Lachmannian Archetype Reconstruction for Ancient Manuscript Corpora

Armin Hoenen
Goethe University
Robert- Mayer Strasse 10
60486 Frankfurt, Germany
hoenen@em.uni-frankfurt.de

Abstract

Two goals are targeted by computer philology for ancient manuscript corpora: firstly, making an edition, that is roughly speaking one text version representing the whole corpus, which contains variety induced through copy errors and other processes and secondly, producing a stemma. A stemma is a graph-based visualization of the copy history with manuscripts as nodes and copy events as edges. Its root, the so-called archetype, is the supposed original text or *urtext* from which all subsequent copies are made. Our main contribution is to present one of the first computational approaches to automatic archetype reconstruction and to introduce the first text-based evaluation for automatically produced archetypes. We compare a philologically generated archetype with one generated by bio-informatic software.

1 Introduction

In philology, oftentimes more than one single manuscript of the same *tradition* (that is, the same text) has survived. These manuscripts often differ in their wording in various places since copy errors, corrections, and other processes have led to deviation from the original text. This causes two problems: uncertainty about the original wording and uncertainty about which manuscript has been copied from which other.

The reconstruction of the copy history of manuscript texts is largely similar to that of DNA, which is why phylogenetic approaches have been adopted (Robinson and O'Hara, 1996; Robinson et al., 1998; van Reenen et al., 1996; van Reenen et al., 2004; Spencer et al., 2004; Roos and Heikkilä, 2009; Roelli and Bachmann, 2010; Andrews and Macé, 2013). However, the main goal of the philological work on ancient manuscripts is not the reconstruction of the copy history but compiling an edition of a historical text. This entails reducing the variation encountered so that one single main text as the prototypical representation of the manuscript corpus emerges. Ideally, this text is believed to be the author's original or closest possible to it. Two model approaches to making an edition are most widespread. The earlier one by Lachmann (see for instance Lachmann (1853)) opts for reconstructing an urtext actively, that is if needed by means of *emendation*, which refers to inferring the original (authorial) wording from the extant variants even if the so-inferred form is not itself extant and thus not attested in any of the manuscripts. The later approach after Bédier (see for instance Bédier (1928)) bases the edition directly on the text of the best available manuscript.

In this paper, we will present a first automatic implementation of the earlier approach. Additionally, an algorithm for evaluation of a so-reconstructed text will be presented and applied to artificial benchmark data sets (gold standard). First, we will introduce the data sets. Then, two methods, rule-based (using philological principles) and statistical (likelihood-based using bio-informatic software) will be explained in detail before the results are being presented, followed by a general discussion, a field specific discussion, and a conclusion and outlook.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1209–1214,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

Text	Lang	MS	Tok
Parzival	English	21	957
Notre Besoin	French	13	1029

Table 1: The artificial traditions. MS = number of manuscripts, Tok = number of tokens, Lang = language

2 Artificial Traditions

An *artificial tradition* is a fully digitized set of manuscripts which have been produced through manual copying in recent times whilst recording the true copy history/genealogical relationships. Three of these corpora have been published to date, *Parzival* by (Spencer et al., 2004), *Notre Besoin* by (Baret et al., 2004), and *Heinrichi* by (Roos and Heikkilä, 2009). They are provided in a fully word-aligned tabular version by the authors, so that collation must not be performed anymore. We excluded the Heinrichi tradition from computation as Old Finnish data could not be interpreted by us. Table 1 displays the composition of the artificial traditions we used.

3 Method

Contrasting a rule-based and a statistical approach, we automatically reconstruct the archetype text for the two aforementioned traditions. To the best of our knowledge, this paper treats automatic archetype reconstruction in connection with evaluation for the first time and applies suitable bio-informatic programs for the first time. While the transfer of biological software to philology, especially in stemmatics, is done since the 1990ies, (O'Hara, 1996), purely philological automatic reconstructions are a recent development. As input for both reconstruction methods, we use the same tree, generated by bio-informatic software. Corresponding to this tree, we reconstruct the archetype using a) a philological rule-based majority-vote bottom-up algorithm and b) statistical bio-informatic software.

3.1 Philological Reconstruction

A lost manuscript text can be reconstructed in different ways given a precomputed stemma. The key question is how to resolve variation. Depending on a concrete decision rule for disambiguation of variation among the direct descendents of a lost manuscript node, there are several possible reconstructions, which means one stemma can correspond to several possible archetype texts (depending on the decision rule). We implement the majority-vote decision rule as referred to frequently in philological discourse, see for instance (West, 1973). The algorithm simply assigns the most frequent variant of all direct descendents. If more than one variant is most frequent, our reconstruction retains all for later disambiguation either through majority-vote in subsequent recursion steps or for manual disambiguation through the expert if more than one variants end undisambiguated in the reconstructed archetype. Lost text of leaf nodes should be pruned since their texts are more corrupt than the one of their ancestors, thus unnecessarily corrupting the tradition. This pruning roughly parallels the philological practice of *recensio*, the prior exclusion of uninformative manuscripts.

In mathematical terms, let S be a rooted stemmatic tree (directed acyclic) consisting of a set of vertices V, a set of edges E, and one root node v_0; Each vertex has an indegree of 1, only v_0 has no incoming edges. Let each vertex v_i be associated with a text $T_i \in T_{all}$, which is its textual content, $|V| = |T_{all}|$. While T_{all} is the set of all texts that really existed in the tradition, T' initially is the set of surviving texts, $T' \subseteq T_{all}$. The texts in T_{all} are aligned, that is each text consists of a sequence of tokens or reconstructed tokens $\{T_{i_i}, .., T_{i_n}\}$ and all T_i have the same length. A recontructed token can be a set of to be disambiguated tokens. For each $T_j \in T_{all} \setminus T'$, we reconstruct the lost text in the following way:

1. Collect all texts $\{T_k, .., T_m\}, 1 \leq (k, m) \leq |T_{all}| - 1; k, m \neq j$ associated with the vertices $\{v_k, .., v_m\}$ which are direct descendents of vertex v_j, which is associated with T_j. If one of the texts in $\{T_k, .., T_m\}$ is itself not in T', delay the actual reconstruction and start a new reconstruction for the next unreconstructed text until all texts $\{T_k, .., T_m\}$ are reconstructed

2. Text reconstruction for each token T_{j_i}: $T_{j_i} = majority(\{T_{k_i}, .., T_{m_i}\})$; in case $|majority(\{T_{k_i}, .., T_{m_i}\})| > 1$ assign all variants to T_j (using a separator), if one of $\{T_{k_i}, .., T_{m_i}\}$ does already carry multiple variants, treat each one as one variant

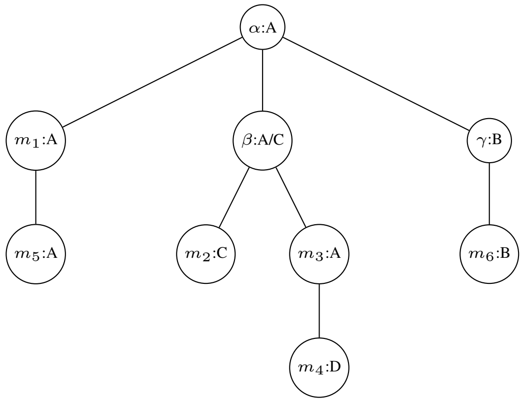

Figure 1: A simple stemma of a one word tradition, Greek letters denoting lost (hype)archetype(s). The Roman upper case letters refer to the observed variant (manuscripts m_i) or to the reconstructed variant ((hype)archetype(s)) inferred bottom-up.

m_1	m_2	m_3	m_4	**Variants**
this	t'	it	dis	A-D
is	is	is	is	A
a	an	a		A-C
text	text	text	text	A
AAAA	BABA	CAAA	DACA	**PseudoDNA**

Table 2: Conversion of a word-aligned example tradition to a pseudo-DNA.

and compute the majority variant (example: $\{T_{2_i}, T_{5_i}\} = \{this/tis, tis\}$ then $majority(\{T_{2_i}, T_{5_i}\}) = tis$)

3. $T' = T' \cup T_j$; if $T' = T_{all}$ end else start next reconstruction

For an example see Figure 1.

3.2 Bio-informatic Reconstruction

In bio-informatics, reconstruction of ancestral genomes has been undertaken. The most famous case is presumably that of the mammoths, (Miller et al., 2008). However, bacterial asexual reproduction is generally more similar to manuscript copying than mammalian sexual reproduction. Therefore and for other methodological reasons, it is more appropriate to transfer bacteriological reconstruction to philology. Bacteriologists sucessfully reconstructed the predecessors of yeast, (Voordeckers et al., 2012). Given a pregenerated tree, they used Bayesian marginal and joint probabilities to generate the sequences best explaining the tree. A program performing this is PAML, (Yang, 2007).

We converted the already word-aligned traditions into sequences of letters for each manuscript, see Table 2. Each letter encodes the variant the current manuscript carries at the current position. For the so-produced pseudo-DNA sequences, the PAML software generates a stemmatic tree using the maximum likelihood (ML) criterion without resolving variation at inferred internodes. This ML tree or any other can be used as input to generate ancestral sequences at the internodes and the root node using an optimization with a) marginal probabilities, for details see (Yang et al., 1995) and b) an optimization with joint probabilities, for details see (Pupko et al., 2000). Yang (2007) states that the results of a) and b) differ only in borderline cases. Indeed we found, that the respectively produced archetypes were identical in our case.

The ML tree is usually bifurcating, generating many internodes and an extra-corporal root. Although in our context, it challenges performance of automatic reconstruction, bifurcations are not a circumstance necessarily paralleling philology closely, (Howe et al., 2012).

4 Evaluation

For each position of the alignment, a produced archetype (PA) is compared to the original archetype present in the benchmark data sets. Whenever the PA has at least one variant at the current position corresponding to the archetype, an agreement score is incremented by one divided through the number of current variants in the PA. This assumes implicitly that manual disambiguation is at random and represents therefore a baseline evaluation. The agreement score is divided by the number of positions in the alignment to give the total precision of the PA text. This evaluation is called whole text evaluation (WTE). It serves as an orientation point towards the overall reliability of the reconstructed text as a whole. A second evaluation concerns the proportion of correctly disambiguated positions of variation. That agreement score is produced in the same way as described above, but only for positions where the corpus had variation. This evaluation is called po-

Tradition	**WTE**	**PVE**	**ASD**
Parzival(PAML)	0.91	0.73	51.38
Parzival(MV)	0.96	0.88	
Notre Besoin(PAML)	0.95	0.9	54.95
Notre Besoin(MV)	0.97	0.94	

Table 3: Evaluation of the archetypes by PAML and majority-vote (MV). ML trees evaluated with ASD.

sition of variation evaluation (PVE). Formally, both WTE and PVE can be represented by

$$\frac{\sum_{i=0}^{n} \frac{\mathbb{1}_{C_i}(a_i)}{|C_i|}}{n}, \mathbb{1}_{C_i}(a_i) = \begin{cases} 1 & \text{if } a_i \in C_i \\ 0 & \text{if } a_i \notin C_i \end{cases} \tag{1}$$

where n is either the number of words in the alignment (WTE) or the number of positions of variation (PVE), a is the archetype, a_i the i-th token in the archetype, and C_i the set of variants of the PA at the i-th position.

5 Results

We evaluated the PAs (rule-based and statistical) with WTE and PVE and additionally evaluated the initial ML trees against the true stemma by means of the graph-based Average Sign Distance (ASD),[1] a measure of distance of genealogical trees using vertex triple distances as described in (Roos and Heikkilä, 2009). In order to achieve this, we converted the stemmas automatically from the Newick output format by PAML into the required format of the ASD. Results are displayed in Table 3. The philological reconstruction outperformed the PAML one in both cases.

In order to assess the quality of these results, we produced the majority archetype (MA), which at each position carried the majority variant. Additionally, we produced a random archetype (RA), where a randomizer as implemented in the Java programming language chose one variant at random for every place of variation. The RA was then evaluated. This procedure was repeated 1000 times and then we averaged over the RA evaluation results. As a point of orientation, we additionally provide the evaluation score of the maximally wrong archetype (MW), which is the one archetype that has a non-archetypical variant at each position where variation occurred. Note that none of these automatically produced archetypes requires a stemma beforehand. The results are displayed in Table 4. The RAs are considerably better than the MWs, but are clearly outperformed by both reconstructions and the MAs.

α(P/NB)	**WTE**	**PVE**
MW(P)	0.68	0
RA(P)	0.8	0.38
MA(P)	0.96	0.87
MW(NB)	0.52	0
RA(NB)	0.76	0.49
MA(NB)	0.98	0.95

Table 4: Evaluation of archetypes. α as used in philology denotes the archetype. We evaluated the majority archetypes (MA), the averaged random archetypes (RA) and the maximally wrong archetypes (MW) for the Parzival (P) and Notre Besoin (NB) traditions.

6 Discussion

The MA and MV archetypes are the most accurate ones. Hence, the true distribution of variants considering all manuscripts is such that the majority variant in the majority of cases is the archetypical one. Whether this conclusion holds for historical corpora remains to be shown. The PAML-generated archetypes performed well and were considerably better than the random condition. Note that the ML tree's ASD score was relatively low as compared to other algorithms evaluated in (Roos and Heikkilä, 2009). The limitation of the current reconstruction is that at each position the reconstructed text can only carry one of the variants of the extant manuscripts. This makes any reconstruction with many reconstructed internodes, such as the MVs or the PAML reconstructions by definition quite similar to the MAs.

6.1 Comparing Stemma and Archetype Production

From a bad stemma, disambiguation rules can nevertheless produce an accurate archetype and vice versa. One stemma can correspond to several possible archetypes and one archetype can be consistent with several different stemmas. The two tasks

[1] For details, data sets and evaluation scripts, browse `http://www.cs.helsinki.fi/u/ttonteri/casc`.

and evaluations should therefore be considered separately. This is of utter importance as it points to an imbalance of computational effort in the field.

6.2 Implications for Computer Philology

In biology the establishment of genealogical relations is in its own rights a primary goal, reconstruction of ancestral sequences being rather secondary. On the contrary in philology, compiling an edition is not only historically preceeding stemmatology, but can be considered the main target of dealing with manuscript corpora. Stemma construction is rather a secondary goal.[2] Despite this imbalance between the fields, the technological emphasis is on genealogical trees in both. This may be seen as a computer philological co-loan from bio-informatics. On the other hand it might correspond to a more Bédierian edition practise, where stemmatology is emphasized because it can point to the most important manuscript, which is however implicit and unlikely. Another reason for a benefit resulting from a shifting focus onto automatic archetype reconstruction is the problem of having two vorlages for one copy called *contamination*. Whilst especially excessive contamination is a bad problem for stemmatology, (Maas, 1960), considering automatic archetype reconstruction, on a word level it doesn't increase diversity and should therefore be a less severe problem.

For both traditions, the produced archetype was reasonably accurate. The automatisation of this process could thus accelerate the production of editions, making them most dependent on the indispensable digitization of corpora.

6.3 Implications for Bio-Informatics

In biology, reconstructive algorithms such as the one by Yang (2007) have been developed alongside biological benchmark data sets enumerated by Linder et al. (2010). However, in stemmatology, the probability to have a manuscript and its copy in the corpus at the same time is disproportionately higher than in the biological case making archetype production more transparent here. The resulting algorithmic conclusions from philology could therefore enrich the field of ancestral sequence reconstruction.

7 Conclusion and Outlook

We have presented an automated reconstruction of an archetypical text through philological rules and phylogenetic software. Additionally, we invented an evaluation for the produced archetypes, where we found the reconstruction to produce results considerably above chance level.

The inversion of the process is implicit. From a (manually or automatically) constructed archetype, all possible corresponding stemmas on the given set of manuscript digitizations can be computed and evaluated, which would be a new approach to stemmatology. Both tasks could thus profit from each other provided they are understood as separate and developed each in its own right.

Many issues remain unaddressed. The phenomena encountered in manuscripts are much more varied than word substitutions as modelled in this paper; an enumeration of some of the phenomena will corroborate this: word deletions, word separation errors, whole passages missing, text on the margins, unreadable or destructed text, crossing out of sections, oral variation, contamination. Trovato (2009) claims that manuscript loss of far more than 90% is realistic. In linguistics, historical-comparative studies have engaged in using bio-informatic software for instance in connection with the recontruction of language family trees. For automatic emendation these studies as well as the reconstruction of unattested word forms are a valuable source. Stemmatology itself offers ever new algorithms, artificial traditions and tools for electronic editing.

In the light of these manifold possibilities for elaboration and cooperation, the current study presents but one entry point into automatic archetype reconstruction.

Acknowledgments

We greatfully acknowledge the financial support of the Goethe University, the Federal State of Hesse and the Federal Ministry of Education and Research BMBF. We express our gratitude to our colleagues of the Texttechnology Lab for valuable comments, help and advice.

[2] For more details, consider for instance Pasquali (1988), Timpanaro (2005), and Reynolds & Wilson (2013).

References

T. L. Andrews and C. Macé. 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, December.

P. Baret, C. Macé, and P. Robinson (eds.). 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.

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

C. Howe, R. Connoly, and H. Windram. 2012. Responding to criticism of phylogenetic methods in stemmatology. *SEL studies in English Literature*, 52:51–67.

Karl Lachmann. 1853. *In T. Lucretii Cari De rerum natura libros commentarius: Index*. Georg Reimer.

C. Randal Linder, Rahul Suri, Kevin Liu, and Tandy Warnow. 2010. Benchmark datasets and software for developing and testing methods for large-scale multiple sequence alignment and phylogenetic inference. *PLoS Currents*, 2.

P. Maas. 1960. *Textkritik*. B. G. Teubner.

Webb Miller, Daniela I. Drautz, Aakrosh Ratan, Barbara Pusey, Ji Qi, Arthur M. Lesk, Lynn P. Tomsho, Michael D. Packard, Fangqing Zhao, Andrei Sher, Alexei Tikhonov, Brian Raney, Nick Patterson, Kerstin Lindblad-Toh, Eric S. Lander, James R. Knight, Gerard P. Irzyk, Karin M. Fredrikson, Timothy T. Harkins, Sharon Sheridan, Tom Pringle, and Stephan C. Schuster. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, 456(7220):387–390.

R. J. O'Hara. 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.

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

Tal Pupko, Itsik Pe'er, Ron Shamir, and Dan Graur. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, 17(6):890–896.

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

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

P. Robinson, A. Barbrook, N. Blake, and C. Howe. 1998. The Phylogeny of The Canterbury Tales. *Nature*, 394:839.

Philipp Roelli and Dieter Bachmann. 2010. Towards Generating a Stemma of Complicated Manuscript Traditions: Petrus Alfonsis Dialogus. *Revue dhistoire des textes*, 5(4):307–321.

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

M. Spencer, E. Davidson, A. Barbrook, and C. Howe. 2004. Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology*, 227:503–511.

Sebastiano Timpanaro. 2005. *The Genesis of Lachmann's Method*. University of Chicago, Chicago.

Paolo Trovato. 2009. *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.

P. T. van Reenen, M. van Mulken, and J. Dyk. 1996. *Studies in Stemmatology I*. Studies in Stemmatology. John Benjamins Publishing Company.

P. T. van Reenen, A. A. den Hollander, and M. van Mulken. 2004. *Studies in Stemmatology II*. Studies in Stemmatology. John Benjamins Publishing Company.

K. Voordeckers, C.A. Brown, K. Vanneste, E. van der Zande, A. Voet, S. Maere, and K. J. Verstrepen. 2012. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. *PLoS Biol*, 10(12).

Martin L. West. 1973. *Textual Criticism and Editorial Technique: Applicable to Greek and Latin texts*. Teubner, Stuttgart.

Ziheng Yang, Sudhir Kumar, and Masatoshi Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141:1641–1650.

Ziheng Yang. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24(8):1586–1591.