

influence of context onto reading, see for instance (Rayner et al., 2012, chapter context experiments, paradigms) and sources mentioned therein. One effect is that contextually primed targets appear to have quicker lexical access times. An interesting aspect is the interaction of script and language specific confusability patterns and general, rather universal cognitive processes active in reading.

Although the TASCFE in number of tokens is tiny in comparison to the other artificial corpora, many deviative events happened allowing for abstractions. The next section contrasts the patterns of the other artificial traditions.

Comparison to the error patterns of the other artificial traditions: Some of the patterns were rather writing system specific, such as the copy of vocalization characters; thus more research is needed and larger artificial traditions at least on all types of writing systems (especially Chinese logographic writing differs much) are an immediate desideratum.

The types of errors which occur are similar, their relative frequencies are not. There are **non genealogical** changes (occurring multiply and independently) in the TASCFE corpus more than in any other artificial corpus, because the vocalization and punctuation was handled individually. In Arabic writing, those aspects are considered to be a more or less free choice of each scribe, depending on the intended audience and other factors such as holiness of the text. This is similar not only to other scripts applying vocalization but, for instance to the application of Japanese Furigana and writing direction in Japanese and therefore a phenomenon that has to be accounted for when dealing with stemmatology in a cross linguistic sense. In Latin based writing systems, the phenomenon might be comparable with to date accepted alternative spellings, for instance in German *Panter* and *Panther*. Non genealogical variation should thus be considered carefully.

3.1.6 Summary and Conclusion

In this chapter, we have presented the data sets which will be used for evaluation in the rest of the work. Two of them have been analysed extensively regarding their errors by recalling such results from the literature and by conducting another analysis. Then, the generation of a new test dataset has been described, where especially writing system specific errors have been focussed. Another peculiarity has been the presence of versions. Being small (number of alignment

positions) and having special characteristics (writing system, versions, transition from print to handwriting), this dataset is computationally more challenging than the others and could for instance serve as a robustness testset.

Chapter 4

Stemma Generation

4.1 Weighting in Stemmatology

In classical manual stemmatology, philologists largely try to determine certain significant types of variation they believe to be genealogically informative (Andrews and Macé, 2013). Without recourse to another version of the text it is rather impossible to recover the content of a lost line which a copyist had accidentally skipped when copying the vorlage. Thus lineskips count as one of the most famous examples of a genealogically significant error whereas a difference in punctuation could simply stem from the introduction of punctuation to different copies at different places. Against this background, the finding of Spencer et al. (2004b) that weighting in their scenario made few differences implies the question for a more elaborated analysis of this phenomenon which this subchapter, which has been almost verbatim published as Hoenen (2018) tries to do – at least in part by proposing and analysing the effects of a new kind of weighting.

4.2 Multi Modal Distance

Stemma generation can be understood as a task where an original manuscript M gets copied and copies – due to the manual mode of copying – vary from each other and from M. Copies $M_1, .., M_k$ which survive historical loss serve as input to a mapping process estimating a directed acyclic graph (tree) which is the most likely representation of their copy history. One can first tokenize and align the texts of $M_1, .., M_k$ and then produce a pairwise distance matrix

between them. From this, one can finally derive a tree with various methods, for instance Neighbor-Joining (NJ) (Saitou and Nei, 1987). For computing those matrices, previous research has applied unweighted approaches to token similarity (implicitly interpreting each token pair as a binary observation: identical or different), see Mooney et al. (2003). The effects of weighting have then been investigated and Spencer et al. (2004b) found them to be small in their (not necessarily all) scenario(s). The present approach goes beyond the token level and instead of a binary comparison uses a distance model on the basis of psycholinguistically gained distance matrices of letters in three modalities: vision, audition and motorics. Results indicate that this type of weighting have positive effects on stemma generation.

We primarily target stemmata for closed traditions (Pasquali and Pieraccioni, 1952) that have no multiple originals, as is probable for orally transmitted epics (Lord, 1960). Hoenen (2017) has attempted to reconcile tree and network perspectives on stemmata. We describe a new method which uses external data in the form of psycholinguistically generated letter and phoneme distance matrices in order to a) generate and evaluate stemmata and b) assess how large the influence of low level perceptual processes is. From the alignments of the artificial traditions, pairwise distance matrices of the single manuscripts (texts, nodes) are built. Each manuscript pair is compared tokenwise using some metric resulting in an overall distance. This metric can be described as weighted, where the external data serves for determining the weights. Concerning token comparison, philology describes a whole range of types of variation and their implications, see e.g. Roelli and Macé (2015); Andrews and Macé (2013). Philologically motivated classification has been used for weighting token pairs upon distance computation. Categories such as "Word variant, changes meaning" or "Word change affecting rhyme" Mooney et al. (2003, p.287) have been applied. A stemmatologically relevant distinction and driving force behind the will to weight variants is that between genealogically informative and accidental variation (Andrews and Macé, 2013). The implication is that some innovations in the text induced by copying are idiosyncratic and hardly revertable, for instance when some non syntactically crucial word is accidentally left out: *this is a really big challenge* → *this is a big challenge* or when some content word gets replaced by one equally fitting into the context: *the clay dust shimmered* → *the day dust shimmered.* Such

errors imply,[1] that the whole subbranch rooted by the manuscript having the new version at first will have it. In this way the innovation is genealogically informative. That is the information helps us locate the manuscript on the stemmatic tree, whereas other innovations could easily happen independently in different copy processes such as the introduction of punctuation at some point in time or some shift in definiteness *I heard the magpie* → *I heard a magpie*. Often variation can be multicausally explained and is analyzed on a case-by-case basis. One process which could be responsible for both kinds of innovations is the confusion of letters. In philological discussions on the complex processes which can lead to variation, the confusion of letters such as <cl> with <d> has been discussed early on, probably already in antiquity (Vanek, 2007, p.276). Reynolds and Wilson (2013, p.222) identify conscious and inadvertent processes as underlying such and other processes.

In this section, we try to determine how well the true stemmatic trees for artificial benchmark datasets in stemmatology (gold standards) can be approximated from external data on the confusion of letters. As Spencer et al. (2004b), we use manually provided alignments, derive pairwise distance matrices and then use the Neighbor-Joining (NJ) algorithm for stemma generation from the distance matrix. In computing the distance matrix of pairwise variant text distances, we compare each position of the alignments and implement three metrics, the simple binary (same or different variant?) Hamming distance (Hamming, 1950), the Levenshtein distance (Levenshtein, 1965) and the weighted Levenshtein distance.[2] For weighting, we do not consider philological classes of variation but distance matrices from psycholinguistic research on letter distances. These have been gained in experimental set-ups and do thus suffer less from a weighting bias introduced through subjectivity as mentioned in Spencer et al. (2004b). Comparing stemma generation with philologically inspired weighting against unweighted stemma generation (Hamming distance), Spencer et al. (2004b) found no crucial

[1] Terminologically, there are some slightly differing terms which imply similar things: variant, innovation, error, change, alteration. Since 'error' implies a knowledge of the correct form, the term can sometimes lead to controversies. Here, we use all terms quasi-interchangeably.

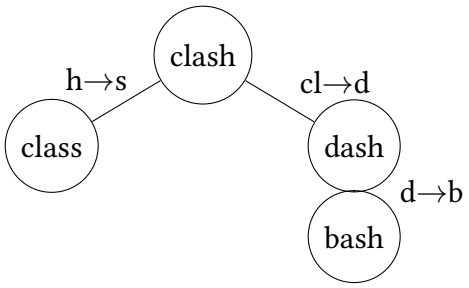
[2] Weights for transpositions are not immediately derivable from psycholinguistic letter confusion matrices. Additionally, there are long distance transpositions or transpositions of vowels of adjacent syllables which would require some additional linguistically carefully modelled distance. The java library debatty `info.debatty.java.stringsimilarity` was used for implementation of the weights.

differences in the resulting stemmata for their data set but stated (p. 236) that 'different weightings could lead to completely different stemmata' concluding (p.238) that 'Determining appropriate weightings in these cases is an open problem'. The main aim of this section is to assess a part of this problem through using external data for weighting.

For evaluation, we use the Parzival PRZ (English), Notre Besoin NB (French) and Heinrichi HR (Finnish) (Baret et al., 2004; Spencer et al., 2004a; Roos and Heikkilä, 2009) both in their entirety. TASCFE has a) multiple roots, b) is written in the Arabic writing system of Persian usage and c) there are considerably less psycholinguistic resources for this constellation. Additionally the text is rather short. Consequently, results are only briefly summarized below. We include the PRZ_loss challenge data set (17 ms) for comparison. From a machine learning perspective, these data sets are quite small and from a historical perspective, they may not represent but a tiny fraction of possible scenarios. Results are thus to be taken with utter caution. Nevertheless, these are the only data in the field for which an indisputable gold standard exists.

4.2.1 Method and Model

Of all pairwise manuscript comparisons, in large numbers of cases both manuscripts do not share an edge in the true stemma. Hence, those comparisons will include word pairs which stem from remotely distant manuscripts on the stemmatic tree. On each edge on the shortest path between them some event(s) may have happened with the implication that one is most often looking at variation reflecting more than one copy step and a back and forth of directionality. This is unfortunate but unavoidable if one doesn't know the true relations in advance. To illustrate, Figure 4.1 gives a small toy example of a tradition, where each manuscript contains only one word and where thus the comparison of the concurrent word pairs correspond to manuscript pair comparisons. Looking at Figure 4.1, when comparing and aligning words, all pairs are different in terms of a binary classification. Using the Levenshtein distance, token pair c gets the same distance as pair b, but only b corresponds to an edge. Counting differing alignment positions assigns the same distance to b and c as well. Only carefully chosen weights achieve an overall weighting that assigns the three lowest values to the pairs corresponding to edges and the largest to that with the longest path in the true stemma. Weights in the example are set intuitively to mimick



Index	comp	pos	Lev	MMD*	l(path)	EDGE
a	c-l-a-s-**h** c-l-a-s-**s**	1	1	0.4	1	T
b	**cl**-a-s-h **d** -a-s-h	1	2	0.2	1	T
c	**cl**-a-s-h **b** -a-s-h	1	2	0.8	2	
d	**cl**-a-s-**s** **d** -a-s-**h**	2	3	0.2+0.4 0.6	2	
e	**cl**-a-s-**s** **b** -a-s-**h**	2	3	0.8+0.4 1.2	3	
f	**d**-a-s-h **b** -a-s-h	1	1	0.3	1	T

Figure 4.1: An example stemma and all corresponding (text=) word pair comparisons. All word pairs are manually aligned, corresponding comparisons (column comp.) highlighted. The number of such comparisons or positions (column pos) is compared to the Levenshtein distance (lev) and a modally weighted version of it (*Multi Modal Distance, MMD, with one addend for each comparison). Path length (l(path)) between the nodes of a pair and whether this corresponds to an edge serve evaluation. Only MMD achieves an optimal ranking.

confusability of the aligned letter units. Realistic confusability patterns of letters and phonemes have been researched and can be inferred from psycholinguistic experiments, see next section, which brings external data into the model and which might help to avoid overfitting and subjectivity. Another question is how much these linguistically speaking low-level phenomena are responsible for the variation observed in copies.

Model

We operate with a number of observed (survived) manuscript variant texts $\mathbb{M}$, which are arranged in a provided token level alignment A. For each variant text pair $(M_i, M_j), i \neq j$, we sum the weighted Levenshtein distances of all words $M_{i,j_{1..k}}$ (implying different letter level alignments) according to the different weighting schemas of the modalities and then weight again each modality with a linear factor.

$$\begin{aligned} &\Delta(M_{i_k}, M_{j_k}) = \\ &\alpha \cdot wLev_{vis}(M_{i_k}, M_{j_k}) + \\ &\beta \cdot wDist_{ac}(M_{i_k}, M_{j_k}) + \\ &\gamma \cdot wLev_{mot}(M_{i_k}, M_{j_k}), i \neq j, \end{aligned} \tag{4.1}$$

where $w(Lev|Dist)_{modality}$ is the weighted (Levenshtein) distance according to the values from modally determined (**vis**ual, **ac**oustic, **mot**oric) psycholinguistic letter distance matrices, M_{i_k} is the $k-$th token (alignment position, often word) of the $i-$th manuscript and α, β, γ are the respective weights for the modalities. The final distance of a variant text pair is then simply

$$\sum_{k=1}^{length(A)} \Delta(M_{i_k}, M_{j_k}). \tag{4.2}$$

In the even simpler conditions for comparison, the distance function Δ simply returns 1 (in case of difference) or 0 (in case of identity) of the elements in (M_{i_k}, M_{j_k}) or in the other condition $Lev(M_{i_k}, M_{j_k})$.[3]

[3] In the design of the formulas, especially the linear weighting factors go back to the supervisor of this thesis.

Modalities

Copying is a very complicated process and builds on many cognitive processes, compare Hoenen (2014a), amongst others reading, retaining the read in memory and writing are involved. These make use of vision, probably acoustics (as far as retention in memory is involved) and motor innervation of the muscles responsible for the movements leading to writing. Among human modalities or senses, those three are assumed to be the decisive ones for the copy process.

Whilst human languages differ profoundly in a number of parameters, the basic receptory and cognitive apparatus is essentially the same for all humans. Consequently, basic confusability patterns should across time and language be roughly stable. Therefore, we believe one can use psycholinguistically derived confusability information for a weighting regardless of the time period or language from which the textual material may stem.

Vision and Reading In comparison to the other modalities, vision is not only the most important one, but witnesses by far the largest body of research on confusability of letters. In order to model the values of the visual modality, matrices of visual confusability have to be used. Müller and Weidemann (2011) have compared 55 papers from 1886 until 2011 that describe 74 experiments (the majority using psycholinguistic approaches (ca. 82%)) to establish letter discriminability matrices for the Latin alphabet. As many tables as were readily available from the supplied paper links have been extracted and it was ensured that they were labelled for

1. modality (visual, motoric, acoustical)
2. directionality (symmetric matrix?, $\Delta(< a >, < d >)$ $= \Delta(< d >, < a >)$?)
3. letter set (upper case, lower case, numbers, mixed case)
4. polarity (similarity or distance)

However, some matrices or data reported in the papers were not used, since they either analysed irrelevant data (perception in pigeons (Blough, 1985), discrimination of the Braille alphabet, (Gilmore et al., 1979)), reported a poor predictive performance (Coffin, 1978), provided incomplete data (Uttal, 1969), featured very few observations (Banister, 1927) or were hardly extractable due to age or condition of the pdfs. We ended up with 27 matrices.

In order to make all matrices comparable, the values were normalized to a number between 0 and 1 by using the largest value as 1, if and only if the reported values were not already in that range explicitly representing percentages. The values were transformed if necessary turning similarities into distances. Furthermore, distances were averaged if directional differences existed: $\Delta(a, b) = \Delta(b, a)$. This was primarily done since the direction of copy when comparing two manuscripts is not apriori known. All non observed letter combinations would receive the maximal distance. For numbers Keren and Baggen (1981) provided a table and for mixed case only Boles and Clifford (1989) reported confusability values. This gave 1 matrix for numbers and mixed case visual confusion and 6 matrices for lower case to lower case letters and 17 for upper case to upper case letters. We combined those and obtained and tested 102 combinations of visual uppercase, visual lowercase, visual mixed case, visual number confusabilities with acoustic and motoric confusion matrices. Matrices have been made available on GitHub.[4]

Acoustic Modality For acoustic confusion, the process of modal transition from and into the visual medium must be modelled as an additional step. Naturally, one could choose grapheme-to-phoneme (g2p) and p2g based approaches. However, since the aim of the present study is to analyse explicitly modally motivated errors, we alternatively do the following and leave g2p/p2g as an alternative for future research.

Cutler et al. (2004) provide phoneme-based confusability matrices. We use the ones for initial vowels and consonants discriminated by natives. In word initial position, the phonemes do usually not become subject of heavy coarticulation.[5] Additionally, there was a high canonical correlation between initial and final confusability values (vowels initial and final:0.99, consonants in onset and coda: 0.81). For the mapping of phoneme pair distances to graphemic units (GU), Van Berkel (2005)'s *basic*, *contextual* and *word specific* spellings were used for English.[6] For instance, the presumably confusable GU pairs potentially

[4] `https://github.com/HoenenA/MultiModalDistance/`. References to all in Müller and Weidemann (2011).

[5] Coarticulation is a linguistic phenomenon whereby some phonemes are influenced by previous or subsequent ones.

[6] Van Berkel (2005) analyses the English spelling system postulating for each phoneme a *basic spelling* which reflects the most frequent spelling for this phoneme, a *contextual spelling* repre-

representing /aʊ/ and /aɪ/ constructed from this were: <ou>:<i>, <ou>:<y>, <ou>:<ie>, and <ow>:<i>, <ow>:<y>, and <ow>:<ie>. The same corresponding normalized distance value from the matrix of phoneme distances was assigned to each of them and used with the acoustically weighted Levenshtein distance. If one GU pair could represent multiple phoneme pairs, all of its values were averaged. For Finnish and French similar resources were used to obtain GUs (Lyytinen et al., 2013; Lehtonen, 2013; Wiik, 1965; International-Phonetic-Association, 1999; Guex and Pithon, 1975; Dryer and Haspelmath, 2013; O'Grady et al., 1997).

Those acoustic distances between phonemes for which Cutler et al. (2004) have provided no values have been estimated using the average of the values of all observed pairs, which had a similar distance. This distance was measured in terms of numbers and qualities (backness, height, roundedness) of edges in the vowel diagram or number and quality (place, manner and voice) of steps in the consonant table of the International Phonetic Alphabet (IPA). See Figure 4.2.1, here, to come from /a/ to /ɜ/ requires a shift in height (to reach ɛ) and then one in backness summing 2 steps. To estimate the value of a pair where at least one phoneme was not observed by Cutler et al. (2004), and where their distance was one shift in height and one in backness, we sum and average all distances for observed vowel pairs from the chart which have that very distance. Analogously, for consonants, for instance /n/ to /m/ has 4 place steps, /n/ to /d/ 4 place and 1 manner steps, /n/ to /t/ an additional voice step etc. For diphthongues with missing values, the distinction was made between such diphthongue pairs which shared at least one sound and such which did not and the concurrent observed averaged values were assigned. In order to obtain the corresponding $n : m$ sequences, the FileDiff algorithm (Miller and Myers, 1985) was used to align tokens at corresponding positions.

Motor Modality While biologically muscular neurology is well-understood, research focussing on letter production from a motor-perspective is comparatively rare. In Müller and Weidemann (2011), the only mentioned study focussing on letter production is Miozzo and Bastiani (2002), where production errors of one patient are reported, who suffered a brain damaging intoxication. They

senting a frequent but not the most frequent spelling and *word specific spellings*. Corresponding phonemes have been mapped from American to British English in the process.

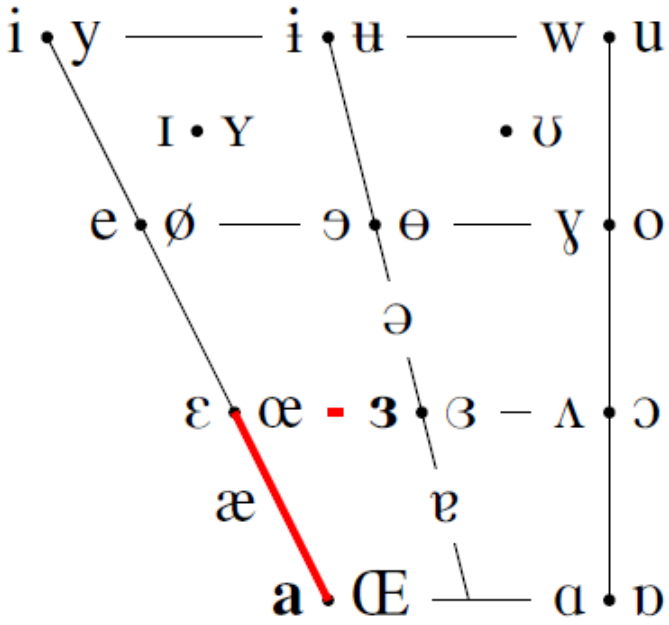


Figure 4.2: The vowel diagram from the IPA. Highlighted in red and bold is a path starting in /a/.

found letter substitutions to occur predominantly between letters with common strokes such as <b> and <p> and remarked that letter frequency, consonant-vowel status and letter gemination were affecting such errors. Their data was used to define all motoric weights and truly corresponds to handwriting. Non observed pairs received the maximum distance.[7]

4.2.2 Stemmatological Application of MMD

We used the artificial datasets to compute a pairwise distance matrix with the MMD (ignoring gaps in the alignment) using each combination of matrices (102 combinations). From the pairwise distances, we computed a stemma using NJ from the R package *ape* and then scored it using the so called Average Sign Dis-

[7]Values on $< a/e >$ were not used.

Trad.	Bin	Lev	MMD	RH09
NB	69.35	58.74	69.35	**77**
PRZ	72.11	67.58	72.11	
PRZ_loss	76.04	71.28	76.04	**87**
HR	72.71	72.9	74.32	

Table 4.1: Comparison of stemmatological evaluation results with ASD on the percentage of shared words (Bin), the Levenshtein distance (Lev) and the MMD (psycholinguistically weighted Levenshtein distance). RH09 gives the best achieved results of the 2009 study of Roos and Heikkilä.

tance (ASD),[8] an accuracy value introduced by Roos and Heikkilä (2009) as:

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

A, B and C are nodes present in both the true and the estimated stemma, $d(A,B)$ is the distance of the two nodes in the true stemma defined as the number of edges on the shortest path between them, $d'(A,B)$ the same distance for the estimated tree. $sign(d(A,B) - d(A,C))$ returns so to speak only the sign, discarding length, thus -1 if $d(A,B) < d(A,C)$, 1 in the opposite case and 0 if both are equal. The index equals 1 if both stemmata agree and 0 if they differ ($\frac{1}{2}$ if they partly agree, for details see the formula or Roos and Heikkilä (2009)) and is computed and turned into a proportion for all such triples. ASD is the to date most used evaluation metric for stemmata on the artificial data sets used for instance in Lai and O'Sullivan (2010); Roos and Zou (2011b); Hoenen (2015a).[9]

Experiment and Results

For each of the artificial traditions, we tested 102 combinations of uppercase and lowercase confusion matrices and for each such combination, we tested 66 different parameter settings, including such where the weight for any one parameter

[8] While on the level of path comparison operating on distance, in terms of the overall manuscript comparison, the ASD is rather a similarity and referred to as Average Sign Similarity by other authors.

[9] A python and a C++ implementation are available from Roos and Heikkilä (2009) through the stemmatology challenge website.

was 1.0. In all, these were $6,732$ combinations per tradition, thus roughly $27,000$ results. Since this is far too much to be displayed in a simple table, we give results in several different ways. Table 4.1 contains the *best achieved results* of the MMD for each tradition (including the further not focussed loss scenario). Ranges between the best and worst results in the whole grid of $6,732$ configurations were roughly 24% ASD for NB with the worst result $45,5$ for PRZ (worst:67) and 26 for HR (worst:48).

As what regards the combinations of uppercase and lowercase matrices, it must be said that those transitions truly involving only uppercase letters were rare and those involving mixed case still very infrequent (in NB roughly 10% and for PRZ roughly 14%) in respect to lowercase to lowercase transitions. All matrix combinations (1 uppercase, 1 lowercase confusion matrix) witnessed parameter settings for which the respective best results were produced. Averages per matrix (over all parameter settings) produced roughly similar results and no matrix combination was an extreme outlier.

As for the values of the modalities, we looked at the weighted average contributions of the parameters, that is for each modality: $\sum_{i=1}^{6732} \omega * ASD[i]$, where all 6,732 ASD values are in one array and each position of the array is conditioned by four parameters: the matrix combination, the weight for vision, acoustics and motorics and where ω is the corresponding weight for the modality under investigation.

Values were almost identical for the modalities and a desired significant difference was not visible. Looking only at those results, where one of the parameters had been set to 1, only for NB some significant pattern emerged: vision (contributing 39%) worked slightly better than audition (contributing 34%) while motorics performed a little worse (contributing 27%) and deviated from the mean significantly (t-test, significance level 0.01). Getting a deeper insight, while this was not possible for the matrix combinations due to the categorical character of these data points, for the weighting factor values additionally a Pearson correlation analysis with the ASD array could be conducted, which yielded more interesting results, see Table 4.2.

There is a strong positive correlation of the acoustic modality with the result for PRZ and NB and a weak one for HR. It must be said however that these correlations are to be understood as on the conjunction of parameter settings. This means that a larger negative value does not automatically mean that there is a bad

Trad.	vis	ac	mot
NB	$-0.18(58)$	$\mathbf{0.56}(66)$	$-0.39(45)$
PRZ	$-0.43(69)$	$\mathbf{0.83}(72)$	$-0.4(68)$
PRZ_loss	$-0.6(67)$	$\mathbf{0.51}(76)$	$0.03(71)$
HR	$-0.12(73)$	$\mathbf{0.29}(74)$	$-0.17(73)$

Table 4.2: Pearson correlations between ASD values and modal parameters, strongest per row highlighted. In brackets, ASD value when modality was used exclusively (weighting factor set to 1 and all other weighting factors to 0, average over combinations of letter case confusion matrices).

effect of this modality but solely that in conjunction with at least one better performing modality, the contribution to the overall result was moderate. In other words, the reason why ASD values suffer if the negatively correlated modality gets stronger may be the result of the more effective ones getting weaker not necessarily because of a bad fit of the modality itself. This is corroborated by the values where the single modalities were used exclusively and by the fact that the best overall results were often only reached in settings where the modalities had been combined.

We conducted all the above analyses in the same way also for the so called TASCFE corpus (Hoenen, 2015a) which has some special characteristics such as being based on 4 different initial versions (entailing multiple roots or 4 clusters) and can be used as a testset to robustness of a stemma generating algorithm. Secondly and most importantly, TASCFE is written in Persian making all letter distance matrices from Müller and Weidemann (2011) unuseful. Not least because of the small length (the alignment features only 137 positions) this data set is the most challenging and produced not unexpectedly the worst results. Best ASDs were roughly between 56 and 63 (still far from chance) on the 4 complete single subsets. For the MMD, visual confusion resulted in only 1 matrix which had been modelled based on the similarity of letter features in Wiley et al. (2016), motoric similarity came from the same source but used the reported stroke similarity between the letters. Acoustic to graphemic mappings were deducted from the International Phonetic Alphabet (International-Phonetic-Association, 1999).

4.2.3 Discussion

Results are generally negative in that they did not outperform the best of some previously reported values although they are in the range of many of the there presented approaches.[10] However, to a certain extent this was expectable, given that only a subset of the innovations found which occurred along the textual transmission are estimated to be the direct result of simple modal or multi-modal confusion on the token level. Consequently, we first looked at the data in more detail to find out how many of the deviations were possibly such captured by the MMD. We conducted a tentative and surely partly subjective classification to this end. In order to better be able to interpret this data, the analysis was confined to NB (French) and PRZ (English). Some of the confusions occurred many times in different copy processes not always of the same source manuscript. Some of them, in the French case, are presumably reverts of a dictation copy where a non-native speaker had misrecorded some silent endings. A large number of cases involved deletions or insertions of letters, which possibly in part explain why the results of the binary distance and the MMD are not differing for NB and PRZ. Those are not subject to the MMD weighting schema but will make token distance coincide with the Hamming value. Another reason can be that some distances in the matrices (for some matrices, the majority of distances) were so small that they could hardly make a big difference as compared to the Hamming distance.

Overall, we found roughly a third of the differences to be applicable to a visual (127 of 422) or acoustic (115) weighting. Motoric confusion was deemed possible for roughly a sixth (56) of the cases. In conjunction roughly half (206) of the variation was subject to MMD weighting.

As for the matrices, generally their performance was not extremely different with some interesting observations. Müller and Weidemann (2011) comparing 11 of the matrices found a mean correlation of 0.68 to the generated average matrix (p.30), which well aligns with our observation. The matrices were all qualitatively roughly similar but some would have a large range between smallest and largest values, some would give differing values for self similarity. In

[10]Note, that our results on the binary distance combined with NJ achieved slightly worse results than those obtained in the challenge using the same data, metric and algorithm, for NB (theirs:76.2, ours:69.35), PRZ (theirs:81.5, ours:76.04) different implementations and/or manuscript text orderings may output different trees.

cases with no self similarity reported, we had assigned unity regardless of the magnitude of differences with and in the rest of the matrix. Overall, the matrix of Geyer (1977) performed best by a very small margin. Courrieu and De Falco (1989)'s matrix although on average best performer for English clearly performed worst for French. The data had been gained from the confusions of preschoolers and showed a presumably acoustically decoupled confusability component (<p> with <q>). Such relations might have influenced the distance matrix in a way as to overwrite some genealogical relations for French so that NJ, which is a greedy algorithm found some tree quite different from the others. Generally, the information from the differences which are not measurable by MMD may additionally crucially determine the schema of information reduction from distance matrix to stemma and thus obscure the fit.

Comparing the different metrics, interestingly, the Levenshtein distance was clearly outperformed for NB and PRZ by the binary distance despite Levenshteins ability to measure the degree of difference between two tokens. For HR however, this was not the case. Furthermore, for HR MMD was outperforming the binary distance. This result may be due to the writing systems of the languages involved. More specifically, Katz and Frost (1992) introduce the notion of *orthographic depth*. English and French in this sense are deep orthographies, that is their g2p and p2g relations contain many n:m relationships, whereas Finnish is a shallow system (Joshi and Aaron, 2013). Illustrating the difference between a deep (English) and a shallow (Finnish) orthography, it may suffice to look at the following two examples:
P2G: $/k/ \rightarrow$ {<c>, <k>, <ck>}^{EN}, {<k>}^{FIN}
G2P: <a>$\rightarrow$ {/ɑ/, /ɑː/, /, /ɒ/, /æ/, /eɪ/}^{EN},{/ɑ/}^{FIN}
The values from our confusion matrices in the MMD cover 1:1 letter confusion values. If now confusion took place also on some levels of graphemic units, these would not be captured by the visual and motoric confusion values, albeit by the acoustical ones. For instance the confusion of <their> and <there> could entail such a larger-unit-based confusion, where not one letter is cofused with one other letter. Acoustic distances as modelled however take into account such units since there is the above-described mapping between phonemes and graphemic units. In fact, the positive correlations of the acoustic weighting factors seem to support such an interpretation. Moreover, since Levenshtein may assign too large a value to confusions which involve n:m relations that correspond to just

one confusion it may introduce noise, especially for deeper orthographies, so much so that its overall result becomes worse than the binary distance.[11] For the same reason, MMD is distinguishable from the binary distance for Finnish. In this vein, results all seem to be most consistent with an interpretation which suggests that the proposed method currently works best for texts written in languages with a shallow writing system (e.g. Latin). Confusion matrices for more complex orthogaphic units could be useful.[12] It also suggests that the level of graphemic units could be quite important in analysing confusion (on token level). This interpretation would be consistent also with the Persian data, where taking into account abstract letter identities[13] produced some better values than the graphemic distances.

However, utter caution must be taken since the data sets are by all means small and not representative of historical data as such. Their size entails a grave danger of overfitting, which is why using methods of machine learning to optimize the weights may be dangerous and surely much more effective on larger data sets. Additionally, our model of an interplay of the modalities is not the only possibility and ideally each position of a manuscript would require some different weighting input or an entirely different model (for instance if not modal confusion on the token level but contextual priming effects paired with some degree of visual similarity cause miscopying (Hoenen, 2015b)). There are confusions, where one single modality is to be held responsible. When Spencer et al. (2004a) mention the exapmle of <cl> and <d>, it is unlikely that the reason for the confusion lie in any other modality than vision. Thus modelling each modality separately and summing them, apart from having neurological correlates, is not unreasonable but the presented approach is surely just a first step to investigate a complex and data sparse object.

[11] A similar explanation may hold for the observation of Spencer et al. (2004b) who found that subjective weights had made few difference. Here, weights might have accidentally obscured the genealogical information although the weighting, quite like Levenshtein may not have been unreasonable in itself.

[12] To this end, some experiments with OCR error data on n:m confusions showed positive effects.

[13] An abstract letter identity is a cognitive entity which connects different elements of a writing system behaving in the same way, for instance the lowercase 'incarnation' and the uppercase 'incarnation' of a letter $\{a, A\}$.

4.2.4 Conclusion

We presented an approach to weighted stemma generation from pairwise manuscript text distance matrices. In the approach, external data in the form of psycholinguistically generated letter and phoneme distance matrices in the visual, acoustic and motoric modalities was used to model weights for a weighted version of the Levenshtein distance. We tested and evaluated the approach producing stemmata from manuscript pair distances of three artificial data sets with known ground truth. matrices for each tradition and compared the MMD results. Results were not outperforming the best results reported in Roos and Heikkilä (2009), but in all cases were better than many other approaches. Which external input matrix to choose was found not to be crucial in our setting and all combinations of matrices performed very similarly. Regarding the contribution of the single modalities, acoustics as modelled performed very well, but best results were often only achieved when the modalities were combined in a weighting schema. We additionally found that most likely orthographic depth was the reason why MMD outperformed the binary distance only for Finnish and why the unweighted Levenshtein distance was outperformed by the binary distance for the French tradition NB and the English tradition PRZ. The main contribution of the section is thus in corroborating an argument in the discourse. That argument is that weighting beyond the word level may make sense, but weights must be carefully elicited and theoretically grounded for instance using psycholinguistically derived confusability matrices. Approaches to weighting which are not confined to the comparison of manuscript and token pairs, but which take into account additional distributional information of each variant, such as the one presented by Roelli and Bachmann (2010) could improve results of weighted approaches in another vein, which a quantitative assessment for instance against the benchmark datasets could reveal. We conclude with a word of caution, that all results have been obtained on relatively small data sets.

4.3 Minimum Spanning Trees

This section represents work, which in a substantially extended version is being prepared for publication together with A. Mehler. Formula (3.3) and the use of Graph Edit distance as additional evaluation metric is among his contributions.

There are different methods to derive trees from distance matrices. Among those, the ones originating in bio-informatics have been widely used and influenced the field of computational stemmatology deeply. To date, three methods have formed the de facto standard for this purpose: the *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA), *Fitch-Margoliash* (FM) (Fitch and Margoliash, 1967) and *Neighbor Joining* (NJ) (Saitou and Nei, 1987). While UPGMA makes some simplifying assumptions, NJ is computationally less demanding than FM. Due to its competitive results and computational efficiency, NJ is probably being most widely used (Osborn and Smith, 2005, p.42). However, all those methods produce trees which have two rather 'unstemmalike' characteristics: they are *bifid* meaning that all their internodes have a degree of three and secondly, they place all input entities at leaf positions. Christopher J. Howe and Windram (2012) extensively discuss non-technical ways of accommodating these properties with stemmatology. The reason for the need to accommodate are the differing prequisites of the two disciplines. While in biology, the most applied species concept postulates splits into two (Hoelzer and Meinick, 1994), there is no reason to believe that with the same rigour manuscripts have been copied only twice. Secondly, the case that a vorlage and at least one copy of the subbranch rooted by that witness (thus its descendent) have survived appears to be so common in stemmatology that the stemmatological method of Lachmannianism includes an obligatory step where witnesses are excluded if an ancestor of them is contained in the corpus (*eliminatio*). Hence, placing all input entities at leaf positions would be inappropriate or unstemmalike. In contrast to this and as an additional possibility, apart from interpretational or accommodation through manual effort, one can use an algorithm operating on a pairwise distance matrix which does apriori produce trees which do not have the undesirable properties. Such a tree could be a Minimum Spanning Tree (MST) which is a spanning tree – that is a tree spanning all nodes of the fully connected weighted graph obtained as a representation of any pairwise distance matrix– with a minimal sum of edge weights (of all possible spanning trees). An MST is neither bifid nor does it place input entities only at leaf positions. A complication might be seen in that no

hypothetical nodes are assumed (Marmerola et al., 2016) which however for the complete data sets of the artificial traditions does not hold. Another problem is that for a given matrix corresponding to a fully connected weighted graph, there can in principle be several MSTs iff at least two weights in the matrix are the same. Generally, multiple MSTs could be empirically rare depending on the data type. MSTs are in use in biology as well (Teixeira et al., 2015) for instance for viral strains Spada et al. (2004).

4.3.1 On MSTs

MSTs conceptually developed in the 1920s (Nešetril and Nešetrilová, 2012) and since then several different efficient algorithms have been proposed to compute/find them, the most well-known (and most implemented) of which are Kruskal (Kruskal, 1956) and Prim (Prim, 1957).

MSTs may symbolize various kinds of networks from streets to chemical compounds. It is thus not surprising to find them in many domains. According to Bazlamaçcı and Hindi (2001, p. 768) MSTs have "direct applications in the design of computer and communication networks, power and leased-line telephone networks, wiring connections, links in a transportation network". In NLP, they have been applied variously, for instance in Word Sense Disambiguation (Tsatsaronis et al., 2008) or clustering with an application scenario in improving language models (Manning and Schütze, 1999, pp. 504). Marmerola et al. (2016) widely apply MSTs to multimedia phylogenies (including image trees) and plagiarsim, but also for the reconstruction of the versioning histories of Wikipedia articles with a section testing applicability in stemmatology. MSTs have also been basis for further developments on the theoretical level in NLP, consider for instance Minimum Spanning Markovian Trees (Mehler, 2010).

In stemmatology, Minimum Spanning Trees have been used by Lai and O'Sullivan (2010) on the matrix of the normalized Hamming distances.

4.3.2 MSTs and distance matrices in stemmatology

Although in both domains – distance matrix construction and stemma generation from distance matrices – there have been studies evaluating presented methods such as Roos and Heikkilä (2009), a systematic comparative survey combining both has not yet been conducted. This is the task of the present chapter. It

can help to decide which methods to choose for which data and application scenarios. To this end, we show that MSTs are an effective, easy to compute tool for generating stemmata and evaluate our approach using five different data sets. For evaluating our approach, we use the three widely used artificial datasets. We use the manually aligned texts obtained from the authors. Apart from not being representative of the whole gamut of historical variety, these datasets have each some particularities. NB contains, for example, a dictation copy, while HR has Latin text in some places and a very low degree of contamination as a result of the confluence of readings from several original manuscripts. For the present study, these particularities are to be regarded as noise.

4.3.3 Approach

In this experiment, we survey 4 methods to generate distance matrices from the above mentioned data and 4 algorithms for generating spanning trees from this data. We apply them to 3 data sets and, thus, consider 48 different evaluation scenarios.[14] Furthermore, we provide a method for both postprocessing and visualization in case more than one MSTs are generated. Lastly, we evaluate all trees using Average Sign Distance and all MSTs additionally using Tree Edit Distance.

Distance Matrix Generation

The first method to generate a distance matrix is the Hamming distance (Hamming, 1950) (H) counting simply the number of differing alignment positions (thus, typically tokens) between any pair of texts. It is used for instance in Lai et al. (2010).

The second method is the Levenshtein distance (Levenshtein, 1965) (LEV) summed for all token pairs of a manuscript text pair.

The third is a weighted variant thereof, the Multi Modal Distance (MMD), a Levenshtein distance weighted according to psycholinguistic distance matrices (wLev) is used to obtain modally governed token distances and additionally those are weighted with linear factors α, β, γ. The distance matrix of Geyer (1977) for

[14]The results of a new method for distance matrix generation based on normalized pointwise mutual information as proposed in Bouma (2009) suggest that this method does not improve distance matrix generation in stemmatology.

lowercase letters performed very well, which is why it is used here together with the matrix for uppercase letters from Townsend (1971) since all of the uppercase matrices performed similarly. We set the linear factors to $\alpha = 0.2$, $\beta = 0.7$ and $\gamma = 0.1$ since this is the setting where on the conjunction of the data sets, the overall best results are achieved when all three modalities are involved.

The fourth method is based on an interpretation of a philological entity called *leitfehler* (LF). A leitfehler according to Roelli and Macé (2015) is understood as a genealogically significant error. The only algorithm implementing a method based on the leitfehler comes from Roelli and Bachmann (2010), see also Roelli (2014b). According to Roelli and Macé (2015, p.129) the algorithm still needs development and characterize it meanwhile as:

> [...] a subcategory of distance-based methods. Thus the traditional scholarly concept of Leitfehler is taken to be a quantitative one: a variant's usefulness as Leitfehler may be assigned a number or weight. In classical stemmatology the Leitfehler is the most important tool to arrive at a filiation of witnesses that is believed to be most correct representation. [..., evaluating how good a leitfehler a variant is] for every pair of them [...] If one of the four combinations of absence / presence of any of these two candidates is not represented in any witness, this is taken to be a hint that both variants suffered their change from absence to presence (or vice versa) exactly once in the tradition, which is characteristic for good traditional Leitfehler (Maas 1937). Such a comparison can be made for all combinations of potential Leitfehler while both Leitfehler in pairs with only three combinations get their score increased.

The algorithm of Roelli and Bachmann (2010) produces a list, where every variant in the textual tradition gets a value that indicates how good a leitfehler it is. This list can be and in our case is pruned to roughly the upper third (using the initial threshold of Roelli and Bachmann (2010)). The idea is to compute a similarity (inverse: distance) for each manuscript pair based on how many good leitfehler they share. The score for each variant is based on the idea to compare it with all variants, where each of the other variants per position counts only once. If we call the array of variants V, the score for each $v_i \in V$ is computed by comparing it to all other $v_j \in V, i \neq j$. For each such variant pair (v_i, v, j), each manuscript pair is considered and the numbers of occurrences are tabulated

	v_1	$\overline{v_1}$
v_2	100	0
$\overline{v_2}$	73	37

Table 4.3: An exemplary tabulation for a variant pair in the leitfehler method. The sum of values equals the number of manuscript pairs in the tradition and hence the number of fields in one half of the distance matrix (diagonal omitted).

with respect to cooccurrence in terms of truth values. For an illustration, see Table 4.3. Only in case, one of the 4 fields is empty, the score for being a good leitfehler is increased for v_i. In this way each variant $\in V$ obtains a score. Yet Roelli and Bachmann (2010) apply some additional weighting depending on the truth table configuration, which is where they see further space for development. We use the original code provided by them used for generating results in their concurrent publication. The advantage of the leitfehler method is that it uses distributional information on the variants, while all other methods mentioned above do not leave the context of strict pairwise text comparisons.

Tree Generation

For tree generation from distance matrices there are several methods available. To date the most widespread methods in bio-informatics – constituting the majority of applications and research on transformations from distance matrices to trees – to solve this task are UPGMA , FM and NJ. All three have been documented extensively elsewhere.

Another possibility is to view the matrix as fully connected graph and generate an MST. An MST is a spanning tree (covering all nodes in the graph) whose edge weight sum is minimal (given all possible trees the node set of which is identical to that of the complete graph). Compared to NJ, MSTs have been applied scarcely in bio-informatics, because of some –taking the systematicists perspective –impractical features. First, MSTs do put the input species at internode positions where in taxonomical studies extant taxa should all occupy leaf positions. Secondly, MSTs are not limited to outputting exclusively bifurcating trees as are preferred but not exclusively used in phylogeny, see for instance Slowinski (2001). For stemmata however both of these drawbacks are not to be seen as drawbacks, since extant manuscript texts can well occupy internode positions and since furthermore multifurcation is rather rule than exception in stemma-

tology. From the point of view of stemmatology, rather the fact that an MST is usually unrooted appears slightly impractical since philologists before using computers at least in mainstream philology did root their trees. However, since at least Haigh (1970) rooting algorithms for stemmatology are known. Rooting for stemmatology may be seen as harder than in the biological case since the concept of outgroup is inapplicable in stemmatology. Sinsheimer et al. (2012) remark that "there are no general methods available for determining roots" but present one which targets cases where the outgroup is absent. Historical and psycholinguistics may provide some remedies for stemmatology with respect to rooting in the near future, see also Marmerola et al. (2016).

Furthermore, MSTs have another property which may cause problems in both biological and stemmatological applications, namely the fact that there can be more than one MST for any given fully connected graph with weighted edges (corresponding to a pairwise distance matrix) Yamada et al. (2010). For multiple MSTs to exist however, there must be at least 2 edges with the same weight, compare Wright (1997, Lemma 2.1). This means, that for distance matrices with only unique edge weights there can only be one MST which empirically certainly is the most probable scenario. For distance matrices using smaller and similar integer distances in general the existence of more than one such trees is more probable. In our scenario, the H, LEV and LF algorithms produce such matrices. We implement the first approach of Yamada et al. (2010)to generate all MSTs using Java and the library *jgrapht*.

Evaluation

Once a tree has been built, it remains to be evaluated. In the last decade evaluation against benchmark datasets (or artificial traditions) has been conducted in stemmatology Baret et al. (2004); Spencer et al. (2004a); Roos and Heikkilä (2009). These datasets have been generated by first giving one "root" text to volunteers to be handcopied (or dictated). Its copies have then been handcopied again and so forth. The true *vorlage*[15]-copy relations (edges in the true stemma) have been recorded. Texts have been digitized and manually aligned. For evaluation, Spencer et al. (2004a) use partition distance (PD) (Penny and Hendy, 1985), which they summarize (p. 505) as "the number of edges on the first tree for which

[15]Vorlage is a German loan, which is used in philology to describe the model or the original of a copy.

there is no edge on the second tree whose removal divides the manuscripts into the same two subsets [as those obtained by removing the respective edge on the original tree]". As a second method they use triplet symmetric distance (TSD)[16], where all triplets of nodes in both trees are compared for topology. Spencer et al. (2004a) remark that partition distance can be influenced by few "rogue species", while TSD is defined for trees which have exactly the same number of labeled leafs with the same labels which would require a reduction and transformation of the original gold standard tree for stemmata, where both extant manuscripts can be found at internodes and where the gold standard tree ususally has more labels than the estimated tree. In a subsequent study, Roos and Heikkilä (2009) have introduced the Average Sign Distance (ASD), which compares all triples of shared nodes between two trees looking at rough topological features. ASD is thereby not dependent on having the same labeled node set in both trees nor on having all nodes at leaf positions. It further is agnostic to directionality. A drawback is that general topological properties of the estimated tree (such as number of hypothetical internodes etc.) do not get evaluated. ASD has been used further for instance by Lai and O'Sullivan (2010); Roos and Zou (2011a); Hoenen (2015a) and is defined as:

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

A, B and C are nodes present in both the true and the estimated stemma, $d(A,B)$ is the distance of the two nodes in the true stemma defined as the number of edges on the shortest path between them, $d\prime(A,B)$ the same distance for the estimated tree. $sign(d(A,B) - d(A,C))$ returns so to speak only the sign, discarding length, thus -1 if $d(A,B) < d(A,C)$, 1 in the opposite case and 0 if both are equal. The index equals 1 if both stemmata agree and 0 if they differ and is computed and turned into a proportion for all such triples.

Another well established measures include the Graph Edit Distance (GED) or in case of dealing with trees, the Tree Edit Distance (TED). Since the ASD does not take into account true path lengths between the nodes in a triple, since furthermore unlabeled nodes are not being considered, naturally, many quite different graphs could get the same ASD value. One of the reasons behind this is surely

[16]as implemented in COMPONENT, `http://taxonomy.zoology.gla.ac.uk/rod/cpw.html`

the application of bio-informatic algorithms which infer unlabeled internodes so as to produce an entirely bifurcating tree. This implies that in such a tree all labeled nodes are leafs and no two labeled nodes are connected through an edge. Since for MSTs in stemmatology, this restriction is not holding, here, we use a GED which will evaluate the inferred tree on the basis of the edges. Generally, let GED for a true Graph $G = \{V_1, E_1\}$ and an estimated Graph $G' = \{V_2, E_2\}$:

$$\begin{aligned} GED((G(V_1, E_1), G'(V_2, E_2)) = \\ |V_1| + |V_2| - 2(V_1 \cap V_2) + \\ |E_1| + |E_2| - 2(E_1 \cap E_2) \end{aligned} \tag{4.3}$$

In our case, the upper half of the equation, concerning the nodes will always equal 1 since we test on the entire traditions. In order to normalize to a value $\in [0, 1]$, we normalize with $|E_1| + |E_2|$.

4.3.4 Results

Table 4.4 summarizes the results which have been obtained from combining all methods to generate a distance matrix with all tree building methods. One immediately sees that MSTs produce the best results overall often by large margins. UPGMA trees perform clearly poorest. For the distance matrices, the results are not so clear with all methods lying closer together in many cases. Bootstrapping on the complete datasets deteriorated results. There were few MSTs in the H, LEV and LF conditions with a maximum of 16 for NB.

4.3.5 Discussion

For complete traditions deriving a tree is equivalent to an edge classification task and MST performs very well here. All MSTs must be connected by a chain where subsequently only one edge is exchanged between each of them, see Wright (1997). Thus, especially when there are few MSTs, those will be sufficiently similar regarding shared edges making the average evaluation value which we give here a good approximator of the goodness of stemmatic fit.

The TED allows to better distinguish between the distance matrices.

For comparison we computed all MSTs also for two reduced data sets used in the challenge of Roos and Heikkilä (2009),[17]. Bootstrapping improved the NJ-

[17]Those datasets had been subsets of the PRZ and HR sets, which had been reduced in order

Distance Matrix	MST	TED	NJ	UPGMA	FM	DM Average
Parzival						
H	*89.129(4)	0.15	72.105(69.687)	51.692	72.258	71.258
LEV	*82.406(1)	0.25	67.581	53.471	69.173	68.158
LF	**96.259**(4)	0.1	80.351+	60.476+	80.05+	*79.284*
MMD	*82.406(1)	0.25	71.028	55.05	71.241	69.932
Method Average	*87.55*		72.766	55.172	72.661	**MST-LF**
Notre Besoin						
H	68.211(2)	$0.\bar{3}$	*69.347(69.289)+	60.839+	65.909+	*66.077*
LEV	**72.086**(1)	$0.\bar{3}$	60.082	55.711	60.373	62.063
LF	*68.779(16)	0.375	57.809	56.527	54.196	59.328
MMD	*71.97(1)	$0.\bar{3}$	69.347+	56.061	67.832	66.303
Method Average	*70.261*		64.146	57.285	62.078	**MST-H**
Heinrichi						
H	*86.519(8)	0.295	72.705(71.604)	54.798	68.697	70.68
LEV	*86.445(2)	0.341	72.898	51.015	77.376+	71.933
LF	**86.669**(2)	0.258	72.425	55.523	65.601	70.055
MMD	* 86.407(1)	0.288	73.863+	55.701+	76.715	*73.172*
Method Average	*86.51*		72.973	54.259	72.097	**MST-MMD**

Table 4.4: Evaluation for all combinations of distance matrix (DM) generation metrics combined with tree generation methods for 3 benchmark data sets. In brackets for MST: number of MSTs, for NJ-H: result when applying 100 bootstrap samples and collapsing bifurcations with support less than 50. Bold numbers represent the best result per tradition, besides this, each best result per row in the other rows of the tradition carries an initial star (best tree generation method on current distance matrix) and in each column other than the one with the winning value, the best distance matrix carries a final plus. Best averages (Av.) are cursivized.

H condition by small margins there from 76.042 to 77.946 for PRZ and from 62.57 to 63.227 for HR. Bootstrapping is expectably more effective on reduced sets recovering the original topology. The best result on PRZ(loss) was 80.506 for the LF matrix with the NJ algorithm. Best result on HR(loss) was an average of 78.76 for 4 MSTs on the H matrix, on par with best previous results (Lai and O'Sullivan, 2010) who use the normalized Hamming distance (they do not mention multiple MSTs), however here, we show that there are 4 equally likely MSTs in our case. The results were generally similar in that UPGMA performed clearly worst, the MST best (on ASD, TED not being applicable) and with more heterogeneous results for the distance measures.

4.3.6 Multiple MSTs – Strategies

Regarding the use of MSTs, we have seen that in our setting, in slightly more than half of the cases, the MST algorithm provided one single tree. The average number of MSTs was x but removing the largest y outliers reduced this to z.

In order to disambiguate between MSTs, there are multiple possible strategies. One could simply take an algorithm to find one MST, such as Kruskal's or Prim's (Kruskal, 1956; Prim, 1957). However, primarily those algorithms have been designed on the bases of principles to effectively find an MST of a graph. The choice of which of all MSTs these algorithms find is thus determined by algorithmic design choices or by the non-meaningful sequential orders of nodes and edges in the data structure of the underlying graph. Thus, in case that there are many MSTs, one cannot argue that the one produced by one of those algorithms has a superior quality over the others.

MST postprocessing

In order to work with the results from all MSTs nonetheless, one can apply some post- or reprocessing. Since in stemmatology the notion of contamination exists (cross fertilization), we can construct a consensus network containing all edges from all minimum spanning trees, albeit putting them into two disjoint sets. The first one constitutes a base graph with all edges contained in all MSTs. The second set comprises contamination prone relations (edges present in at least one but not all MSTs). Ideally, both types are graphically displayed in different ways.

to simulate one scenario of historical loss.

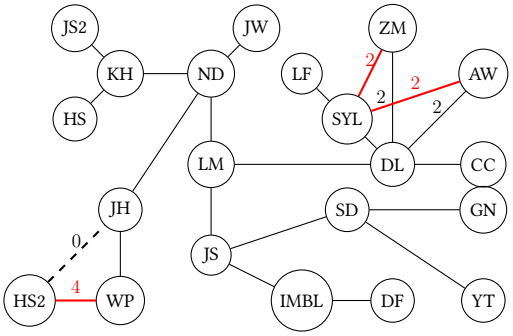


Figure 4.3: The consensus graph of all MSTs for LF. All unlabeled edges have weight 4, that is they occur in all 4 MSTs. The red edges are not present in the true graph, the dashed edge is present in no MST.

Additonally, just like in a bootstrap consensus tree, we can print onto the edges the number or proportion of MSTs which contain it and their original weights (e.g. below and above). For an example see Figure 4.3.

For two reasons, such a network can be statisfying. Firstly, phylogenetic networks (Bandelt and Dress, 1992) have become popular ever since their introduction which entails that researchers are not exclusively interested in the strictly hierarchical relations between taxa, but also in the networks of their interrelations, which an MST network also alludes to. Secondly, especially in stemmatology cross fertilization, a process called contamination in philology, is said to be widespread. The same goes for some biological subdisciplines such as botany and bacteriology.

If one is however not interested in a network, there are numerous ways of how to obtain one unanimous single MST from all MSTs. Here, we want to mention two of them applicable in our case.

Postprocessing: Post-weighting of ambiguous edges in the consensus network of all MSTs. A single tree is only guaranteed if these weights are truly different for the other methods. This possibility complicates interpretability however.

Reprocessing: For methods which are non-binary on token level comparison (all other than H), a strategy is a non-linear transformation for the token level distances prior to summation for the manuscript pair value. For instance, one can use squaring or more generally, one can find the smallest possible natural number power raised to which token level distances sum up to all unique weights in the matrix. This would mean a very unspecific but not counterintuitive down-

weighting of smaller differences, which can be consistent with the data, but there is no good motivation why to use potentiation instead of other non-linear transformations. Furthermore, there can be cases where such a procedure will still produce more than one MSTs. For the Hamming distance potentiation will not change givens, since we deal with bitvectors of zeros and ones both of which are not affected by potentiation. Here, instead of the Hamming distance, one could try the Jaccard coefficient, which could yield different results, but which could essentially also lead to a matrix with even more MSTs. Two other possibilities would be to align the manuscripts pairwise before taking the Hamming distance, instead of using the global alignment of all manuscripts. Again, the effects must not necessarily play to the advantage of having fewer MSTs. Another commonly excerted possibility is to ignore gaps, with the same caveat.

4.3.7 Conclusion

We surveyed the generation of a stemma from a distance matrix, where we used 4 different approaches to distance matrix generation and 4 algorithms for stemma generation. Results have shown that MSTs clearly outperformed the other approaches on our stemmatological data sets. As for distance matrix generation results have been less straightforward. We discussed strategies for dealing with multiple MSTs postprocessing and visualizing them.

Although results have been quite promising, there is room for further improvement concerning distance matrices (new methods), post-processing, evaluation, input data etc. and for elaboration not only in the stemmatological case considering for instance Marmerola et al. (2016); Mehler (2002, 2005); Dehmer and Mehler (2007); Dehmer et al. (2007); Mehler (2009, 2010).

Chapter 5

Closing Remarks

This thesis has presented a series of experiments on stemmatology and the evaluation of stemma generating algorithms using the artificial benchmark data sets, where one has been compiled by the author.

From the viewpoint of the humanities, this thesis has hopefully contributed some clarifying aspects of numerical arguments in the long standing debate on root bifurcativity shaking the very foundations of stemmatology. A dynamic visualisation has been proposed. The thesis has offered an extension for the explanation of the philological principle of *lectio difficilior* based on psycholinguistic knowledge.

The thesis has tried to show the benefits of incorporating psycholinguistic data as sources of external training into Natural Language Processing, in particular stemmatology. Although this is basically no new concept, especially in the context of historical languages for which data sparseness is often a problem, this possibility might be thought of as worthwhile.

Various algorithms blending humanistic, psycholinguistic and statistical knowledge have been presented and evaluated. Especially for the evaluation of stemmata, this thesis has introduced the archetype evaluation scenario, which is more robust than graph based evaluations, since contamination does not bias it to the extent that it influences stemma generation and since furthermore no mapping of reconstructed nodes is required.

Additionally, a new artificial tradition benchmark data set has been compiled, introduced and made available. The results on the MMD and especially on the Minimum Spanning Trees for the leitfehler induced distances have been shown to be competitive. Nevertheless, more artificial traditions are needed to provide

more stable generalizations for automated stemmatology.

It has been tried to combine a text heavy hermeneutic interpretation of the data, with illustrative conciseness and a table and formula loaden style of quantitative analyses. Some of the chapters adhere more to the one others to the other paradigm, where the ideal forms of those chapters may not have been attained. It might therefore rightfully be claimed that one or the other style dominated on the whole, which the author regrets insofar as a truly balanced style should be a hallmark of digital humanism. The candidate hopes with regard to this, that the current thesis, which closes with this sentence, is a step into the right direction.

Ressources (created corpora, tables and source code) have been made available to the public through the TTLab- Website or on Github.

- `https://github.com/ArminHoenen/KFurcatingRootedGregTrees`
- `https://github.com/ArminHoenen/dynamicStemma`
- `https://github.com/HoenenA/MultiModalDistance`
- `https://www.texttechnologylab.org/wp-content/uploads/2015/08/TASCFECorpusDownload.zip`

Bibliography

Alberti, G. B. (1979). Problemi di critica testuale. *Paideia*, 23.

Amodio, M. (2004.). *Writing the oral tradition*. Poetics of orality and literacy. University of Notre Dame Press, Notre Dame, Ind.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research*, 9(13):3015–3027.

Andrews, T., Blockeel, H., Bogaerts, B., Bruynooghe, M., Denecker, M., De Pooter, S., Macé, C., and Ramon, J. (2012). Analyzing manuscript traditions using constraint-based data mining. In *Proceedings First Workshop on Combining Constraint Solving with Mining and Learning (ECAI 2012 workshop)*, pages 15–20.

Andrews, T. L. (2014). Analysis of variation significance in artificial traditions using stemmaweb. *Digital Scholarship in the Humanities*, 31(3):523–539.

Andrews, T. L. and Macé, C. (2013). Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmas. *Literary and Linguistic Computing*, 28(4):504–521.

Andrews, T. L. and Macé, C. (2014). *Analysis of ancient and medieval texts and manuscripts: digital approaches*. Brepols Publishers.

Balduino, A. (1989). Manuale di filologia italiana. *Biblioteca universale Sansoni*, 1.

Bandelt, H.-J. and Dress, A. W. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular phylogenetics and evolution*, 1(3):242–252.

Banister, H. (1927). Block capital letters as tests of visual acuity. *The British journal of ophthalmology*, 11(2):49.

Barabucci, G., Di Iorio, A., and Vitali, F. (2014). Stemma codicum: analisi e generazione semi-automatica. *Quaderni DigiLab*, 3(1):129–145.

Barbrook, A. C., Howe, C. J., Blake, N., and Robinson, P. (1998). The phylogeny of the canterbury tales. *Nature*, 394:839.

Baret, P., Macé, C., and Robinson, P. (2004). Testing methods on an artificially created textual tradition. In *Linguistica Computationale XXIV-XXV*, volume XXIV-XXV, pages 255–281, Pisa-Roma. Instituti Editoriali e Poligrafici Internationali.

Bazlamaçcı, C. F. and Hindi, K. S. (2001). Minimum-weight spanning tree algorithms a survey and empirical study. *Computers & Operations Research*, 28(8):767–785.

Bédier, J. (1928). La tradition manuscrite du 'Lai de l'Ombre': Réflexions sur l'Art d'Éditer les Anciens Textes. *Romania*, 394:161–196, 321–356.

Bergel, G., Howe, C. J., and Windram, H. F. (2015). Lines of succession in an english ballad tradition: The publishing history and textual descent of the wandering jew's chronicle. *Digital Scholarship in the Humanities*, 31(3):540–562.

Blough, D. (1985). Discrimination of letters and random dot patterns by pigeons and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, 11(2):261–280.

Bod, R. (2013). *A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present*. OUP Oxford.

Boles, D. B. and Clifford, J. E. (1989). An upper- and lower case alphabetic similarity matrix, with derived generation similarity values. *Behavior Research Methods, Instruments, & Computers*, 21:597–586.

Bolter, J. D. and Grusin, R. (1999). *Remediation: Understanding New Media*. MIT Press.

Bordalejo, B. (2015). The genealogy of texts: Manuscript traditions and textual traditions. *Digital Scholarship in the Humanities*, 31(3):563–577.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL (2009)*, pages 31–40.

Brotzman, E. (1994). *Old Testament Textual Criticism: A Practical Introduction*. Baker Publishing Group.

Bryant, D. and Moulton, V. (2004). Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution*, 21(2):255–265.

Cameron, H. D. (1987). The upside-down cladogram: problems in manuscript affiliation. In *Biological Metaphor and Cladistic Classification: an Interdiscplinary Approach*, pages 227–242. University of Pennsylvania.

Canfora, L. (2002). *Il copista come autore*. Sellerio, Palermo.

Casson, L. (2002). *Bibliotheken in der Antike*. Artemis & Winkler.

Castellani, A. E. (1957). *Bédier avait-il raison?: La méthode de Lachmann dans les éditions de textes du moyen age: leçon inaugurale donnée à l'Université de Fribourg le 2 juin 1954*. Number 20 in Discours universitaires. Éditions universitaires.

Cayley, A. (1889). A theorem on trees. *Quarterly Journal of Mathematics*, 23:376–378.

Christopher J. Howe, R. C. and Windram, H. F. (2012). Responding to criticism of phylogenetic methods in stemmatology. *SEL studies in English Literature*, 52(1):51–67.

Cisne, J. L., Ziomkowski, R. M., and Schwager, S. J. (2010). Mathematical philology: Entropy information in refining classical texts' reconstruction, and early philologists' anticipation of information theory. *PloS one*, 5(1):e8661.

Coffin, S. (1978). Spatial frequency analysis of block letters does not predict experimental confusions. *Perception & Psychophysics*, 23(1):69–74.

Coltheart, M. (1981). Disorders of reading and their implications for models of normal reading. *Visible language*, 15(3):245.

Cook, V. and Bassetti, B., editors (2005). *Second Language Writing Systems*. Multilingual Matters.

Costard, S. (2011). *Störungen der Schriftsprache*. Thieme.

Courrieu, P. and De Falco, S. (1989). Segmental vs. dynamic analysis of letter shape by preschool children. *Cahiers de psychologie cognitive*, 9(2):189–198.

Cutler, A., W., A., Smits, R., and Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, 116:3668–3678.

Darwin, C. (1859). On the origins of species by means of natural selection. *London: Murray*.

Dearing, V. A. (1970). Computer aids to editing the text of dryden. In Gottesman, R. and Bennett, S., editors, *Art and Error: Modern Textual Editing*, pages 254–278. Methuen.

Dehmer, M. and Mehler, A. (2007). A new method of measuring the similarity for a special class of directed graphs. *Tatra Mountains Mathematical Publications*, 36:39–59.

Dehmer, M., Mehler, A., and Emmert-Streib, F. (2007). Graph-theoretical characterizations of generalized trees. In *Proceedings of the 2007 International Conference on Machine Learning: Models, Technologies & Applications (MLMTA '07), June 25-28, 2007, Las Vegas*, pages 113–117.

Dekker, R. H. and Middell, G. (2011). Computer-supported collation with collatex: Managing textual variance in an environment with varying requirements. *Supporting Digital Humanities*, pages 17–18.

den Hollander, A. (2004). How shock waves revealed successive contamination: A cardiogram of early sixteenth-century printed dutch bibles. In van Reenen, P., den Hollander, A., and van Mulken, M., editors, *Studies in Stemmatology II*, pages 99–112. John Benjamins.

Drogin, M. (1983). *Anathema!: medieval scribes and the history of book curses*. Medieval Scribes and Books. Allanheld, Osmun.

Dryer, M. S. and Haspelmath, M., editors (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Eagleton, C. and Spencer, M. (2006). Copying and conflation in geoffrey chaucer's treatise on the astrolabe: a stemmatic analysis using phylogenetic software. *Studies in History and Philosophy of Science Part A*, 37(2):237 – 268.

Efron, B. (1979). Bootstrap methods:Another look at the jackknife. *Annals of Statistics*, 7:1–26.

Ellison, J. W. (1957). *The use of electronic computers in the study of the Greek New Testament text*. Harvard University.

Erbse, H. (1959). Review. Paul Maas: Textkritik 3., verb. und verm. Auflage. Leipzig: Teubner 1957. 34 s. 2,30 dm. *Gnomon*, 31(2):97–103.

Felsenstein, J. (1978). The number of evolutionary trees. *Systematic Zoology*, 27(1):27–33.

Ferdoussi, A.-l.-Q. (1966-1967). *The Shahname - the book of kings*. The Great Islamic Encyclopaedia.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760):279–284.

Flight, C. (1990). How many stemmata? *Manuscripta*, 34(2):122–128.

Flight, C. (1992). Stemmatic theory and the analysis of complicated traditions. *Manuscripta*, 36(1):37–52.

Flight, C. (1994). A complete theoretical framework for stemmatic analysis. *Manuscripta*, 38(2):95–115.

Foley, J. (2002). *How to Read an Oral Poem*. University of Illinois Press.

Fourquet, J. (1946). Le paradoxe de Bédier. *Mélanges*, 1945(II):1–46.

Froger, J. (1968). *La critique des textes et son automatisation*, volume 7 of *Initiation aux nouveautés de la science*. Dunod.

Geyer, L. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, 22(5):487–490.

Gilmore, G., Hersh, H., Caramazza, A., and Griffin, J. (1979). Multidimensional letter similarity derived from recognition errors. *Perception & Psychophysics*, 25(5):425–431.

Gjessing, H. K. and Pierce, R. H. (1994). A stochastic model for the presence/absence of readings in niðrstigningar saga. *World Archaeology*, 26(2):268–294.

Greetham, D. C. (2010). *The Pleasures of Contamination: Evidence, Text, and Voice in Textual Studies*. Indiana University Press.

Greg, W. W. (1931). Recent theories of textual criticism. *Modern Philology*, 28(4):401–404.

Grier, J. (1989). Lachmann, Bédier and the bipartite stemma: towards a responsible application of the common-error method. *Revue d'histoire des textes*, 18(1988):263–278.

Griffith, J. (1984). A three-dimensional model for classifying arrays of manuscripts by cluster analysis. *Studia Patristica*, XV(I).

Griffith, J. G. (1968). A taxonomic study of the manuscript tradition of juvenal. *Museum Helveticum*, 25:101–138.

Guex, A. and Pithon, M. (1975). *Manuel de phonétique française*. Ecole de français moderne de l'Université, Lausanne.

Haigh, J. (1970). The recovery of the root of a tree. *Journal of Applied Probability*, 7(1):79–88.

Haigh, J. (1971). *Mathematics in the Archaeological and Historical Sciences*, chapter The manuscript linkage problem, pages 396–400. Edinburgh University Press, Scotland, UK.

Halonen, M. (2015). Computer-assisted stemmatology in studying paulus jusʼten's 16th-century chronicle catalogus et ordinaria successio episcoporum finlandensium. *Digital Scholarship in the Humanities : DSH*, 31(3):578–593.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29(2):147–160.

Harper Jr, C. W. (1976). Phylogenetic inference in paleontology. *Journal of Paleontology*, pages 180–193.

Haugen, O. E. (2002). The Spirit of Lachmann, the Spirit of Bédier: Old Norse Textual Editing in the Electronic Age. In *Annual Meeting of The Viking Society, University College London*, volume 8.

Haugen, O. E. (2010). Is stemmatology inherently dichotomous? On the silva portentosa of Old Norse stemmata. *Studia Stemmatologica*.

Haugen, O. E. (2015). The silva portentosa of stemmatology Bifurcation in the recension of Old Norse manuscripts. *Digital Scholarship in the Humanities*, 31(3):594–610.

Heikkillä, T. (2014). The possibilities and challenges of computer-assisted stemmatology: the example of vita et miracula s. symeonis treverensis. In Andrews, T. L. and Macé, C., editors, *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, pages 19–42. Brepols, Turnhout.

Hering, W. (1967). Zweispaltige Stemmata. *Philologus-Zeitschrift für antike Literatur und ihre Rezeption*, 111(1-2):170–185.

Hillis, D. M., Heath, T. A., and John, K. S. (2005). Analysis and visualization of tree space. *Systematic Biology*, 54(3):471–482.

Hockey, S. M. (1980). *A guide to computer applications in the humanities*. Johns Hopkins University Press.

Hoelzer, G. A. and Meinick, D. J. (1994). Patterns of speciation and limits to phylogenetic resolution. *Trends in ecology & evolution*, 9(3):104–107.

Hoenen, A. (2014a). Simulation of scribal letter substitution. In Andrews, T. L. and Macé, C., editors, *lectio 1*, pages 119–139. Brepols, Turnhout.

Hoenen, A. (2014b). Stemmatology, an interdisciplinary endeavour. In *Book of Abstracts zum DHd Workshop Informatik und die Digital Humanities*. DHd.

Hoenen, A. (2015a). Das artifizielle Manuskriptkorpus TASCFE. In *DHd 2015 - Von Daten zu Erkenntnissen - Book of abstracts*. DHd.

Hoenen, A. (2015b). Simulating misreading. In *Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems (NLDB)*.

Hoenen, A. (2016a). Das erste dynamische Stemma, Pionier des digitalen Zeitalters? In *DHd 2016 Konferenzabstracts*.

Hoenen, A. (2016b). Silva Portentosissima - Computer-Assisted Reflections on Bifurcativity in Stemmas. In *Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University*, pages 557–560.

Hoenen, A. (2017). Beyond the tree - a theoretical model of contamination and a software to generate multilingual stemmata. In *Book of Abstracts*, pages 155–159. AIUCD.

Hoenen, A. (2018). Multi Modal Distance - An Approach to Stemma Generation with Weighting. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, LREC 2018.

Hoenen, A., Eger, S., and Gehrke, R. (2017). How many stemmata with root degree k? In *Proceedings of the 15th Meeting on the Mathematics of Language*, pages 11–21.

Howe, C. J., Barbrook, A. C., Spencer, M., Robinson, P., Bordalejo, B., and Mooney, L. R. (2001). Manuscript evolution. *TRENDS in Genetics*, 17(3).

Howe, C. J. and Windram, H. F. (2011). Phylomemetics—evolutionary analysis beyond the gene. *PLoS Biol*, 9(5):e1001069.

Huson, D. H. (1998). Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73.

International-Phonetic-Association (1999). *Handbook of the International Phonetic Association*. Cambridge University Press.

Irigoin, J. (1954). Stemmas bifides et états de manuscrits. *Revue de Philologie, de Littérature et d'Histoire Anciennes*, 28:211.

Joshi, R. M. and Aaron, P. (2013). *Handbook of Orthography and Literacy*. Routledge.

Josuat-Vergès, M. (2015). Derivatives of the tree function. *The Ramanujan Journal*, 38(1):1–15.

Katz, L. and Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. *Haskins Laboratories Status Report on Speech Research*, SR-111:147–160.

Kelemen, E. (2009). *Textual Editing and Criticism: An Introduction*. WW Norton.

Kendall, D. G. (1948). On the generalized" birth-and-death" process. *The annals of mathematical statistics*, pages 1–15.

Keren, G. and Baggen, S. (1981). Recognition models of alphanumeric characters. *Perception & Psychophysics*, pages 234–246.

Kleinlogel, A. (1968). Das Stemmaproblem. *Philologus-Zeitschrift für antike Literatur und ihre Rezeption*, 112(1-2):63–82.

Knuth, D. E. (2005). The art of computer programming, volume 4: Generating all combinations and partitions, fascicle 3.

Kruskal, J. B. (1956). On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. In *Proceedings of the American Mathematical Society, 7*.

Lachmann, K. (1853). *In T. Lucretii Cari De rerum natura libros commentarius: Index*. Georg Reimer.

Lai, P.-H. (2012). *Information theoretic methods for biometrics, clustering, and stemmatology*. Washington University in St. Louis.

Lai, P.-H. and O'Sullivan, J. A. (2010). Mdl hierarchical clustering with incomplete data. In *Information Theory and Applications Workshop (ITA), 2010*, pages 1–5. IEEE.

Lai, P.-H., Roos, T., and O'Sullivan, J. A. (2010). Mdl hierarchical clustering for stemmatology. In *2010 IEEE International Symposium on Information Theory*, pages 1403–1407. IEEE.

Langosch, K., Micha, A., Avalle, D. S., Folena, G., Ineichen, G., Quaglio, A. E., Mengaldo, P. V., Steiger, A., Brunner, K., Neumann, F., et al. (1964). *Geschichte der Textüberlieferung der antiken und mittelalterlichen Literatur: Überlieferungsgeschichte der mittelalterlichen Literatur*. Atlantis.

Lantin, A.-C., Baret, P. V., and Macé, C. (2004). Phylogenetic analysis of gregory of nazianzus' homily 27. In Purnelle, G., Fairon, C., and Dister, A., editors, *Le poids des mots (JADT vol.1 et 2). Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, pages 700–707. Presses universitaires de Louvain, Louvain-la-Neuve.

Le Pouliquen, M. (2010). Filiation de manuscrits sanskrits et arbres phylogénétiques. *Mathématiques et sciences humaines. Mathematics and social sciences*, 192(4):57–91.

Le Pouliquen, M. and Csernel, M. (2010). Stemma codicum et relation d'intermédiarité, utilisation de la méthode de don quentin. In *Statistical Analysis of Textual Data Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles Lexicométrica*, volume 1, pages 309–320.

Lee, A. R. (1989). Numerical taxonomy revisited: John griffith, cladistic analysis and st. *Augustine's Quaestiones in Heptateuchem. Studia Patristica*, 20:24–32.

Lehtonen, A. (2013). *Handbook of Orthography and Literacy*, chapter Sources of Information Children Use in Learning to Spell: The Case of Finnish Geminates, pages 63–80. Routlegde.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Lima, M. (2014). *The book of trees: visualizing branches of knowledge*. Princeton Architectural Press New York.

Lin, Y.-J. (2016). *The Erotic Life of Manuscripts: New Testament Textual Criticism and the Biological Sciences*. Oxford University Press.

Lord, A. B. (1960). *The Singer of Tales*. Harvard University Press.

Lyytinen, H., Aro, M., Holopainen, L., Leiwo, M., Lyytinen, P., and Tolvanen, A. (2013). *Handbook of Orthography and Literacy*, chapter 4, pages 47–62. Routlegde.

Maas, P. (1927). *Textkritik*. Teubner.

Maas, P. (1937). Leitfehler und Stemmatische Typen. *Byzantinische Zeitschrift*, 37(2):289–294.

Maas, P. (1958). *Textual Criticism*. Clarendon Press.

Maas, P. (1960). *Textkritik*. 4. Auflage. Leipzig: Teubner.

Macé, C., Baret, P. V., and Lantin, A.-C. (2004). Philologie et phylogénétique: regards croisés en vue d'une édition critique d'une homélie de grégoire de nazianze. In Bozzi, A., Cignoni, L., and Lebrave, J.-L., editors, *Digital technology, philological disciplines*, pages 305 – 341. Istituti Editoriali e Poligrafici Internazionali.

Macé, C., Schmidt, T., and Weiler, J.-F. (2003). Le classement des manuscrits par la statistique et la phylogénétique : le cas de grégoire de nazianze et de basile le minime. *Revue d'histoire des textes*, 31(2001):241–273.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Marmerola, G. D., Oikawa, M. A., Dias, Z., Goldenstein, S., and Rocha, A. (2016). On the reconstruction of text phylogeny trees: Evaluation and analysis of textual relationships. *PloS one*, 11(12):e0167822.

Mazza, R. (2009). *Introduction to Information Visualization*. Springer.

Mehler, A. (2002). Text mining with the help of cohesion trees. In Gaul, W. and Ritter, G., editors, *Classification, Automation, and New Media. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation, March 15-17, 2000, Universität Passau*, pages 199–206, Berlin/New York. Springer.

Mehler, A. (2005). Lexical chaining as a source of text chaining. In Patrick, J. and Matthiessen, C., editors, *Proceedings of the 1st Computational Systemic Functional Grammar Conference, University of Sydney, Australia*, pages 12–21.

Mehler, A. (2009). Generalized shortest paths trees: A novel graph class applied to semiotic networks. In Dehmer, M. and Emmert-Streib, F., editors, *Analysis of Complex Networks: From Biology to Linguistics*, pages 175–220. Wiley-VCH, Weinheim.

Mehler, A. (2010). Minimum Spanning Markovian Trees: Introducing Context-Sensitivity into the Generation of Spanning Trees. In Dehmer, M., editor, *Structural Analysis of Complex Networks*, pages 381–401. Birkhäuser Publishing, Basel.

Mehler, A. (2011). Social Ontologies as Generalized Nearly Acyclic Directed Graphs: A Quantitative Graph Model of Social Ontologies by Example of Wikipedia. In Dehmer, M., Emmert-Streib, F., and Mehler, A., editors, *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*, pages 259–319. Birkhäuser, Boston/Basel.

Mehler, A., Abramov, O., and Diewald, N. (2011). Geography of Social Ontologies: Testing a Variant of the Sapir-Whorf Hypothesis in the Context of Wikipedia. *Computer Speech and Language*, 25(3):716–740.

Merivuori, T. and Roos, T. (2009). Some observations on the applicability of normalized compression distance to stemmatology. In *Proceedings of 2nd Workshop on Information Theoretic Methods in Science and Engineering*.

Miller, W. and Myers, E. W. (1985). A file comparison program. *Softw., Pract. Exper.*, 15(11):1025–1040.

Mink, G. (2004). Problems of a highly contaminated tradition: the new testament: Stemmata of variants as a source of a genealogy for witnesses. In van Reenen, P., den Hollander, A., and van Mulken, M., editors, *Studies in Stemmatology II*, pages 13–86. John Benjamins.

Miozzo, M. and Bastiani, P. D. (2002). The organization of letter-form representations in written spelling: Evidence from acquired dysgraphia. *Brain and Language*, 80(3):366 – 392.

Mooney, L. R., Barbrook, A. C., Howe, C. J., and Spencer, M. (2003). Stemmatic analysis of lydgate's "kings of england" : a test case for the application of

software developed for evolutionary biology to manuscript stemmatics. *Revue d'histoire des textes*, 31(2001):275–297.

Müller, S. and Weidemann, C. (2011). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica*.

Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L., and Zhou, Y. (2003). *TreeJuxtaposer: scalable tree comparison using Focus+ Context with guaranteed visibility*, volume 22(3), pages 453–462. ACM.

Najock, D. (1995). Letter distribution and authorship in early greek epics. *Revue informatique et Statistique dans les Sciences Humaines*, XXXI(1-4):129–154.

Najock, D. and Heyde, C. (1982). On the number of terminal vertices in certain random trees with an application to stemma construction in philology. *Journal of Applied Probability*, pages 675–680.

Nešetril, J. and Nešetrilová, H. (2012). The origins of minimal spanning tree algorithms–boruvka and jarnık. *Documenta Mathematica, vol. Extra Volume ISMP*, pages 127–141.

O'Grady, W., Dobrovolsky, M., and Katamba, F. (1997). *Contemporary Linguistics*. Longman, St. Martin's.

O'Hara, R. J. (1996). Trees of history in systematics and philology. *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano*, 27(1):81–88.

Oliaei, S. (2010). *L'art du conteur dans les cafés traditionnels en Iran*. Heinemann.

Ong, W. J. (2012). *Orality and Literacy*. Routledge.

Osborn, A. M. and Smith, C. J. (2005). *Molecular microbial ecology*. Garland Science.

O'Hara, R. J. (2006). Trees of history in systematics, historical linguistics, and stemmatics: A working interdisciplinary bibliography. *Historical Linguistics, and Stemmatics: A Working Interdisciplinary Bibliography (February 1, 2006)*.

Paradis, E. (2011). *Analysis of Phylogenetics and Evolution with R*. Springer Science & Business Media.

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290.

Pasquali, G. (1988). *Storia della tradizione e critica del testo*. Casa editrice Le lettere, Firenze.

Pasquali, G. and Pieraccioni, D. (1952). *Storia della tradizione e critica del testo*. Le Monnier.

Penny, D. and Hendy, M. (1985). The use of tree comparison metrics. *Systematic Zoology*, 34(1):75–82.

Phillips-Rodriguez, W. J., Howe, C. J., and Windram, H. F. (2009). Some considerations about bifurcation in diagrams representing the written transmission of the mahābhārata. *Wiener Zeitschrift für die Kunde Südasiens/Vienna Journal of South Asian Studies*, 52:29–43.

Platnick, N. I. and Cameron, H. D. (1977). Cladistic methods in textual, linguistic, and phylogenetic analysis. *Systematic Zoology*, 26(4):380–385.

Plotree, D. and Plotgram, D. (1989). Phylip-phylogeny inference package (version 3.2). *cladistics*, 5:163–166.

Poole, E. (1974). The computer in determining stemmatic relationships. *Computers and the Humanities*, 8(4):207–216.

Pouliquen, M. L. (2007). Using lattices for reconstructing stemma. In *Proceedings of the Fifth International Conference on Concept Lattices and Their Applications, CLA 2007, Montpellier, France, October 24-26, 2007*.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technology Journal*, 36:1389–1401.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series 3. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 45(3), pages 483–484. Cambridge Univ Press.

Quentin, H. (1926). *Essais de critique textuelle:(ecdotique)*. Picard.

Rayner, K., McConkie, G., and Zola, D. (1980). Integrating information across eye movements. *Cognitive Psychology*, 12:206–226.

Rayner, K., Pollatsek, A., Ashby, J., and jr. Clifton, C. (2012). *Psychology of Reading*. Psychology Press, New York/Hove.

Reynolds, L. and Wilson, N. (2013). *Scribes and Scholars, A Guide to the Transmission of Greek & Roman literatures*. Oxford University Press.

Robinson, P. (1996). Computer-assisted stemmatic analysis and 'best-text' historical editing. In van Reenen, P. and van Mulken, M., editors, *Studies in Stemmatology*, pages 71–104. John Benjamins.

Robinson, P. (2000). The one text and the many texts. *Literary and linguistic computing*, 15(1):5–14.

Robinson, P. (2015). Four rules for the application of phylogenetics in the analysis of textual traditions. *Digital Scholarship in the Humanities*, 31(3):637–651.

Robinson, P. and O'Hara, R. J. (1996). Cladistic Analysis of an Old Norse Manuscript Tradition. *Research in Humanities Computing (4)*.

Robinson, P. M. (1994). Collate: A program for interactive collation of large textual traditions. *Research in humanities computing*, 3:32–45.

Robinson, P. M. and O'Hara, R. J. (1992). Report on the textual criticism challenge 1991. *Bryn Mawr Classical Review*, 3(4):331–337.

Roelli, P. (2014a). Genealogical variant locations and simplified stemma: a test case. In Andrews, T. L. and Macé, C., editors, *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, pages 69–95. Brepols, Turnhout.

Roelli, P. (2014b). Petrus alfonsi, or on the mutual benefit of traditional and computerised stemmatology. In Andrews, T. L. and Macé, C., editors, *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, pages 43–68. Brepols, Turnhout.

Roelli, P. and Bachmann, D. (2010). Towards Generating a Stemma of Complicated Manuscript Traditions: Petrus Alfonsi's Dialogus. *Revue d'histoire des textes*, 5(4):307–321.

Roelli, P. and Macé, C. (2015). Parvum lexicon stemmatologicum. a brief lexicon of stemmatology.

Roos, T. and Heikkilä, T. (2009). Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*, 24:417–433.

Roos, T., Heikkilä, T., and Myllymäki, P. (2006). A compression-based method for stemmatic analysis. In *Proceedings of the 2006 Conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29 – September 1, 2006, Riva Del Garda, Italy*, pages 805–806, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Roos, T. and Zou, Y. (2011a). Analysis of textual variation by latent tree structures. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, pages 567–576.

Roos, T. and Zou, Y. (2011b). Analysis of textual variation by latent tree structures. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, pages 567–576.

Rubanovich, J. (2011). *Medieval Oral Literature*, chapter Orality in Medieval Persian Literature, pages 653–680. De Gruyter.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Salemans, B. (1996). Cladistics or the resurrection of the method of lachmann. In van Reenen, P. and van Mulken, M., editors, *Studies in Stemmatology*, pages 3–70. John Benjamins.

Salemans, B. (2000). *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian, Way: The Case of Fourteen Text Versions of Lanseloet Van Denemerken.* Nijmegen University Press.

Schmidt, D. and Colomb, R. (2009). A data structure for representing multi-version texts online. *International Journal of Human-Computer Studies*, 67(6):497–514.

Senoner, R. (1981). *Die römische Literatur.* C.H. Beck'sche.

Shanzer, D. (1986). Felix capella: Minus sensus qum nominis pecudalis. *Classical Philology*, 81(1):62–81.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Shaw, D. J. (1974). Mss: manuscript stemma simulator. *ALLC Bulletin*, 2(2):27–29.

Simson, W. J. (2006). *Die Geschichte der Aussprüche des Konfuzius (Lunyu)*, volume 10. Peter Lang.

Sinsheimer, J. S., Little, R. J., and Lake, J. A. (2012). Rooting gene trees without outgroups: Ep rooting. *Genome biology and evolution*, 4(8):821–831.

Slowinski, J. B. (2001). Molecular polytomies. *Molecular phylogenetics and evolution*, 19(1):114–120.

Snowling, M. and Hulme, C., editors (2014). *The Science of Reading: A Handbook (Handbooks of Developmental Psychology)*. Wiley-Blackwell, 1 edition.

Spada, E., Sagliocca, L., Sourdis, J., Garbuglia, A. R., Poggi, V., De Fusco, C., and Mele, A. (2004). Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis c virus infection. *Journal of clinical microbiology*, 42(9):4230–4236.

Spencer, M., Bordalejo, B., Robinson, P., and Howe, C. J. (2003a). How reliable is a stemma? an analysis of chaucer's miller's tale. *Literary and Linguistic computing*, 18(4):407–422.

Spencer, M., Bordalejo, B., Wang, L.-S., Barbrook, A. C., Mooney, L. R., Robinson, P., Warnow, T., and Howe, C. J. (2003b). Analyzing the order of items in manuscripts of the canterbury tales. *Computers and the Humanities*, 37(1):97–109.

Spencer, M., Davidson, E. A., Barbrook, A. C., and Howe, C. J. (2004a). Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology*, 227:503–511.

Spencer, M. and Howe, C. J. (2001). Estimating distances between manuscripts based on copying errors. *Literary and Linguistic Computing*, 16(4):467–484.

Spencer, M. and Howe, C. J. (2002). How accurate were scribes? a mathematical model. *Literary and linguistic computing*, 17(3):311–322.

Spencer, M., Mooney, L., Barbrook, A., Bordalejo, B., Howe, C., and Robinson, P. (2004b). The effects of weighting kinds of variants. In van Reenen, P., den Hollander, A., and van Mulken, M., editors, *Studies in Stemmatology II*, pages 227–240. John Benjamins.

Spencer, M., Wachtel, K., and Howe, C. J. (2002). The greek vorlage of the syra harclensis: A comparative study on method in exploring textual genealogy. *TC: A Journal of Biblical Textual Criticism*, 7:8–2.

Spencer, M., Wachtel, K., and Howe, C. J. (2004c). Representing multiple pathways of textual flow in the greek manuscripts of the letter of james using reduced median networks. *Computers and the Humanities*, 38(1):1–14.

Spencer, M., Windram, H. F., Barbrook, A. C., Davidson, E. A., and Howe, C. J. (2006). *Phylogenetic analysis of written traditions*, pages 67–74. McDonald Institute for Archaeological Research.

Sproat, R. W. (2000). *A computational theory of writing systems*. MIT Press.

Staden, R. (1979). A strategy of dna sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610.

Stolz, M. (2003). New philology and new phylogeny: Aspects of a critical electronic edition of wolfram's parzival. *Literary and linguistic computing*, 18(2):139–150.

Strasburger, H. (1977). *Umblick im Trümmerfeld der griechischen Geschichtsschreibung*, pages 3–57. Leuven University Press.

Stussi, A. (1994). *Introduzione agli studi di filologia italiana*. Il mulino.

Swofford, D. L. (1990). *PAUP: Phylogenetic Analysis Using Parsimony Version 3.0, May 1990*. Illinois Natural History Survey.

Taylor, I. and Taylor, M. M. (2014). *Writing and Literacy in Chinese, Korean and Japanese*. Number 14 in Studies in Written Language and Literacy. John Benjamins Publishing.

Taylor, W. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Teixeira, A. S., Monteiro, P. T., Carriço, J. A., Ramirez, M., and Francisco, A. P. (2015). Not seeing the forest for the trees: size of the minimum spanning trees (msts) forest and branch significance in mst-based phylogenetic analysis. *PloS one*, 10(3):e0119315.

Timpanaro, S. (2004). *La genesi del metodo del Lachmann*. UTET Università.

Timpanaro, S. (2005). *The Genesis of Lachmann's Method*. University of Chicago Press, Chicago.

Tov, E. (1982). Criteria for evaluating textual readings: The limitations of textual rules. *Harvard Theological Review*, 75(4):429–448.

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception and Psychophysics*, 9(1A):40–50.

Trovato, P. (2014). *Everything You Always Wanted to Know about Lachmann's Method, A Non-Standard Handbook of Genealogical Textual Criticism in the Age of Post-Structuralism, Cladistics, and Copy-Text*. libreriauniversitaria.it.

Trovato, P. and Guidi, V. (2004). Sugli stemmi bipartiti - decimazione, asimmetria e calcolo delle probabilità. *Filologia Italiana*, 1:9–48.

Tsatsaronis, G., Varlamis, I., and Vazirgiannis, M. (2008). Word sense disambiguation with semantic networks. In *Text, Speech and Dialogue*, pages 219–226. Springer.

Tukey, J. W. (1957). Variances of variance components: Ii. the unbalanced single classification. *The Annals of Mathematical Statistics*, pages 43–56.

Uttal, W. R. (1969). Masking of alphabetic character recognition by dynamic visual noise (dvn). *Perception & Psychophysics*, 6(2):121–128.

Van Berkel, A. (2005). *Second Language Writing Systems*, chapter The Role of Phonological Strategy in Learning to Spell, pages 97–121. Multilingual Matters.

van Reenen, P. T., den Hollander, A. A., and van Mulken, M. (2004). *Studies in Stemmatology II*. Studies in Stemmatology. John Benjamins Publishing Company.

van Reenen, P. T., van Mulken, M., and Dyk, J. (1996). *Studies in Stemmatology I*. Studies in Stemmatology. John Benjamins Publishing Company.

Vanek, K. (2007). *Ars corrigendi in der frühen Neuzeit: Studien zur Geschichte der Textkritik*, volume 4. Walter de Gruyter.

Wachtel, K. (2004). Kinds of variant in the manuscript tradition of the greek new testament. In van Reenen, P., den Hollander, A., and van Mulken, M., editors, *Studies in Stemmatology II*, pages 87–98. John Benjamins.

Wegner, P. (2006). *A Student's Guide to Textual Criticism of the Bible: Its History, Methods and Results*. InterVarsity Press.

Weitzman, M. P. (1982). Computer simulation of the development of manuscript traditions. *ALLC Bulletin. Association for Library and Linguistic Computing Bangor*, 10(2):55–59.

Weitzman, M. P. (1987). The evolution of manuscript traditions. *Journal of the Royal Statistical Society. Series A (General)*, pages 287–308.

West, M. L. (1973). *Textual Criticism and Editorial Technique: Applicable to Greek and Latin texts*. Teubner, Stuttgart.

Whitehead, F. and Pickford, C. E. (1951). The two-branch stemma. *Bulletin Bibliographique de la Société Internationale Arthurienne*, 3:83–90.

Whitehead, F. and Pickford, C. E. (1973). The introduction to the lai de l'ombre: Sixty years later. *Romania*, 94:145–56.

Wiik, K. (1965). *Finnish and English vowels*. Turun Yliopisto.

Wiley, R. W., Wilson, C., and Rapp, B. (2016). The effects of alphabet and expertise on letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, 42(8):1186–1203.

Windram, H. F., Charlston, T., and Howe, C. J. (2014). A phylogenetic analysis of orlando gibbons's prelude in g. *Early Music*, 42(4):515–528.

Windram, H. F., Shaw, P., Robinson, P., and Howe, C. J. (2008). Dante's monarchia as a test case for the use of phylogenetic methods in stemmatic analysis. *Literary and linguistic computing*, 23(4):443–463.

Woerther, F. and Khonsari, H. (2003). L'application des programmes de reconstruction phylogénétique sur ordinateur à l'étude de la traduction manuscrite d'un texte: l'exemple du chapitre xi de l'ars rhetorica du pseudo-denys d'halicarnasse. *Revue d'histoire des textes*, 31(2001):227–240.

Wright, P. (1997). Counting and constructing minimal spanning trees. *Bulletin of the Institute of Combinatorics and its Applications*, 21:65–76.

Yamada, T., Kataoka, S., and Watanabe, K. (2010). Listing all the minimum spanning trees in an undirected graph. *International Journal of Computer Mathematics*, 87(14):3175–3185.

Yamamoto, K. (2003). *The Oral Background of Persian Epics*. Brill, Leiden.

Yorav, A., Dagan, T., and Graur, D. (2005). An exploratory study on the use of a phylogenetic algorithm in the reconstruction of stemmata of halachic texts. *Hebrew Union College Annual*, pages 273–288.

Zarri, G. P. (1976). A computer model for textual criticism? *The Computer in Literary and Linguistic Studies, Jones, A. and Churchhouse, RF, eds., Cardiff*, pages 133–155.

Zipf, George, K. (1949). Human behavior and the principle of least effort.

Appendices

I. Figures & Tables

I.1 Figures (including Images)

Page	Numbering	Figure	Source if Image
10	1.1	First Stemma by Schlyter 1827	O'Hara (1996)
13	1.2	Bifurcativity Illustration	author
17	1.3	Types of stemmata for 3 labelled nodes, as in Maas (1960)	author
19-21	1.4-1.6	Root of Stemmata, argument by Fourquet (1946)	author
37	1.7	Variant graph from `stemmaweb.net`	Andrews (2014)
39	1.8	Avestan stemma	`ada.usal.es/videvdad/manuscripts.htm`
39	1.9	Innovative stemma	`nttextualcriticism.blogspot.de/2010/12/variations-in-genealogical-stemma.html`
40	1.10	Typical stemma	`it.wikipedia.org/wiki/Stemma_codicum`
40	1.11	Phylegenetic stemma	Howe et al. (2001)
41	1.12	Cladogram	`en.wikipedia.org/wiki/Cladogram`
42	1.13	Neighbour Nets	Bryant and Moulton (2004)
45	1.14	Circular Tree Map with Map Underlay	`https://www.mirrorservice.org/sites/gutenberg.org/3/2/6/2/32624/32624-h/images/illus-001.png`
46	1.15	DynStemGen User Dialogue	author
46/47/48	1.16-1.21	DynStemGen Example Slides	author
73	2.1	Example tree with loss probabilities	author
75	2.2	k-furcations diagram	author
77/78	2.3/2.4	Root uni-/bi- & bifurcations diagrams	author
79	2.5	Root uni-/bi- & bifurcations diagram, small large tradition	author
80	2.6	Root uni-/bi- & bifurcations diagram, loss conditions	author
81	2.7	Stemma from Shanzer (1986)	Shanzer (1986)
91	3.1	Variation types from Andrews and Macé (2013)	Andrews and Macé (2013)
104	3.2	Correct stemmata of the 4 TASCFE versions	author
111	3.3	Copying event from afsar to achtar	author
118	4.1	Multi Modal Distance toy tradition	author
123	4.2	Vowel diagram	author, modified and expanded from LaTeX code in `http://tex.stackexchange.com/questions/156955/tikz-pgf-linguistics-vowel-chart`
141	4.3	Consensus graph all MSTs	author
171	5.1	Unrooted labelled tree topologies	`https://en.wikipedia.org/wiki/Cayley%27s_formula`
172	5.2	Topologies for 2 surviving nodes	author

I.2 Tables

II. Abbreviations

ms - Manuscript
p - Probability
sd - Standard deviation
ALLC - Association for Literary and Linguistic Computing
ASD - Average Sign Distance
ATC - Archetype Text Congruency
AVC - Archetype Variant Congruency
CRT - Cathode Ray Tube
DNA - Desoxyribonucleic Acid
Dr. - Doktor
DynStemGen - Dynamic Stemma Generator
EADH - European Association for Digital Humanities
ERP - Event Related Potentials
HIWI - Hilfswissenschaftler = student assistant
HTML - Hyper Text Markup Language
IPA - International Phonetic Alphabet
LB - lectio brevior
LD - lectio difficilior

LF - leitfehler
MA - Majority Archetype
ML - Maximum Likelihood
MMD - Multi Modal Distance
MR - Majority Reconstruction
MW - Maximally Wrong Archetype
MST - Minimum Spanning Tree
NJ - Neighbour Joining (algorithm)
OFT - Oral Formulaic Theory
PAML - Phylogenetic Analysis by Maximum Likelihood
Prof. - Professor
PF - Position Faithfulness
RA - Random Archetype
ReAV - Recurrence Analytics Visualization
RHM - an algorithm, name based on the names of the inventors (Roos, Heikkilä, Myllymäki)
TASCFE - Tehran Artificial Shahname Corpus with Frankfurt Extension
TEI - Text Encoding Initiative
TTLab - Text Technology Laboratory, Uni Frankfurt
TTR - Type Token Ratio
VVR - Versetype-Verse Ratio
WB - word-based: refers to pairwise token comparison
XML - Extensible Markup Language

IV. Appendices on Trees

Proof through contradiction of the formula for Root Bifurcativity in Arbres

A way to proof the formula would be through a combinatiorial proof or contradiction, showing that the formula is a bijective projection of the desired set. The to be proven sentence is:

By the aforedescribed procedure, exactly and only all root bifurcating trees over V have been produced/counted exactly once. The equation determines thus the proportion of root bifurcating labelled rooted trees for n nodes.

If the procedure would not induce exactly and only all possible root bifurcating trees over V, in other words, if the relation was not bijective, then there must be either:

1. a root bifurcating tree T_b, which has not been generated by the aforedescribed procedure
2. a tree T_b, which has been generated at least twice
3. a tree T_b, which is not a root bifurcating labelled tree over $|V|$ which has been counted

Reductio ad absurdum:

1. a rooted root bifurcating labelled tree T_b over V exists, which has not been generated:
   (a) Since T_b is a root bifurcating tree, cutting the edges from ν to its two children, three components of T_b must emerge: the root ν, the non-empty subset a and the non empty subset b. a and b together must be of sizes $|a| + |b| = |V \setminus \nu|$ and range in size from 1 to $|V| - 2$.
   (b) Since any unordered partition of $V \setminus \nu$ has been generated in the procedure above; since for any of those partitions any possible tree has been generated, T_b must have been generated.
   (c) THEN, from a and b follows that no root bifurcating rooted labelled tree T_b can exist, which has not been produced.
2. A tree T_b exists, which has been counted at least twice
   (a) Each root has been used separately, no trees of two different roots can be equal.
   (b) The procedure generates all possible unordered partitions into two. Each partition induces different subsets and since the sets are unordered, each partition is unique. Then, no partition can contain a tree, counted by another partition.
   (c) Within each partition $\binom{n}{k}$ generates only different sets (binomial coefficient). Then, within each partition all subset vertex combinations are unique. Each subset combination is unique and none occurrs twice (a,b).
   (d) Any such subset induces only different trees (rooted labelled), apply-

ing n^{n-1}, see Harper Jr (1976).

(e) Combining any tree of subset a once with any tree of subset b, each combination of trees is unique.

(f) THEN, from a to e follows, that no tree can have been produced twice.

3. There is a tree T_b which is not a rooted labelled root bifurcating tree over V, which has been counted

(a) Choosing a root, all produced trees are rooted; no unrooted tree is produced

(b) Any produced tree is composed of a root ν, and two subsets a, b, which are constituting the set $V \setminus \nu$. Then, no produced tree is not a tree over V.

(c) All nodes are labelled, there is no unlabelled tree, which has been produced.

(d) To all possible roots, exactly two rooted subtrees have been attached. Then no non root bifurcating tree has been generated.

(e) THEN, from a to d it follows that no tree T_b can have been produced, which is not a root bifurcating rooted labelled tree over V.

From 1. to 3. follows that only any possible root bifurcating tree has been generated/counted exactly once. q.e.d.

Examples and visual counts

Starting from the equation for root bifurcating arbres, we can set $n = 3$ and get (details omitted):

$$\frac{3 * \frac{\binom{2}{1} * 1^0 * 1^0}{2}}{3^2} = \frac{3}{9}$$

which we can show to be true, displaying all possible trees, see Figure 5.1.

Likewise, if we set $n = 4$:

$$\frac{4 * \frac{\sum_{k=1}^{2} \binom{3}{k} * (k^{k-1} * (3-k)^{2-k})}{2}}{4^3}; n > 2$$

which is:

$$\frac{4 * \frac{\binom{3}{1} * (1^0 * 2^1) + \binom{3}{2} * (2^1 * 1^0)}{2}}{4^3} = \frac{24}{64}$$

which can equally be shown to be true by displaying the possible trees, see Figure 5.1.

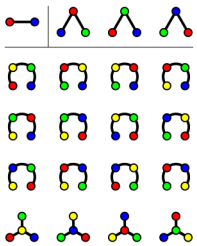


Figure 5.1: Depicted are all unrooted labelled trees for n = 2,3,4. Image from the English Wikipedia.

It suffices for each of the displayed possibilities, to take each node as root and check for root bifurcation. One immediately sees, that only rooting the tree at the middle node (1 in 3) results in a root bifurcation. Thus, in consequence, the proportion of root bifurcations is 3 out of 9. For 4 nodes, in each of the upper three rows, the topology is equal and thus in 12 cases 2 nodes would produce a root bifurcation (internodes), whilst 2 (terminal nodes) would not. In the last row, rooting in any terminal node results in a root unifurcation, rooting in the internode results in a root trifurcation. Thus out of $16 * 4 = 64$ possible rooted labelled trees, only $12 * 2 = 24$ are root bifurcating as expressed through the formula. Illustration from `https://en.wikipedia.org/wiki/Cayley\%27s_formula`.

Root bifurcating Greg trees

The formula builds on the assumption that all root bifurcating Greg trees are roots to which two rooted subtrees are being attached. These are furthermore taken to be themselves Greg trees. In any tree of $m + n$ nodes, Greg trees are those, where the unlabelled nodes must be of degree 3, or if root 2. All other n,m trees are no Greg trees (which implies for instance that a single labelled node is, a single unlabelled node is not a Greg tree). Now, since the formula counts only Greg trees for the subsets, the only way in which the whole conglomerate tree would not be a Greg tree, would be through a non conformancy arising from the attachement to root and/or root. Root can be labelled or unlabelled. If labelled, it can take on any degree, certainly 2, so attaching 2 Greg trees must result in a Greg tree. The root of subset 1 can be labelled or unlabelled. If labelled, the degree does not matter, if unlabelled then it must have at least 2 children, since otherwise the

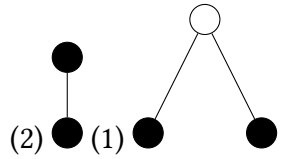


Figure 5.2: The unlabelled rooted topologies of possible stemmas for two surviving manuscripts as thought of by Maas (1960). White nodes symbolize reconstructed, lost manuscripts (unlabelled), whereas black nodes are survivors (to be labelled). The number in brackets refers to the number of possible labelled trees for each topology.

subset tree would be no Greg tree and not produced. Then attaching to root, the node would have degree at least 3 if unlabelled, which fulfills the requirement of being a Greg tree. The same goes for subset 2. So there can be no tree produced, which is not a Greg tree. Did we count all root bifurcating Greg trees? Since for each possible root, for each case (rooted, unrooted) for each possible partition of the subsets into labelled and unlabelled nodes and for each permutation of the labelled ones, we produces the trees, there is no possible tree not produced.

Examples and visual counts

Starting from the formula for root bifurcating Greg trees, we can set $m = 2$ and get:

$$\sum_{n=0}^{1}\left(\frac{\sum_{k=0}^{n}\sum_{l=0}^{1}\binom{1}{l} * ((g^*(l,k) * g^*(1-l,n-k)) + (g^*(l,n-k) * g^*(1-l,k)))}{2} + \frac{\sum_{k=0}^{n-1}\sum_{l=0}^{2}\binom{2}{l} * ((g^*(l,k) * g^*(2-l,n-1-k)) + (g^*(l,n-1-k) * g^*(2-l,k)))}{4}\right)$$

Thus, we have to generate all possible values of k and combine them with all possible values of l, but since for all cases in which either $n >= m$ or $m, n < 0$ there are no Greg trees by definition, some of the factors become 0.

We can have values for $n = 0$ and $n = 1$: For $n = 0$, there are no cases in which root is unlabelled and otherwise only the case of $k = 0$:

$$\frac{(\binom{1}{0} * ((g^*(0,0) * g^*(1,0)) + (g^*(0,0) * g^*(1,0))))}{2}$$
$$+ \frac{(\binom{1}{1} * ((g^*(1,0) * g^*(0,0)) + (g^*(1,0) * g^*(0,0))))}{2}$$

but, since $g^*(0,0)$ is 0, the whole case of $n = 0$ contributes 0 cases. For $n = 1$, we have

$$\sum_{k=0}^{1} \frac{\sum_{l=0}^{1} \binom{1}{l} * ((g^*(l,k) * g^*(1-l,1-k)) + (g^*(l,1-k) * g^*(1-l,k)))}{2}$$
$$+ \frac{\sum_{l=0}^{2} \binom{2}{l} * ((g^*(l,0) * g^*(2-l,0)) + (g^*(l,0) * g^*(2-l,0)))}{4}$$

Here, for the cases with root unlabelled:

$$\begin{aligned} &\frac{(\binom{2}{0} * ((g^*(0,0) * g^*(2,0)) + (g^*(0,0) * g^*(2,0))))}{4} \\ &+ \frac{(\binom{2}{1} * ((g^*(1,0) * g^*(1,0)) + (g^*(1,0) * g^*(1,0))))}{4} \\ &+ \frac{(\binom{2}{2} * ((g^*(2,0) * g^*(0,0)) + (g^*(2,0) * g^*(0,0))))}{4} \end{aligned} \tag{5.1}$$

Again, since $g^*(0,0)$ is 0, only the case of $l = 1$ remains and the expression becomes 1. This is exactly the one tree with an unlabelled root depicted in Figure 5.2. For the cases with root labelled, we get:

$$\frac{\sum_{l=0}^{1} \binom{1}{l} * ((g^*(l,0) * g^*(1-l,1)) + (g^*(l,1) * g^*(1-l,0)))}{2}$$
$$+ \frac{\sum_{l=0}^{1} \binom{1}{l} * ((g^*(l,1) * g^*(1-l,0)) + (g^*(l,0) * g^*(1-l,1)))}{2}$$

However, in all cases of l, regardless of the value of k, we produce a $g^*(0,0)$ on the one and a $g^*(0,1)$ tree on the other side of the sums; both equal 0. This expression is thus 0 and we are left with the overall result of 1, which we know to be true from a look at Figure 5.2.

With $m = 3$ survivors:

n	k	l	root labelled	Res.	root unlabelled	Res.
0	0	0	$\binom{2}{0} * ((g^*(0,0) * g^*(2,0)) + (g^*(0,0) * g^*(2,0)))$	0	-	-
0	0	1	$\binom{2}{1} * ((g^*(1,0) * g^*(1,0)) + (g^*(1,0) * g^*(1,0)))$	4	-	-
0	0	2	$\binom{2}{2} * ((g^*(2,0) * g^*(0,0)) + (g^*(2,0) * g^*(0,0)))$	0	-	-
0	0	3	-	-	-	-
1	0	0	$\binom{2}{0} * ((g^*(0,0) * g^*(2,1)) + (g^*(0,1) * g^*(2,0)))$	0	$\binom{3}{0} * ((g^*(0,0) * g^*(3,0)) + (g^*(0,0) * g^*(3,0)))$	0
1	0	1	$\binom{2}{1} * ((g^*(1,0) * g^*(1,1)) + (g^*(1,1) * g^*(1,0)))$	0	$\binom{3}{1} * ((g^*(1,0) * g^*(2,0)) + (g^*(1,0) * g^*(2,0)))$	12
1	0	2	$\binom{2}{2} * ((g^*(2,0) * g^*(0,1)) + (g^*(2,1) * g^*(0,0)))$	0	$\binom{3}{2} * ((g^*(2,0) * g^*(1,0)) + (g^*(2,0) * g^*(1,0)))$	12
1	0	3	-	-	$\binom{3}{3} * ((g^*(3,0) * g^*(0,0)) + (g^*(3,0) * g^*(0,0)))$	0
1	1	0	$\binom{2}{0} * ((g^*(0,1) * g^*(2,0)) + (g^*(0,0) * g^*(2,1)))$	0	-	-
1	1	1	$\binom{2}{1} * ((g^*(1,1) * g^*(1,0)) + (g^*(1,0) * g^*(1,1)))$	0	-	-
1	1	2	$\binom{2}{2} * ((g^*(2,1) * g^*(0,0)) + (g^*(2,0) * g^*(0,1)))$	0	-	-
1	1	3	-	-	-	-
2	0	0	$\binom{2}{0} * ((g^*(0,0) * g^*(2,2)) + (g^*(0,2) * g^*(2,0)))$	0	$\binom{3}{0} * ((g^*(0,0) * g^*(3,1)) + (g^*(0,1) * g^*(3,0)))$	0
2	0	1	$\binom{2}{1} * ((g^*(1,0) * g^*(1,2)) + (g^*(1,2) * g^*(1,0)))$	0	$\binom{3}{1} * ((g^*(1,0) * g^*(2,1)) + (g^*(1,1) * g^*(2,0)))$	3
2	0	2	$\binom{2}{2} * ((g^*(2,0) * g^*(0,2)) + (g^*(2,2) * g^*(0,0)))$	0	$\binom{3}{2} * ((g^*(2,0) * g^*(1,1)) + (g^*(2,1) * g^*(1,0)))$	3
2	0	3	-	-	$\binom{3}{3} * ((g^*(3,0) * g^*(0,1)) + (g^*(3,1) * g^*(0,0)))$	0
2	1	0	$\binom{2}{0} * ((g^*(0,1) * g^*(2,1)) + (g^*(0,1) * g^*(2,1)))$	0	$\binom{3}{0} * ((g^*(0,1) * g^*(3,0)) + (g^*(0,0) * g^*(3,1)))$	0
2	1	1	$\binom{2}{1} * ((g^*(1,1) * g^*(1,1)) + (g^*(1,1) * g^*(1,1)))$	0	$\binom{3}{1} * ((g^*(1,1) * g^*(2,0)) + (g^*(1,0) * g^*(2,1)))$	3
2	1	2	$\binom{2}{2} * ((g^*(2,1) * g^*(0,1)) + (g^*(2,1) * g^*(0,1)))$	0	$\binom{3}{2} * ((g^*(2,1) * g^*(1,0)) + (g^*(2,0) * g^*(1,1)))$	3
2	1	3	-	-	$\binom{3}{3} * ((g^*(3,1) * g^*(0,0)) + (g^*(3,0) * g^*(0,1)))$	0
2	2	0	$\binom{2}{0} * ((g^*(0,2) * g^*(2,0)) + (g^*(0,0) * g^*(2,2)))$	0	-	-
2	2	1	$\binom{2}{1} * ((g^*(1,2) * g^*(1,0)) + (g^*(1,0) * g^*(1,2)))$	0	-	-
2	2	2	$\binom{2}{2} * ((g^*(2,2) * g^*(0,0)) + (g^*(2,0) * g^*(0,2)))$	0	-	-
2	2	3	-	-	-	-

Table 5.1: Tabled addends for different n,k,l.

$$\sum_{n=0}^{2}(3 * \frac{\sum_{k=0}^{n}\frac{\sum_{l=0}^{2}\binom{2}{l}*((g^*(l,k)*g^*(2-l,n-k))+(g^*(l,n-k)*g^*(2-l,k)))}{2}}{2} + \frac{\sum_{k=0}^{n-1}\frac{\sum_{l=0}^{3}\binom{3}{l}*((g^*(l,k)*g^*(3-l,n-1-k))+(g^*(l,n-1-k)*g^*(3-l,k)))}{2}}{2})$$

For convenience, all addends with all combinations of n, k and l are listed in Table 5.1.

The 4 possibilities, we count for $n = 0$ represent the same tree type 4 times leaving one topology, which can have 3 different roots, thus in total for $n = 0$, there are 3 possible root bifurcating rooted Greg trees. Other possibilities, where the root is labelled do not exist. For an unlabelled root, at $n = 1$, we have then 24 trees, with each tree having 4 equivalents, ending with 6 root bifurcating Greg trees for 3 survivors and 1 unlabelled node. For $n = 2$, we count 12 trees represented 4 times, thus 3 distinct root bifurcating rooted Greg trees for 3 survivors and 2 internodes. In sum, we count 12 root bifurcating rooted Greg trees, which is consistent in both composition and count with Figure 1.3. From this one can

see, that a simpler and more general formalisation is possible. The generalization to k-furcations can be achieved by enumerating all partitions into k non m-empty subsets. The m nodes can then be distributed among the sets, while the table of (n,m)-Greg trees provides the possible combinations for those nodes.