On the Usability of Available Digital Tools for Reconstructive Textual Editing

Philipp Roelli

On the Usability of Available Digital Tools for Reconstructive Textual Editing

Philipp Roelli

University of Zurich, Switzerland

roelli.sglp@yandex.com

Abstract

This article reviews some of the digital tools currently available for reconstructive textual editing. First the main idea of reconstructive textual editing is summarised, then its steps amenable to algorithmic description are compared to similar ones in evolutionary biology. The unequal ability of its variants to be relationship revealing is an important difference between the two fields. Two Latin texts with a complicated transmission are then introduced and used as data to illustrate some available tools *in praxi*. The main focus is on stemma reconstruction. Some steps of the process can already be largely automated, especially collating texts. On the whole it is found that tree-constructing software is of little help in the case of the medical text *Liber Aurelii,* whereas it is somewhat more helpful for Plato of Tivoli's translation of the *Centiloquium*. In a concluding part, the main problems for algorithmic approaches to the stemma are discussed: incomplete witnesses leading to only partly overlapping text samples, contamination in some witnesses, and rooting the automatically generated trees.

Keywords

Critical editing, Textual criticism, Stemmatology, Computer aids, Significant errors, *Liber Aurelii,* Plato of Tivoli.

I THE MAIN IDEA OF RECONSTRUCTIVE TEXTUAL EDITING

This essay shall focus exclusively on tools for reconstructive textual editing, which can be defined as an approach that aims to reconstruct a text known only in copies made from a lost original as closely as possible by using all available data (for a more detailed characterisation [Roelli, 2020: 3–4]).[1] The approach to be outlined was first developed by German and French scholars in the 19th century, Karl Lachmann (1793–1851) and Gaston Paris (1839–1903) being the most famous among them. It was refined in many ways mostly by Italian scholars of the second half of the 20th century (cf. the summary by Paolo Trovato [Roelli 2020: ch. 2.4]). The most recent common ancestor of all surviving copies is known as the 'archetype'. The first goal is to reconstruct this (often lost) archetype from the extant witnesses as far as possible, the second to examine it and to

1 I thank the excellent reviewers at Episciences for improving this essay.

try to correct its errors using the available external information (if any) about the original, the author, and his time. In reality, the situation may be more complicated: for instance, the author may have reworked the text and there may thus be more than one 'original' reading in some places. In antiquity and the early middle ages there are often many centuries between the oldest surviving manuscript and the original; and worse, even between the archetype and the original. Depending on the text in question, there may be from one to hundreds, occasionally even thousands of surviving copies of extant antique or medieval texts.[2] Clearly, digital aids are promising to handle the data of especially abundant traditions. Texts surviving in a handful or fewer witnesses can just as well be dealt with manually. The mentioned goal of approaching the lost text (that is as a first step its archetype), is reached by determining the relationship between the surviving witnesses by evaluating their readings. These relationships can be graphically depicted as a *stemma (codicum),* the genealogical tree which explains the observed variation in the witnesses in the most economical way. It helps us evaluate which readings in the various witnesses are archetypal (primary).

Fig. 1 depicts what is apparently the first printed *stemma codicum* in our field, dating from 1827 [Schlyter, 1827: appendix].[3] Here, the archetype or original is situated at its top end without a label; indeed the concepts 'archetype' and 'original' were not yet kept apart in the time of Schlyter. The lines represent the relation 'was copied to' (not excluding the possibility of lost intermediaries). The stemma provides information about the weight readings should be given in the process of editing the archetypal text depending on their witnesses' position. If the depicted stemma is correct, we will have to grant the readings of A more weight for the reconstructed text than the other witnesses. For argument's sake, let us consider a passage with four readings that occur in the witnesses as indicated in the stemma. Let the readings be: reading 1 (H, K, M, No166), reading 2 (A, L), reading 3 (C), reading 4 (B, G, N). Without the stemma we might be tempted to use a majority criterion and adopt reading 1 or 4. If we have the stemma, we will conclude that readings 1, 3, and 4 are innovations and we will choose reading 2 for the archetypal text as it is the reading of one of the two top branches and also occurs in the other. In cases similar to this one, it becomes thus possible to choose the archetypal reading quite mechanically, that is in a certain sense more objectively than if a philologist chose it intuitively based on its meaning (which is what was usually done prior to the 19th century).

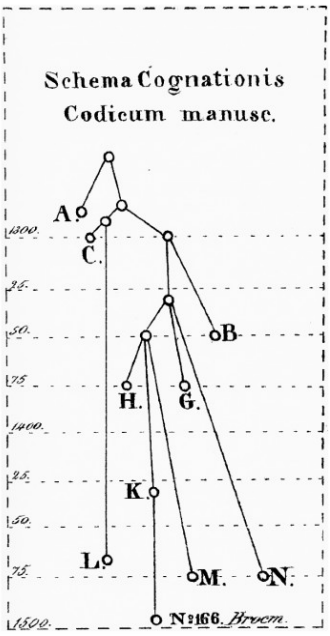


Figure 1: Carl Johann Schlyter's stemma of Västgötalagen (1827).

The crucial question is now: how does one find the correct stemma or genealogical tree for a given set of witnesses? This question led scholarship to the study of variant readings and to the concept of

2 Details on the question of how to edit such texts: [Bourgain, 1992].

3 Apart from a lack of considering contamination, this stemma seems to be largely correct: 'Regarded as a schema that draws up the principal lines and disregards the contamination of the tradition it would actually seem to be almost accurate' [Olrik Frederiksen, 2009: 129]. There are minor mistakes, so according to more recent research the group HKM is doubtful [ibid.: 135]. More on Schlyter's method in [Froger, 1978] and [Holm, 1972].

'significant errors' or *Leitfehler* (a term coined by Paul Maas, cf. Paolo Trovato [Roelli 2020: 117]). A significant error is an edit which can hardly or not at all be undone by a subsequent scribe without access to the original reading. Omissions of more than a word or two are good candidates; especially when they happen involuntarily in the form of eye-skips (also known as *saut du même au même*): a word or phrase occurs twice close-by in the original; the copyist copies up to the first occurrence but then jumps by mistake to the second one, thereby omitting what stood in-between. Although this may happen to more than one scribe at the same place, it is usually out of the question that the missing text can be reconstituted without access to a witness that still has it. A significant error should be directed: it must be clear which variant is the original one and which others are not. Only the secondary readings can serve to establish families, the original does not. In case of eye-skips the direction is usually clear,[4] although, unfortunately, the secondary reading (the omission) may happen more than once. The reconstruction of the stemma becomes much harder if some scribes used more than one source to compile a new manuscript, thus working philologically and trying to improve their new copies. This phenomenon is known as contamination (cf. Tuomas Heikkilä [Roelli, 2020: ch. 4.4]); it will be discussed further below. Matters are especially difficult to deal with if this happens from a lost witness further up the tree than the archetype.[5] As a general rule, the more witnesses, the more contamination is to be expected.

There are several other approaches to editing texts that do not strive to reconstruct an original text. It is important to stress that the best way to edit a text depends both on the textual transmission and on the editor's scholarly interests. In some cases the transmission does not allow the use of the mentioned methodology, for instance because the witnesses differ too greatly from one another or are too fragmentary. There may also be non-original text-forms (recensions) that are of scholarly interest in themselves. These fields have their own computational tools that are not treated here (an example is presented by Dirk van Hulle [Roelli, 2020: ch. 7.9]).

II PARALLELS WITH MOLECULAR BIOLOGY

The procedure of reconstructive editing just outlined can *in praxi* be summarised in these five steps:

- Transcription of available witnesses,
- Evaluating significant errors,
- Reconstructing the stemma,
- Editing the archetypal text according to it,
- Emending the archetypal text to approach the original as far as possible.

As many of these steps are nearly of an algorithmic nature, the idea to use computers to perform them arose early-on and quite naturally when first computers became available. The first such approaches were already done in the late fifties (cf. Armin Hoenen in [Roelli, 2020: ch. 5.1]). For

4 Unless an addition (that may have stood *supra lineam*) moved into the copyist's text and the anchor word was repeated; the context usually shows whether this is the case or not.

5 A process known as 'extra-stemmatic contamination', cf. Paolo Trovato [Roelli, 2020: 123] and Marina Buzzoni [Roelli, 2020: 386].

most of these steps there are today computer aids, but none of them can be fully automated as yet. The computerised approaches to these five steps differ widely in their advancement. Recent years have seen dramatic advances in the automatic reading of handwritten documents (i), for instance the open-source Kraken[6] project in Paris has already reached levels of reading old handwritten material that would have seemed quite unthinkable a few years ago. The second step (ii) has been largely neglected; many digital scholars seem to hope that a large amount of non-significant errors will work just as well as a few rare significant ones. Below, we will see that this is in general not the case. We[7] made a first attempt to automatically identify candidates of significant errors. The third step (iii) is very similar to what evolutionary biologists do when they construct genealogical trees from DNA sequences. Thus, it was possible to borrow much know-how from them. However, we will see that locating the archetype is a specific problem in our field for which the biologists' tools are not helpful. The fourth step (iv), sometimes referred to as *Urtext* reconstruction, is still quite experimental; although given a stemma and a set of texts the task to find the most likely readings for the (lost) intermediary nodes would seem to be a relatively straight-forward task to program (Armin Hoenen: [Roelli, 2020: ch. 5.4.6]) – at least if the stemma is a tree, i.e. does not contain contamination. If the position of the archetype (in biological terminology: the 'root') is also known, the *Urtext's* readings could then be largely reconstructed probabilistically by the computer. A problem is that (as shall be seen below) the archetypal text is located between two nodes if there are only two hyparchetypes, so the last step from these to the archetype remains unclear. Traditional editors face the same problem: it is largely up to their philological judgement which of the two hyparchetypal readings they prefer where they disagree (a point reproved by [Bédier, 1928]). The final step of emending the archetype's text has not been tackled by digital aids.[8] In the remainder of this essay we shall mostly focus on the third step, the automatic finding of the stemma from transcribed data.

There is considerable similarity of our problem with that of molecular biologists. Where we deal with text strings, they deal with DNA or amino acid sequences, thus basically 'texts' consisting of four or twenty letters, respectively. Both try to find the tree that explains the variation in the most economic way. The standard approach in biology is to measure the distances between each pair of taxa (i.e. groups of related organisms) and to use the resulting distance matrix to find the best tree (see Jean-Baptiste Guillaumin [Roelli 2020: ch. 5.5] for a practical example). Alternatively, probabilistic Bayesian methods may omit the step of the distance matrix and optimise an initial tree directly in tree-space (cf. Sara Manafzadeh and Yannick M. Staedler [Roelli, 2020: ch. 5.2] as well as Teemu Roos [Roelli, 2020: ch. 5.3]). Either way, a point that has often not been paid its due attention is the distance function (or metric[9]) one uses to measure the distance between items. In biology this

6 https://escripta.hypotheses.org/tag/kraken. For details see Kiessling et al. 2019.

7 [Roelli & Bachmann, 2010], discussed further below. The software (a perl script) is available (without much documentation) at osf.io/3gr8x.

8 There is an approach for epigraphic texts currently being developed by Yannis Assael et al. See https://ithaca.deepmind.com.

9 In mathematics a 'metric' is a function that measures the distance between two points in a (metric) space. It has to fulfil four axioms: (i) the distance from a point to itself is 0, (ii) if two points are not equal their distance is a positive number, (iii) the distance from *a* to *b* is the same as that from *b* to *a*, (iv) the triangle

point seems to be less crucial than in stemmatology as the rate of mutations happening is more constant and no-one interferes by making deliberate changes. In textual criticism a manual assignment of weights to variants has been attempted [Macé, De Vos, Geuten, 2012]. The mapping from text-space to matrix-space by means of a notion of distance can be formalised in this way:

$$L^{w\times n} \xrightarrow{\text{induced by } d} \mathbb{R}^{n\times n} \xrightarrow{\text{heuristic algorithm}} T_n$$

witnesses' texts ↦ distance matrix ↦ tree graph

L is the set of words, w are the number of words in each text string[10] and n the number of witnesses. T_n is the set of unrooted binary trees with n leaves representing the n witnesses. In this formula, d is a metric that calculates a 'distance' between any two texts $t_1, t_2 \in L^r$, $(t_1, t_2) \mapsto d(t_1, t_2) \in \mathbb{R}$.[11] The idea is to make d correspond as closely as possible to our real-world understanding of scribal accuracy or likelihood of copyists' mistakes happening. The second step, from matrix to tree is done by means of a heuristic algorithm such as those found in the free Phylip software package used in molecular biology, for instance neighbour-joining[12] or the Fitch and Margoliash algorithm.[13] The most natural (naive) choice of the metric d is 'counting variants', i.e. $d(t_1,t_2)$ is the number of operations needed to transform t_1 into t_2 (known as 'edit distance'), where an operation is the deletion or insertion[14] of a word or, better, a reading (that may consist of several words). The Unix `diff` command uses the longest common subsequence algorithm (LCS) repeatedly to detect differences in text-strings and can be used for such a task. Now the insertion or deletion of n consecutive words can be counted as a single operation, n distinct operations, or anything in-between. Something in-between is likely the most fitting approach. But we still have not made a difference between a trivial edit, say of near synonyms *dominus* to *deus* or simple word transpositions (which often do not change the meaning in Latin), and a more significant error. There are up to now few studies on how strongly changing the weighting changes the resulting tree.[15] In my own experience the trees tend to be quite robust when changing weights, especially their outermost groups of witnesses. What does change readily with

inequality holds, asserting that the distance between two points is smaller or equal to the their distance going through any 'detour' of a third point. These criteria leave many possibilities to define 'distance' in a given space.

10 Including the 'zero' word. Instead of words readings could also be used as the basic elements.

11 For all n witnesses this produces a matrix $\mathbb{R}^{n\times n}$. As d is symmetric, this matrix is also symmetric, besides it has only zeros in its diagonal (the distance between any text and itself being zero), thus only the subspace $\mathbb{R}^{n(n-1)/2}$ is relevant.

12 This approach sequentially groups taxa by joining the two most similar ones. Details with an example in the Parvum Lexicon Stemmatologicum: https://www.sglp.uzh.ch/static/MLS/stemmatology/Neighbour-joining_229149983.html.

13 https://evolution.gs.washington.edu/phylip/doc/main.html.

14 Possibly also substitution and transposition: this is known as the Damerau–Levenshtein distance.

15 Spencer et al. 2004 studied the influence of manual weighting in a Middle English text (Lydgate's *Kings of England*). They weighted omissions or changes in meaning strongly but did not use traditional significant errors. This did not change the final tree greatly.

weighting is usually the upper branches in the stemma representing the main families. It is precisely to determine these that significant errors are crucial (also in traditional editing).

To my knowledge there is as yet only one attempt to automatically assign a value on a continuous scale (between 0 and 1) to weight the significance of variants. Bachmann and I devised it in 2010 by assuming that if there is no contamination in a set of texts, then two significant errors' absence and presence will not be found in all four possible combinations in the witnesses. This approach uses the principle that significant errors should happen only once in a specific textual tradition as their very idea is that they can hardly be undone [Roelli & Bachmann, 2010, 317–318]. Variants that fulfil this criterion in combination with many other candidate variants can be weighted more strongly. This approach, thus, tries to assign a numerical value for the fitness of a variant as significant. Unfortunately, it cannot determine which of the variants is the original one as the presence and absence of words or readings is symmetrical. Therefore we use quotes and speak of the '*Leitfehler*'-method. Clearly more work on this approach would be required, but first results were promising.

III SOME EXISTING TOOLS WITH TWO COMPLICATED TEXTUAL TRADITIONS AS EXAMPLES

In order to work with concrete, real-life data I will now use data from two Latin texts that are currently under study to find out how well some of the available software works.[16] All software I used for what follows is freely available. I used it either online or on my rather low-power Linux laptop.

3.1 *Liber Aurelii*

The first example is a text I have recently edited critically for the first time using the traditional 'manual' approach [Roelli, 2021]. This anonymous text, erroneously known as *Liber Aurelii,* is a late antique Latin medical work on acute illnesses that goes back to lost Greek sources, mostly from the Methodic school and Soranus of Ephesus (fl. 2nd century AD). Its original title is unknown. It is extant in three recensions: the main text that had come down to the High Middle Ages in a seriously garbled text form and two reworkings by physicians of probably the 10th and 11th centuries. Especially the later one by the Salernitan author Gariopontus was very successful and has survived in more than 65 witnesses [Glaze, 2005]. Gariopontus still had a more comprehensible copy available, besides he also used the other recension; thus his text is contaminated. The author of this earlier recension had shortened the text drastically, often simply omitting hard to understand content. His text is less than half as long as the other two. As usually in medieval Latin, there is quite a great deal of variability in spelling. The text is about the size of a typical antique book (27 chapters of a total of some 11'000 words). I will use this text's chapter 12, which consists of some 750 words in the main version, some 500 in the shortened one (here Gariopontus took his text mostly from the shortened recension) to provide data for what follows. For the edition I determined the

16 The transcription data of Aurelius is published at: https://osf.io/2pv8r. For that of the *Centiloquium* its editor (Emanuele Rovati) has to be contacted.

stemma of the *Liber Aurelii* manually using the traditional methods of textual criticism (fig. 2; for a key to the sigla, see the edition: [Roelli 2021: lxiv–lxv]).

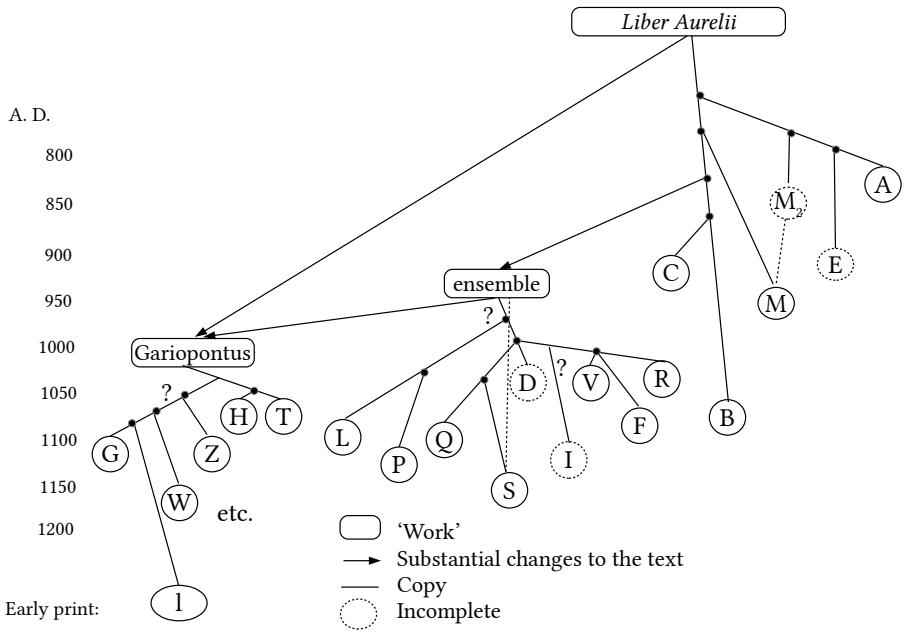


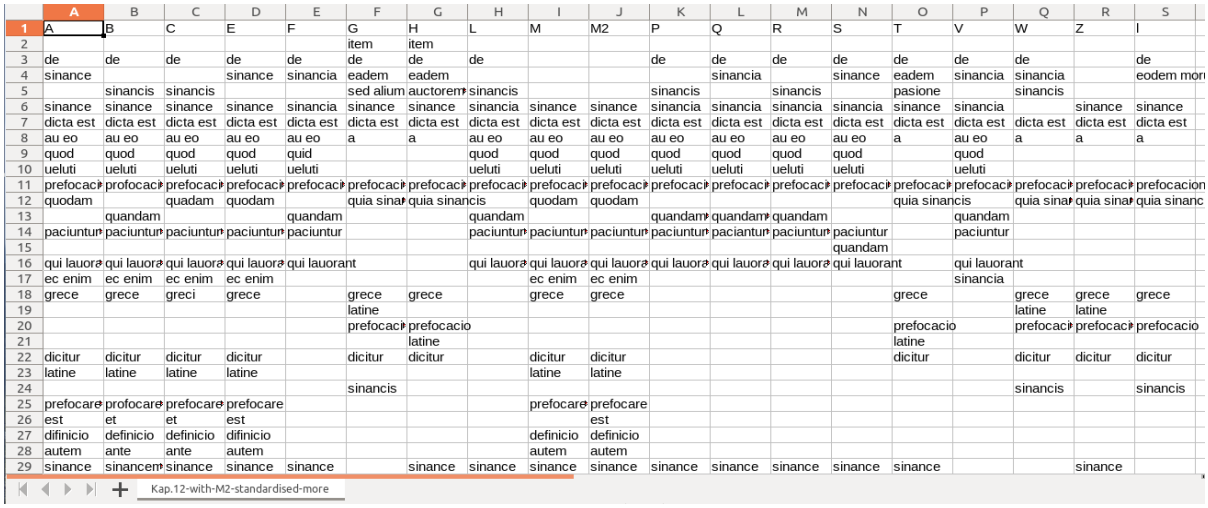
Figure 2: Manually determined stemma of the Liber Aurelii.

I had already transcribed all witnesses manually into a txt file in which each line contains the text of one witness; therefore I will skip the automated reading of the manuscript texts. While transcribing, one should note all orthographic and other peculiarities of the witnesses as this information can come in handy in later steps of editing. But for the input of tree-finding software, it is preferable to use standardised spelling as purely orthographic differences such as *hec* vs. *haec* are very unlikely to be relationship revealing, rather they are what could be termed 'orthographic noise'.[17] Regular expression syntax in a text editor (e.g. the free Edit pad lite on Windows) can be used to quickly standardise the spelling. In our example, among other things, I removed all punctuation and all spelling that tends to be variable in medieval Latin. This can be done drastically and quickly by just removing for instance all *h*'s altogether, or even changing all *b*'s into *u*'s.[18] The resulting txt-file may no longer look like Latin but the important information to determine the relationships between witnesses is preserved and much of the 'noise' is now removed. This step means that such differences are defined as zero distances: e.g. *humiliatio* = *umiliacio*. This step is fully automatic and

17 This is true for Latin and Greek texts. The situation for vernacular medieval texts may be different, according e.g. to Blake and Thaisen's (2004) findings.

18 The letter *h* was silent in medieval Latin and is often optional for scribes. The mixing up of *v* and *b* is a more regional phenomenon known as betacism. Spanish today still pronounces the two letters identically: there is an anonymous saying *beati Hispani quibus vivere est bibere*. It happens to be a common feature of some of this text's witnesses. The letter *v* is a modern invention, Antiquity and the Middle Ages used *u* for both the vowel ('*u*') and the semi-vowel ('*w*') / consonant ('*v*'). Some Latinists print their editions with the letter *v* for the sake of modern readers, some just use *u* either way.

takes no more than a few minutes to accomplish. It is important to note that the exact list of what should be standardised and what not depends on time, language, and textual tradition.



	A	B	C	E	F	G	H	L	M	M2	P	Q	R	S	T	V	W	Z	I
1	A	B	C	E	F	G	H	L	M	M2	P	Q	R	S	T	V	W	Z	I
2						item	item												
3	de	de	de	de	de	de	de	de			de	de	de	de	de	de	de		de
4	sinance			sinance	sinancia	eadem	eadem					sinancia		sinance	eadem	sinancia	sinancia		eodem mor
5		sinancis	sinancis			sed alium	auctorem	sinancis			sinancis		sinancis		pasione		sinancis		
6	sinance	sinance	sinance	sinance	sinancia	sinance	sinance	sinancia	sinance	sinance	sinancia	sinancia	sinancia	sinancia	sinance	sinancia		sinance	sinance
7	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est	dicta est
8	au eo	au eo	au eo	au eo	au eo	a	a	au eo	au eo	au eo	au eo	au eo	au eo	au eo	a	au eo	a	a	a
9	quod	quod	quod	quod	quid			quod	quod	quod	quod	quod	quod	quod		quod			
10	ueluti	ueluti	ueluti	ueluti	ueluti			ueluti	ueluti	ueluti	ueluti	ueluti	ueluti	ueluti		ueluti			
11	prefocaci	profocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocaci	prefocacion
12	quodam		quadam	quodam		quia sina	quia sinancis		quodam	quodam					quia sinancis		quia sina	quia sina	quia sinanc
13		quandam			quandam			quandam			quandam	quandam	quandam			quandam			
14	paciuntur	paciuntur	paciuntur	paciuntur	paciuntur			paciuntur	paciuntur	paciuntur	paciuntur	paciantur	paciuntur	paciuntur		paciuntur			
15														quandam					
16	qui lauora	qui lauora	qui lauora	qui lauora	qui lauorant			qui lauora	qui lauora	qui lauora	qui lauora	qui lauora	qui lauora	qui lauorant		qui lauorant			
17	ec enim	ec enim	ec enim	ec enim					ec enim	ec enim						sinancia			
18	grece	grece	greci	grece		grece	grece		grece	grece					grece		grece	grece	grece
19						latine											latine	latine	
20						prefocaci	prefocacio								prefocacio		prefocaci	prefocaci	prefocacio
21							latine								latine				
22	dicitur	dicitur	dicitur	dicitur		dicitur	dicitur		dicitur	dicitur					dicitur		dicitur	dicitur	dicitur
23	latine	latine	latine	latine					latine	latine									
24						sinancis											sinancis		sinancis
25	prefocare	profocare	prefocare	prefocare					prefocare	prefocare									
26	est	et	et	est						est									
27	difinicio	definicio	definicio	difinicio					definicio	definicio									
28	autem	ante	ante	autem					autem	autem									
29	sinance	sinancem	sinance	sinance	sinance		sinance	sinance	sinance	sinance	sinance	sinance	sinance	sinance	sinance			sinance	

Kap.12-with-M2-standardised-more

Figure 3: The beginning of the automatically generated alignment table of Liber Aurelii*, chapter 12, obtained by using CollateX.*

Critical Apparatus

1 ^item G, H

1 de sinance~ M, M2, Z

1 sinance] sinancis B, C

1 sinance sinance] sinancia sinancia F, L, P, Q, R, V

1 sinance] eadem sed alium auctorem G

1 sinance] eadem H

1 sinance] eodem moruo I

1 sinance] eadem pasione T

1 sinance sinance] sinancia sinancis W

1 sinance] sinancia S

1 au eo quod ueluti prefocacionem quandam paciuntur qui lauorant ec enim] a prefocacione quia sinancis G, I, T, W, Z

1 au eo quod ueluti prefocacionem quandam paciuntur qui lauorant ec enim grece dicitur latine prefocare et definicio autem s impetu circa inguine nam oc tonsillarum impetu differ t quod ea acuta passio est itemque et illo quod non profocant tonsille eis inspiciamus ostendendum est] a prefocacione quia sinancis grece prefocacio latine dicitur sinance est dificilis translacio spirit H

1 quod] quid F

1 prefocacionem] profocacionem B

1 prefocacionem] prefocacione V

1 quandam] quodam A, E, M, M2

1 quandam] quadam C

1 ^paciuntur S

1 paciuntur] paciantur Q

Figure 4: Automatically generated apparatus by Juxta.

The point of aligning the witnesses' texts is to be able to compare readings (corresponding passages) in the various witnesses. The typical collation table that is used in manual editing can now be generated automatically from a txt-file within seconds using the software CollateX (https://collatex.net/doc/#text-input). As a command line tool, CollateX[19] produces from a json-file

19 A simple java -jar collatex-tools-1.7.1.jar -f csv DATA.json > OUTPUT.csv. There is now also a newer python version: https://pypi.org/project/collatex.

(which is easily obtained from our txt-file) a collation table in the form of a csv-spreadsheet (fig. 3). In our case the standardisation of the spelling reduced the number of readings only slightly (from 883 to 860). This software functions well (although not perfectly in all instances, see *sinancia* vs. *sinancis* in row 4 and 5 that should go together) and can save a significant amount of time.

Another useful piece of software is Juxta:[20] it allows to generate 'critical' apparatuses based on a set of witnesses. It compares the witnesses but does not evaluate them, so one has to provide a manual critical text in order to be able to generate a truly critical apparatus. Such lists can be useful to find readings that are shared by a given group of witnesses. The output is an html-file that can be easily searched in any web-browser. The latter's beginning is shown for a part of our text in fig. 4.

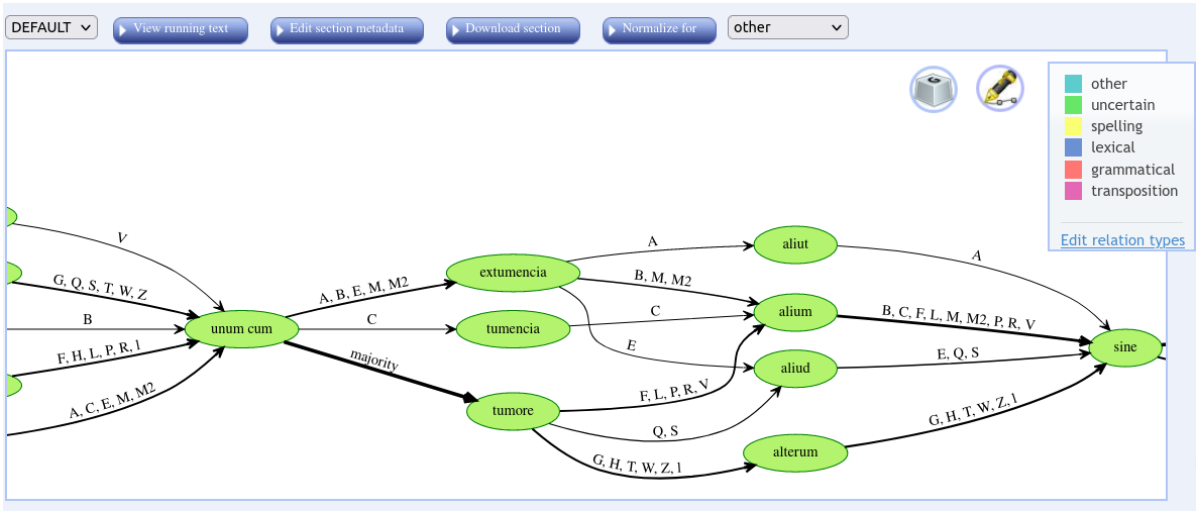


Figure 5: StemmaWeb showing an excerpt from a variant graph.

An entire set of tools is offered by Tara Andrews' Stemmaweb (https://stemmaweb.net). One can visualise a flow diagram with readings as nodes and witnesses as edges (fig. 5), besides genealogical trees can be created using several approaches. Fig. 6 shows such a tree for our example text using the intuitive but simplistic neighbour-joining approach. It does not weight readings philologically, but the result is nonetheless quite realistic: The three recensions are kept apart, but B and C should form a group and stand below M; V should group with F and R. A further negative point of this approach is that all edges are assigned the same length.[21] More sophisticated approaches provide branch length (the longer the branch, the more differences) which enables the user to gauge to what extent texts differ and to decide which bifurcations should actually be understood as multifurcations.[22]

20 Cf. https://wiki.digitalclassicist.org/Juxta. Unfortunately, Juxta is no longer maintained. A 2014 version can be found at Github (https://github.com/performant-software/juxta-service). Sébastien Moureau's ChrysoCollate (https://cental.uclouvain.be/chrysocollate) performs similar tasks.

21 This is due to the implementation, neighbour-joining can produce branch lengths. This may be improved in the future (personal communication by Tara Andrews).

22 The RHM algorithm (by Teemu Roos, Tuomas Heikkilä and Petri Myllymäki) on StemmaWeb (with 100,000 iterations) performed significantly worse than neighbour-joining. For instance SQ and Gl do not

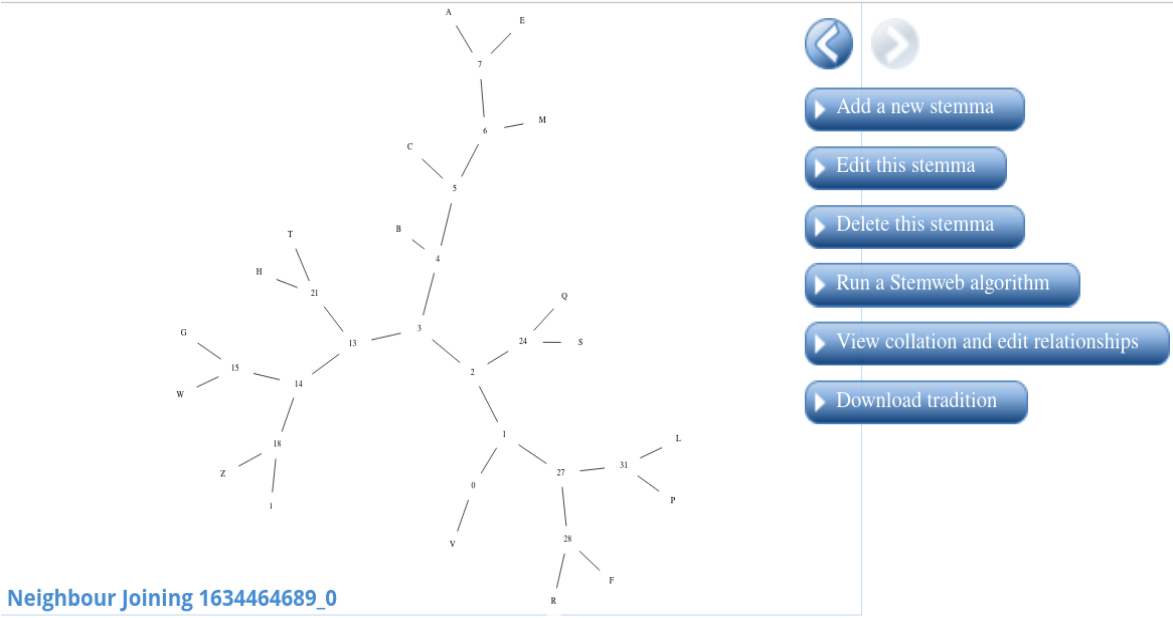


Figure 6: Tree from StemmaWeb based on neighbour-joining.

The Roelli–Bachmann script produced the tree shown as fig. 7 with automatically determined '*Leitfehler*'; it does include variable branch-length. The clear pairs of very similar manuscripts are correctly found: AE, LP, FR, Gl, HT, BC. Three lines are not correct (compared to the stemma in the edition); they are depicted as manually added red arrows in the figure. The most serious mistake (in both plots) is that the abbreviated recension (on top in fig. 7) should belong close to B and C. This artefact is a common problem of most similar algorithms: it moves shorter texts towards the centre of the plot. This happens because texts have – loosely speaking – fewer possibilities to differ when they are compared to a shorter text than when to a longer one. I am not aware of approaches that try to solve this serious issue. In practical life there is the work-about to plot trees of parts of the tradition with more or less the same text amounts, then try to stitch them together. The weighting of variants changes the result significantly in this case. This is less the case for the kinship of leaves in the tree, especially when they are rather distinctive, but the upper branches in the tree are prone to shift easily when the weights are changed. Unfortunately, these are the parts of the evaluation that are also the most difficult for a traditional human textual critic. The result of fig. 7 could be improved by using hand-picked significant errors, but this, of course, will easily become a case of rejoicing after finding again what one had previously hidden behind the bush (as Nietzsche used to say).[23] The '*Leitfehler*' script functions better than simple neighbour-joining with edit distance but would still need significant improvement. Some improvement would be achieved by using entire readings or possibly *n*-grams in order to be able to include series of common words that differ in their combination as potentially significant. Even then, in this case, the improved result would not

form groups (plot not shown).

23 The following list of hand-picked errors: *definitio, tumore, cognoscere, proiciunt, afixia, introrsum, irruit, uisceribus, retentione, aperiant, molitam, catarticum, efficitur, sursum* improves the result somewhat (plot not shown).

be very useful in practical life as an editor because it is of no help for the two most serious problems, rooting and contamination, as will be discussed below.

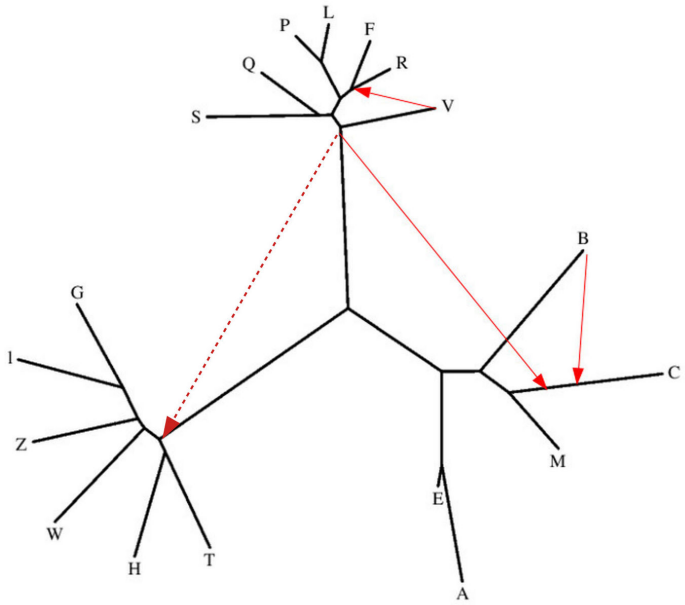


Figure 7: Tree generated by using the Roelli–Bachmann script to calculate a distance matrix and the Fitch–Margoliash algorithm to plot the tree. Red arrows show where the plot differs from the editor's expertise. The dotted line represents contamination (the top group having used witnesses of both others). Minor cases of contamination (especially in M and S) are not shown. Contamination can not be detected by any of the approaches discussed.

3.2 *Centiloquium*

The second text to be used is the Latin translation of Pseudo-Ptolemy's *Centiloquium* by Plato of Tivoli, currently being edited by my colleague Emanuele Rovati, who kindly allowed me to use his transcription data and some of his new insights into the text's transmission. The *Centiloquium*[24] is a collection of one hundred aphorisms on astrology that was considered to be written by the famous antique astronomer Claudius Ptolemaeus. In Arabic it is often transmitted together with the commentary of the tenth-century astrologer Abū Jaʿfar Aḥmad ibn Yūsuf ibn Ibrāhīm ibn al-Dāya (on whom: [Lemay, 1978]). The extant Greek version of the text does not seem be the original version, but rather a Byzantine translation from the Arabic. This short text (of some 16'000 words) was translated several times from the Arabic into Latin, sometimes with, sometimes without Abū Jaʿfar's commentary, and remained influential into early modern times. The text we use here was translated by Plato of Tivoli in 1136 and includes the commentary. There are 101 known manuscripts and three early prints.[25] In total, some 80 of the manuscripts contain a substantial part or the entire

24 Further data about the text [Juste, 2022]. Edition of the Arabic text [Martorello & Beza, 2013], and the *Ptolemaeus Arabus et Latinus* project: https://ptolemaeus.badw.de.

25 The first of these (the editio princeps): Venice, Erhardus Ratdolt, 1484. Ratdolt was a competent editor. The three prints are related to one another and to manuscript Va19.

text. The sample for our plots uses twelve of the one hundred aphorisms and their commentaries (the numbers 9–10, 30, 51–53, 60–62 and 98–100), it spans some 3'000 words per witness.[26] The witnesses' text is much more homogeneous than the one of the *Liber Aurelii.* The most common edits are in orthography, word transpositions, changes of mood, tense, number, or person in verbs, and the substitution of pronouns (e.g. *eius* vs. *illius* vs. *istius*) without changing the meaning significantly.

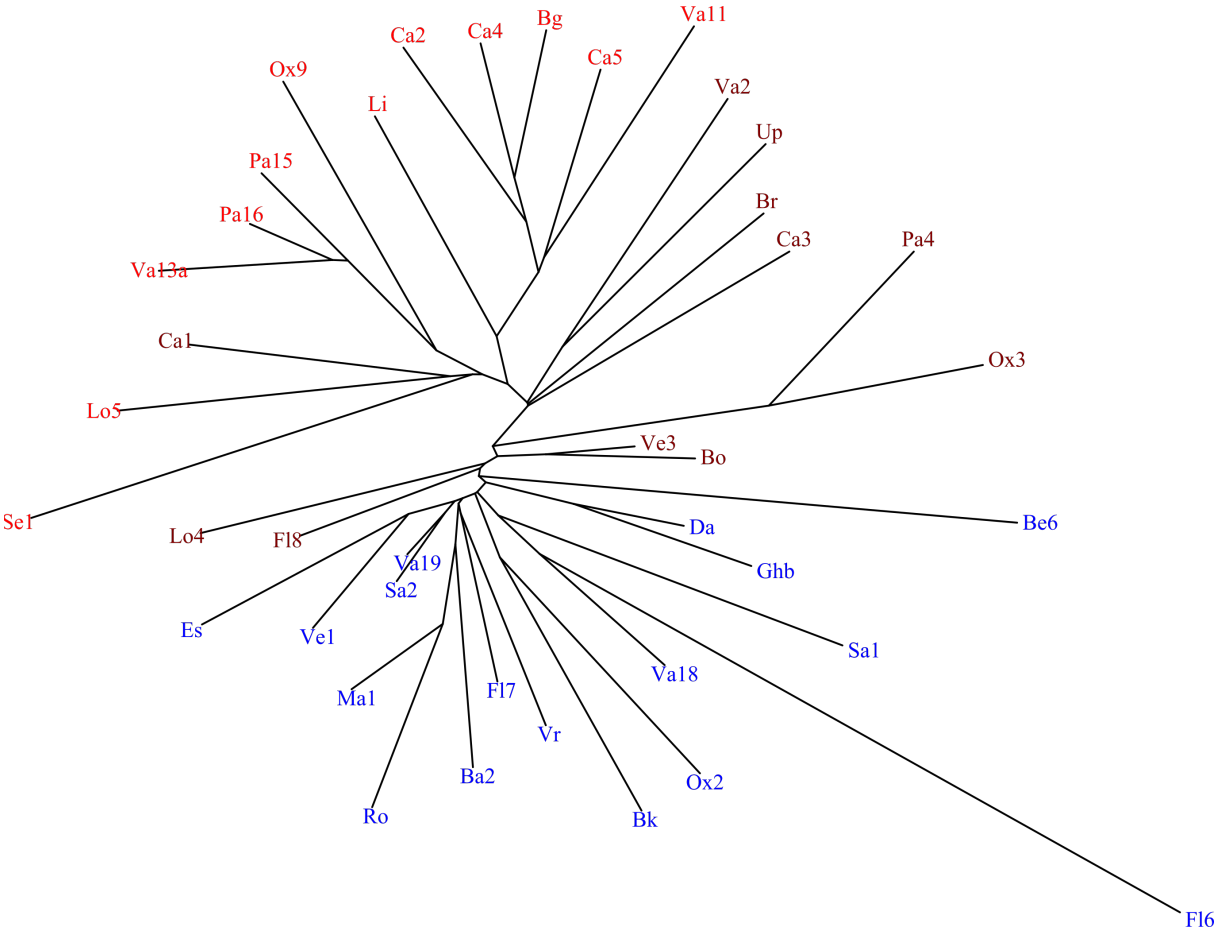


Figure 8: Centiloquium *tree generated by the same method as fig. 6. The α group is depicted in red (witnesses from the threefold version in lighter red), β in blue.*

There are two major groups: α is the text translated by Plato of Tivoli, β a reworked version by Gerard of Cremona, as Rovati has been able to show.[27] The latter is more stable than the former. Gerard checked the Arabic original and changed some 400 passages in Plato's translation. Later copyists often conflated the two versions. The majority of witnesses of Plato's unaltered translation belongs to what is known as the threefold version which quotes each aphorism in three translations

26 The choice of transcribed samples will depend on pragmatic grounds (the ones used here are extant in most witnesses), as well as more theoretical ones (they contain less technical and more narrative passages that exhibit more variance in this case).

27 The data in this paragraph comes from Emanuele Rovati and will be published in detail in his critical edition (around 2025).

(Plato, Adelard of Bath, and the anonymous '*Mundanorum* 1' translation). Obviously only the text of the first of these translations is used here. An example can illustrate the typical differences between α and β: Plato wrote *Obtinebit, inquit, locum patris et erit 10 annis fere in regno, sed erit sicut ille cui iubetur*. Gerard compared the Arabic original and added, apparently in the margin: *in alio* (i.e. in another Arabic manuscript)*: sub potestate vel regimine alterius.* Most β manuscripts now read exclusively: *Obtinebit, inquit, regnum patris et erit 10 annis fere in regno, sed erit sub potestate alterius.*

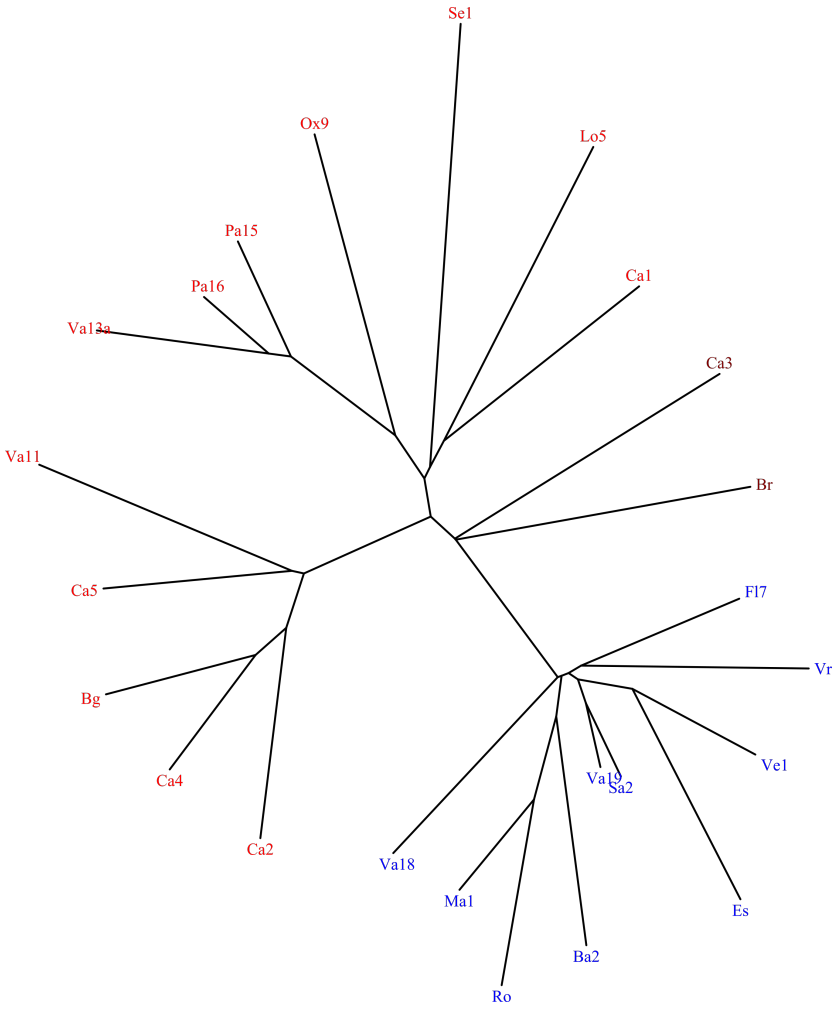


Figure 9: A second plot without the most obvious cases of contamination. Colours as in fig. 8.

I will focus on the automated trees I plotted based on Rovati's transcriptions; thus I skip the process leading to them as it is identical to what was described above for the *Liber Aurelii.* Suffice it to say that CollateX finds 3'190 readings for this sample. A first plot of forty manuscripts (the most complete and mostly the oldest ones) produced the plot in fig. 8. Many of the witnesses can be identified as contaminated at first sight as they contain glosses with readings from other branches or pairs of both the α and the β reading within the main text joined with *uel* or similar. Removing the most obviously contaminated manuscripts from the sample yielded clearer groups (fig. 9). Among β Va18 indeed seems to be the closest to the hyparchetype, although it is not the best witness as it introduced quite a lot of mistakes of its own (*Eigenfehler*). This accounts for its long branch. In

contrast, Va19's text has few *Eigenfehler,* also as depicted. The other groups in the plot are clustered correctly: Ma1–Ro–Ba2, Va19–Es–Ve1. Fl7 is a reworking, hence naturally removed from the rest. It seems that Vr should be closer to Ma1–Ro–Ba2. Only Ma1–Ro–Vr share one of the few eye-skips in the sample. The position of Sa2 is as yet unclear; it reads mostly with Va19–Ve1 and sometimes with Va18–Ox4 or even Vr–Fl7. However, the relationship between these groups does not seem to be as depicted: they do not all stem from one ancestor as the plot might insinuate. The most original manuscripts of α are indeed Br and Ca3 as the plot implies. Contamination is common in α, especially in the non-threefold witnesses, but many details are not yet clear to the editor.

This sample produced the following twenty best scoring '*Leitfehler*' (including their relative score):[28]

removebis -- 100%
auctor -- 85%
accepta -- 85%
fecit -- 82%
divise -- 82%
diutius -- 82%
libros -- 82%
perficitur -- 82%
sumitas -- 82%
sibique -- 81%
proximos -- 81%
contracta -- 81%
fortassis -- 81%
penitus -- 81%
pervenerunt -- 79%
inquisivi -- 79%
nequit -- 76%
relationis -- 76%
potavit -- 75%
perpendi -- 74%

The underlined ones are among the best significant errors Rovati has detected manually. But the two highest scoring ones in the list are not good significant errors, *removebis* stands against *removebit* and *remove, auctor* against *autor,* forms that can easily be changed. Some other promising significant errors, such as *rei pro qua fit electio*, could not be found by the algorithm as each word occurs also elsewhere. This would be different if readings were used instead of words.

Despite the greater number of witnesses, the situation of the *Centiloquium* transmission is less complex and more amenable to computer-aided study than for the *Liber Aurelii*. Despite the widespread contamination, the two main groups were detected correctly (especially in the second plot), and the known sub-groups in the β version were mostly correct too. However, the most contaminated witnesses had to be removed manually and the root cannot be assigned in the plots. Rovati intends to print one of the two versions with an extra apparatus for the other's changes in his forthcoming edition.

IV LIMITATIONS OF DIGITAL TOOLS TODAY

Often, basically every witness has its own individual problems. They may range from physically missing parts to cancelled or otherwise lost text, missing chapters, unreadable characters or abbreviations. One is tempted to say '*chaque manuscrit a son histoire*'.[29] Fig. 10 shows a page of manuscript E from the Aurelius tradition. It had some pages cut off and lacks about a third of each line of text. Another manuscript (M, see ill. 1 in [Roelli, 2021: 173]) was vandalised by a medieval physician who wrote the text from another branch over the original text where he thought it was

28 In this sample *b* was retained (it is used correctly in the witnesses) and the letter *v* was used by Rovati. For details of calculating the values, see Roelli 2014: 46–47.

29 Adapting from Antoine Meillet's slogan '*chaque mot a son histoire*' against the Neogrammarians.

better or more complete. In both cases there are many passages that are irretrievably lost. For the trees I plotted, I pragmatically inserted the readings of the closest relative where there were lacunae – for which, of course, one has to have preconceptions which is the closest other witness. For E this was A, for M it was C. If I had just left the missing words out, the resulting tree would have become much worse, as shorter witnesses tend to move to the centre of the tree. Such problems make a fully automated computerised treatment of the data at present illusory. Apart from such material problems, currently available stemma generating software struggles most with two problems: rooting and contamination (especially between various recensions).

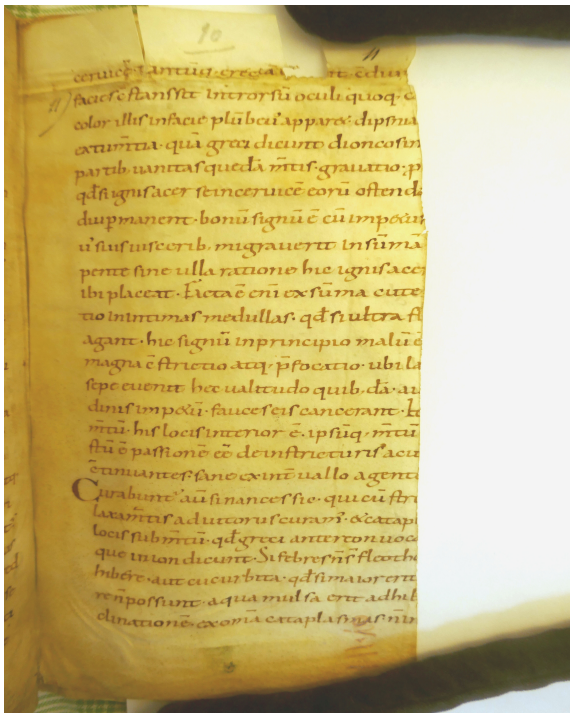


Figure 10: Cut manuscript page, Einsiedeln, Stiftsbibliothek 363, fol. 11r, of the Liber Aurelii *text. Photograph by Philipp Roelli, published by permission.*

4.1 Contamination

Contamination distorts the plots. If we add manuscript M's (*Liber Aurelii*) second contaminating hand to our plot,[30] this 'M²' correctly groups with A and E, but the group MBC loses cohesion and the entire plot becomes distorted (fig. 11). Even at the very other end of the plot, G and l no longer form a group as they should. The basic idea of the Roelli–Bachmann algorithm for weighting variants ceases to function with strongly contaminated witnesses in the sample. But other software also suffers from this problem as it is constrained to find a tree. In a tree the contaminated witness is thus fit somewhere between its two (or more) ancestral groups. Mathematically, a more adequate approach would be to plot not trees but networks that allow several incoming lines to nodes, which would link the contaminated witness to its two (or more) sources. There is software (such as

30 For the sample text M² I used M's original text where it was not changed by the contaminating scribe.

SplitsTree[31]) that can produce such networks (an example by Jean-Baptiste Guillaumin [Roelli, 2020: 351]). But first experiments along this line are not yet very promising for modelling the stemma. The degrees of freedom for the algorithm are too great, one would have to penalise extra edges strongly, otherwise the algorithm tends to plot them for any random identical spelling between unrelated witnesses. A pragmatic solution is to identify contaminated witnesses and add them later manually to the automatically plotted tree of the non-contaminated ones. However, it is not always easy to find out that a manuscript is contaminated and in some cases, like the *Centiloquium*, most witnesses are contaminated.

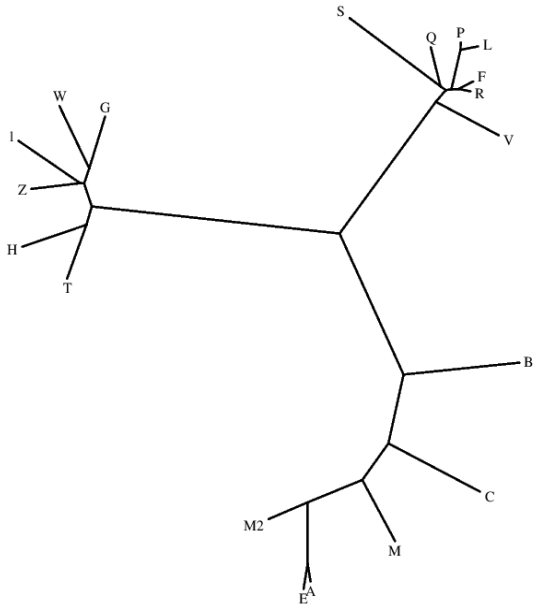


Figure 11: As fig. 7 but including the contaminated text-layer M².

4.2 Rooting

The second serious problem for tree-finding software is rooting. Whereas biologists can usually use an outgroup to root their trees, for instance a shrew for a genealogical tree of primates, this is in most cases not an option in our field as texts are created (more or less) *ex nihilo* by their authors.[32] Procedures using textual distances can by definition not serve for identifying the root: distance is commutative, the information of the direction of change is not contained in the distance data. One would have to add a mathematical notion of direction. In order to decide where the root is located in a tree one needs to know for at least some variants which one is primary and which ones are secondary. It is easy to see that the naive first guess that the root will be in the centre of the tree plot is in general wrong. Fig. 12 shows a constructed example with six passages symbolised by upper and lower-case letters of the alphabet. The obvious idea to start grouping close relatives together again and again in order to bring direction into the stemma does not work in general either (even if we grant that the archetype is not among the witnesses). In the example, one would easily guess that

31 https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/splitstree.

32 In some cases early translations can serve as outgroups, cf. Caroline Macé [Roelli, 2021: ch. 3.2].

the two main groups of readings ABC and bcd are significant variants shared by about half the tradition each and one will be tempted to posit the archetype where they intersect. This would be between α and γ (black line on the right hand side), which is not at all where the archetype is found in truth. We note in passing that in case of a stemma with a bifurcation at the top (a common case, as [Bédier, 1928] showed), the root or archetype is not found at a junction of lines but along one of the lines in the tree, like in this example (red arrow). An idea for future development could be to develop an algorithm that can detect eye-skips and use them to add an element of directionality for parts of the tree by assuming that the texts with the eye-skip represent a derived text-state.

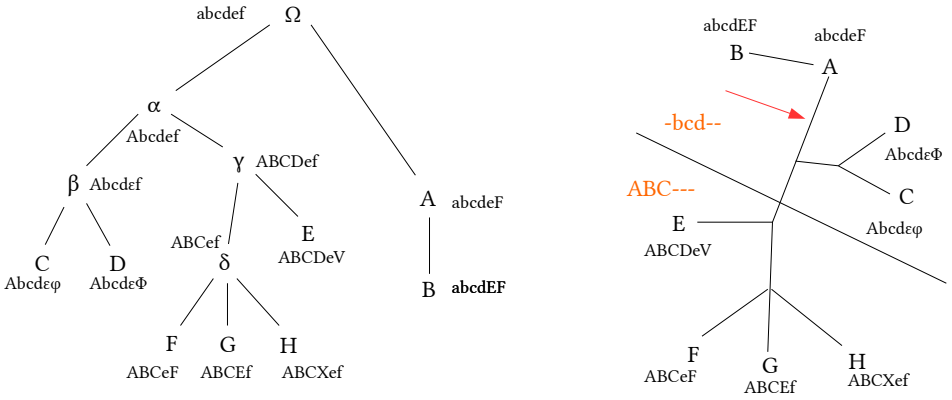


Figure 12: Constructed schematic example: stemma vs. unoriented tree.

V A PRAGMATIC APPROACH FOR THE TIME BEING

The length of the text and the number of manuscripts are the main constraints to determine the time (and thus money) it takes to edit it if we take a skilled editor for granted. Thus it may be that a long text survives in a hundred witnesses but the time and money are only sufficient to use half a dozen manuscripts for the entire text. Similarly the entire variance of a tradition (like the *Centiloquium*) may not be worth recording as it is largely derived and repetitive. In such cases it is crucial to choose the right sample of witnesses for the edition. We can summarise a recommendable approach for texts of abundant traditions: (i) A relatively small text sample – or better two from different parts of the texts – are transcribed from all witnesses, potentially by a team of assistants. (ii) Software can then be used to plot trees (using standardised orthography); together with the other introduced pieces of software (iii) true significant errors can be identified and with them *loci critici*. Analysing these readings only (iv) will help to draw a stemma for the sample texts. (v) A sub-sample of manuscripts can be chosen which contains the most promising ones in each group as well as the solitary and unclear ones. More text is then transcribed from these, more trees are plotted, solid significant errors can be identified and an improved stemma of the sub-sample created. A large percentage of witnesses will often quickly be found to be derivative and of little importance for the reconstruction of the archetypal text (although they may well be interesting for other research questions). (vi) The remaining witnesses can be studied more in depth and a further sub-sample of them can be chosen for the edition. The archetype or root (vii) has to be located manually in the

plotted trees; these will also have to be improved for instance to reflect contamination. With all this data the editor can chose the witnesses (viii) to be used for the edition, from which the archetypal text is then reconstructed as far as possible. Finally it is examined and (ix) *emendatio* may have to be applied to passages that can be shown to have been wrong in the archetype. As this final step is quite speculative, it should be done with circumspection and clearly stated for each passage in the edition so that the reader can decide for himself whether he agrees with the emendation.

For the *Liber Aurelii* edition I decided that I could just as well do the work manually except for the Gariopontus recension where I pragmatically chose five of the oldest manuscripts and the early print from a sample of about a dozen I checked (of the total of some 65). For a critical edition of the entire Gariopontus one will have to study the tradition more in depth but for the *Liber Aurelii* passages (which make up some 12% of Gariopontus' compilation) this seemed good enough as the variation in the Gariopontus witnesses was relatively minor and besides the readings of the full *Liber Aurelii* could be compared. For the *Centiloquium* (having more manuscripts but less variance) software may be of greater use. The approach sketched above can help save time if there are many witnesses of similar length and many of them not contaminated – although there is still a lot of crucial manual work and critical thinking left for the editor. At least the alignment can be done automatically and there is no need to consider readings while transcribing. Indeed, several people can easily transcribe witnesses at the same time. Nonetheless transcription remains the bottle-neck of labour; OCR software for hand-written documents may bring help at this point in the not-too-distant future at least for manuscripts that are nicely written and physically well preserved. Contamination must be identified manually[33] and the tree must be manually rooted.

VI POSSIBLE FUTURE DEVELOPMENTS

OCR software today already uses dictionaries with the help of which *dominus* gets weighted far more strongly than similar looking but nonsensical *dorunus* (if the text is known to be in Latin). This approach could be strongly improved once software can construct syntactic trees automatically and then weight their probabilities. For instance the software would then expect a verb in the subjunctive in an *ut* clause. An automated statistical analysis of vocabulary and style is a further possibility to improve results: if an author always says *scapulae* never *palae* for 'shoulder', an algorithm should consider the first word as the likely primary reading. In case other works of an author are known, their vocabulary and their syntactical trees could additionally be used to weight possible readings or trees more accurately for the author in question. These are basically the things an experienced philologist does intuitively when editing a text. Thus, the goal for software should be to mimic the behaviour of an experienced editor, which may lead to a better understanding of this partly intuitive behaviour and it may in time even improve it. For now computerised methods should strive to approach the gold standard of our reconstructive textual editing – Neo-Lachmannian critical text editing [Trovato, 2017] – one should avoid using software from other fields as black boxes trusting that the results will be fine even though the software was tailored for other problems involving 'descent with modification'.

33 There are special approaches that try to cope with highly contaminated transmissions, such as versions of the Bible: [Mink, 2011].

References

Bédier, Joseph (1928). La Tradition manuscrite du Lai de l'ombre: Réflexions sur l'art d'éditer les anciens textes. *Romania* 54: 161–196, 321–356. [reprint Paris: Champion, 1970. gallica.bnf.fr/ark:/12148/bpt6k8980]

Blake, Norman F. & Jacob Thaisen (2004). Spelling's significance for textual studies. *Nordic Journal of English Studies* 3 (1): 93–108.

Bourgain, Pascale (1992). Sur l'édition des textes littéraires latins médiévaux. *Bibliothèque de l'École des Chartes* 150: 5–49.

Froger, Jacques (1978). Westgöta-Lagen, Edited by H. S. Collin and C. J. Schlyter. Facsimile edition with an addendum by Otto von Friesen, Our Oldest Manuscript in Old Swedish. Edited by Gösta Holm. *Scriptorium* 32/1: 183–188.

Glaze, Eliza (2005). Galen refashioned: Gariopontus in the Later Middle Ages and Renaissance. In: Elizabeth Lane Furdell (ed.). *Textual healing: Essays on Medieval and Early Modern medicine.* Brill (Leiden): 53–75.

Holm, Gösta (1972). Carl Johan Schlyter and Textual Criticism. *Saga och Sed* 1972: 49–80.

Juste, David (2022). Pseudo-Ptolemy, Centiloquium (tr. Plato of Tivoli) (update: 10.03.2022). *Ptolemaeus Arabus et Latinus. Works.* Online: http://ptolemaeus.badw.de/work/41.

Kiessling, Benjamin, Robin Tissot, Peter Stokes, et Daniel Stökl Ben Ezra (2019). eScriptorium: An open source platform for historical document analysis. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 2: 19–24. https://ieeexplore.ieee.org/abstract/document/8893029

Lemay, Richard Joseph (1978). Origin and Success of the Kitāb Thamara of Abū Jaʿfar Aḥmad ibn Yūsuf ibn Ibrāhīm from the Tenth to the Seventeenth Century in the World of Islam and the Latin West'. In: *Proceedings of the First International Symposium for the History of Arabic Science* (Aleppo, April 5-12, 1976). Aleppo University Press (Aleppo): II, 91–107.

Macé, Caroline, Ilse De Vos, and Koen Geuten (2012). Comparing Stemmatological and Phylogenetic Methods to Understand the Transmission History of the 'Florilegium Coislinianum.' In: Alessandra Bucossi and Erika Kihlman. *Ars Edendi Lecture Series.* Stockholm University Library (Stockholm): 2:107–129. http://www.diva-portal.org/smash/get/diva2:551286/FULLTEXT01.pdf

Martorello, Franco, and Giuseppe Bezza (eds.) (2013). Aḥmad ibn Yūsuf ibn al-Dāya. Commento al Centiloquio Tolemaico, Mimesis (Milano).

Mink, Gerd (2011). Contamination, Coherence, and Coincidence in Textual Transmission. In: Klaus Wachtel and Michael W. Holmes. *The Textual History of the Greek New Testament. Changing Views in Contemporary Research.* SBL (Atlanta): 141–216.

Olrik Frederiksen, Britta (2009). Stemmaet fra 1827 over Västgötalagen – en videnskabshistorisk bedrift og dens mulige forudsætninger. *Arkiv för nordisk filologi* 124: 129–150.

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

Roelli, Philipp (2014). Petrus Alfonsi; or, On the Mutual Benefit of Traditional and Computerised Stemmatology. In: Tara Andrews and Caroline Macé (eds). *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, 43–64. Brepols (Turnhout).

Roelli, Philipp (ed.) (2020). *Handbook of Stemmatology. History, Methodology, Digital Approaches.* De Gruyter (Berlin). https://doi.org/10.1515/9783110684384

Roelli, Philipp (2021). Liber Aurelii 'On Acute Diseases', critical edition. *Beihefte zum Mittellateinischen Jahrbuch* 21. Hiersemann (Stuttgart). https://doi.org/10.36191/9783777222035

Schlyter, Carl Johan, and Hans Samuel Collin (eds) (1827). *Westgöta-Lagen.* Häggström (Stockholm).

Spencer, Matthew, Linne Mooney, Adrian Barbrook, Barbara Bordalejo, Christopher Howe, and Peter Robinson (2004). The Effects of Weighting Kinds of Variants. In: Pieter van Reenen, August den Hollander, and Margot van Mulken (eds). *Studies in Stemmatology II*, 227–240. Benjamins (Philadelphia).

Trovato, Paolo (2017). *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 (Padova).