e-ISSN 2724-3923

magazén
Vol. 6 — Num. 2 — Dicembre 2025

LLM-Mining Pre-Stemmatological Philological Literature

Armin Hoenen
Goethe Universität Frankfurt am Main, Germany

Abstract The current article outlines a new research avenue for the analyses of literature from the time before the advent of the stemmatic method in the nineteenth century using large collections of digitized images and texts of historical philological works. The main aim is to understand the dynamics behind the processes leading to the invention of said method. The proposed steps are object recognition (image analysis) with textual clues and relation extraction (text mining). Proof-of-concept-level experiments demonstrate the applicability.

Keywords Stemmatology. Object recognition. Llms. Yolo. Text mining.

Edizioni Ca'Foscari

Peer review

Submitted 2025-09-30
Accepted 2025-11-28
Published 2026-01-07

Open access

Citation Hoenen, A. (2025). "LLM-Mining Pre-Stemmatological Philological Literature". *magazén*, 6(2), 215-232.

DOI 10.30687/mag/2724-3923/2025/02/005

1 Introduction

The current article has two main objectives: demonstrate applications of Large Language Models to stemmatological research (digital [humanities] objective) and outline a research avenue for multimodal (image, text) analytic distant reading of large corpora ([digital] humanities objective). Large collections of recently digitised written sources could be used to explore philological literature from all ages. In this article, proof-of-concept (poc) level experiments demonstrate the feasibility of object recognition and text mining with the objective to explore, quantify and by these tokens improve our understanding of the prehistory of the stemmatic method.

Stemmatology is a subfield of philology occupied with textual evolution. It aims at a visual representation of the history of textual variants. The stemmatological methodology which may include text reconstruction and which is often connected with the name of Karl Lachmann (Trovato 2017) is especially useful for texts which originated in the chirographic age where it sometimes emends towards a more original or authentic text form. This in turn is closely tied to editing. Today, computational methods can be applied and are partly shared with the sister disciplines of phylogenetics/systematics (biology) and historical linguistics (Hoenen 2020).

The relationship and mutual influences of these fields date back much further than to the age of computation. These mutual influences have been variously analysed. Interestingly, all three witness their oldest trees and the onset of tree-drawing in about the same time period, the early-mid nineteenth century:

- Biology: after Darwin published a tree in *The origin of species* 1859 there was a "great burst of tree-making" (O'Hara 1996, 85);
- Linguistics: "The first genuine tree diagram of the history of Indo-European was apparently published around 1800 (Auroux 1990), but linguistic trees of history didn't really become widespread until the 1850s" (O'Hara 1996, 84);
- Philology: Collin and Schlyter (1827) were the first to publish a stemma. Shortly thereafter "Carl Zumpt published a genealogy of the known copies of *Ciceros Verrine Orations* in 1831, and Zumpt's stemma was followed by stemmata drawn by Friedrich Ritschl in 1832, and by J.N. Madvig in 1833" (O'Hara 1996, 85).

The similar time range is somewhat striking, all the more since collaborations between these disciplines, despite existent, are usually not understood as the main driving-force of the epochal changes. What is rather uncontroversial though is the benefit tree-drawing meant.

2 The Enormous Benefit of a Visualisation

Whilst language forces us to express concepts one by one forcing our reasoning into one sequence, the visual domain of a stemma is not that restricted. Describing the relationships between witnesses in words is thus more inherently ambiguous than displaying a clear and simple stemma. And this goes for any tree. Taking a simple example such as the tree **[fig. 1]**, one could describe the relationships in words and find many different narrative sequences. An example could begin as 'From a now believed to be lost archetype, only two copies were made. The descendent of one, A, is now at the royal library of Sweden, …'. At this point alone, language would force the author to decide whether to first mention the other copies of root (breadth-first) or to describe the descendants of A had it had some (depth first). The same would go for every witness node in the stemma and naturally one could jump back and forth between breadth-first and depth-first, between bottom-up and top-down or even jump wildly. The point is that many possible narratives map to the same stemma. If now, in addition to this, authors write about the same tradition with different views of relationships or get iffy, it will become much more difficult to compare two narratives than it is to juxtapose and compare two trees.

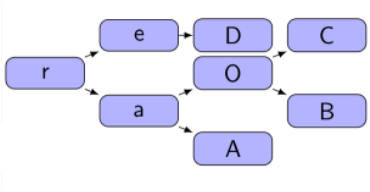


Figure 1
A stemma-like structure with descriptii, capitals for surviving witnesses

The complex genealogical information is much more digestible if displayed as a visualisation rather than if presented as a narrative. The vast success of the tree is in part due to this effect of allowing a simple overview over complicated relations. Scientific exchange is fostered hereby. We shall call the effect visual simplification effect (VSE).

Given the VSE, a valid question is why a tree-like structure for the analysis of mutating units has not been invented before, especially in philology, since staggering amounts of textual variation were known much earlier. Collations, that is side-by-side representations of texts are known very early on, for instance in China where textual criticism is traced back to Liu Xiang (first century BCE) (Fölster, Staack 2021). Although writing system evolution complicated emendation in China, woodblock printing appeared already in the Tang era, around the seventh or eighth century (Barrett 2001) and

would have made stemmata useful for editors. In Western antiquity, the library of Alexandria is known to have hosted many versions of the Homeric epics, compare Nagy (2004). These texts were rather invented orally and may not have one clearly defined original in the same way as born-written works would. Alas, no stemma is known from Alexandria.

In holy texts, variation was present already early on, compare the Qumran manuscripts for instance (Tov 2018). One reaction was that the importance of strict copying for copyists in the Tannaim group (Wegner, 2006, 73) was emphasized, but again, no stemma is known.

Also colophons came into being, recording the local copy histories of single witnesses, but colophons were also copied, sometimes omitted etc. and did not contain any stemmata. Yet, trees as analytic structures have been used since antiquity in a plethora of ways after all, comprising so diverse subjects as the depiction of the descent of aristocratic families and trees of virtues connecting desirable personality traits (Lima 2014). It could thus have been a small step for an early author to transfer this structure.

Being far from exhaustive, this at least shows that there were many possible places where an earlier stemma or graph would already have been useful in order to exploit the VSE. Similar to devices such as the steam engine described by Heron of Alexandria (Roby 2023) or the Baghdad battery (Keyser 1993) there might have been a graph-like structure for text versions ahead of its time which then remained isolated.

Which experiments or research could we conduct to find such a graph, if it had been overlooked? Before diving into this, let us look at the first stemma itself and analyse the invention and the prerequisites.

3 The Invention of the First Stemma

What was the actual reason for Collin and Schlyter to invent their stemma, which they put only into an appendix? It appears, references of philological discourse were rather scarce in their work, but they had a source for the texts they analysed which displayed some family trees of Old Nordic aristocratic families: Fant (1818). Their calling their stemma "et slags stamtre" (Swe. 'a kind of inheritance tree') points to these depictions as their primary inspiration. Their struggle with the terminology, as they called the stemma *schema cognationis* in Latin, corroborates the hypothesis that they had not seen a stemma like graphical depiction for text evolution whatsoever before or something similar. In their case, the coincidental appearance of family trees in one of the secondary sources of their research for the compilation of an edition was just enough to cause the idea of a stemma.

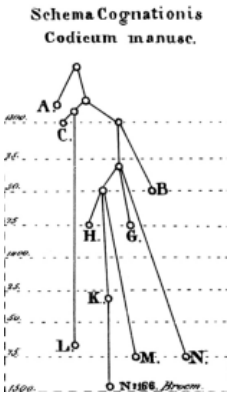


Figure 2
The probably first recorded stemma codicum by Collin, Schlyter 1827

Collin and Schlyter did probably not influence others. What an irony: the mostly tree-shaped stemma has apparently evolved more than once and the visualisation of the genealogy of evolutionary trees in science would thus have multiple roots and be no tree or DAG in a strict sense.

Their invention builds upon previous works and a meticulous collection and analyses of catalogued manuscripts, but in principle, similar conditions could have existed much earlier at least for some works.

4 Was there a More Ancient Stemma or Graph-Like Visualisation?

With the advent of large collections of globally available digitised ancient sources, textual such as the Patrologia Latina (Migne 1993, 1998) or including images and standardized access (IIIF) such as via the VeDPH at Ca' Foscari and on the other hand an impressive increase in technological image recognition capabilities, the time seems right to approach this question with the help of image technology.

LLMs combined with vision encoders could be used without fine-tuning for stemma object recognition. A more conservative approach would be to train an object recognition model for stemmata. Both methods could be used to scan large digital collections for early stemmata. The image recognition could additionally be combined with text extraction. In order to explore if this technology could work, we conduct some poc-level experiments in order to determine whether such an endeavour could be feasible and which technologies seem most promising.

4.1 Dataset and LLM

A small dataset was created, containing:

- one family tree from Fant (1818)
- 125 pages from Collin and Schlyter (1827) with only one page containing the first stemma
- 50 synthetic pages with stemmata and pseudo-text
- Additionally, thanks to the project Open Stemmata (Camps et al. 2021) a corpus of publicly available papers from Persée was available for the experiments containing 61 papers featuring 66 Stemmata and 81 other stemma-like diagrams (false positives)

In the Python programming language, the library *graphviz* was used to generate random trees. These were then placed at a random position onto an artificially generated page with pseudo-text. These pages are not exactly like the ones which occur in the target philological literature, furthermore, text is quadratic and has ragged ends but the appearance is not entirely dissimilar to target structures **[fig. 3]**. For now, if already the poc on synthetic data alone fails, the endeavour would probably be not worth the time and effort. The dataset was forwarded to *GPT4o-mini* via the API from *openAI* for recognition. The prompt combined a role, and some information on how to combine textual and visual evidence.

In a second run, the text of the corresponding page of the image was combined with the image for the Persee dataset where each page was saved as a separate png and each text per page correspondingly (roughly 2500 instances). However, 9 images of graphs had to be excluded due to their being so blurred that even human eyes were not able to distinguish, what kind of diagram that might be.

4.2 Object Recognition with YOLO

LLMs tend to be slow and demanding in hosting. It might be, that one instead want to train one's own object recognition model. Aouinti et al. (2021) used the predominant Object recognition technology YOLO for the detection of illumination. The next poc technique was thus training a YOLO model for stemma recognition. Synthetic training data was generated in the same way as above (1,000 train, 100 val set instances). Additionally, various data augmentation techniques inherent in Yolo were tested, such as rotating by a random angle, overlay, and so forth. In the best condition however training without the augmented examples performed best. In the step placing the trees onto the random text pages label files were generated contemporaneously, indicating the stemma object position by rectangle coordinates. This was enough to train a model with

yolov5. This model was then used to predict all pages of Collin and Schlyter (1827) and the Handbook of stemmatology (Roelli 2020), and the Persee dataset.

4.3 Results

The LLM (GPT4o) recognized trees and distinguished between an ordinary family tree and stemmata (synthetic and real). On the Persee data, the LLM was able to achieve a recall of 0.98, but since almost as many false positives were identified as stemmata, the precision was at a mere 0.44. Given that some of the data were graphically blurred, before the more exact distinction between stemmata and other graphs can be made, the dataset must be improved. The high recall together with the huge number of more than 1400 pages of true negatives which were without any error detected, shows that the LLM is able to recognize graphs. In so far, the poc supports the claim, that a more thorough project set-up will achieve suitable recognition ratios using LLMs.

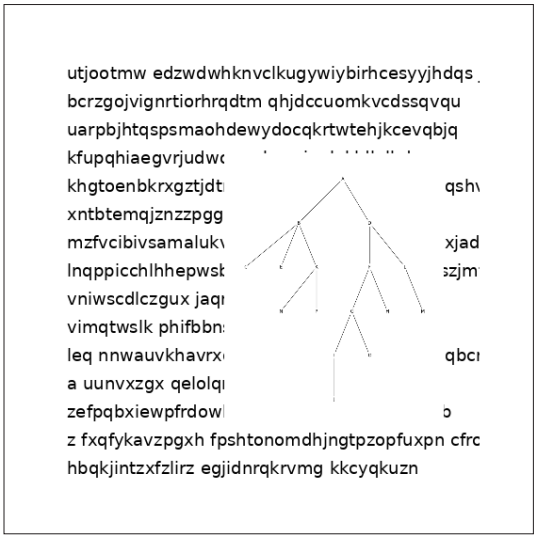
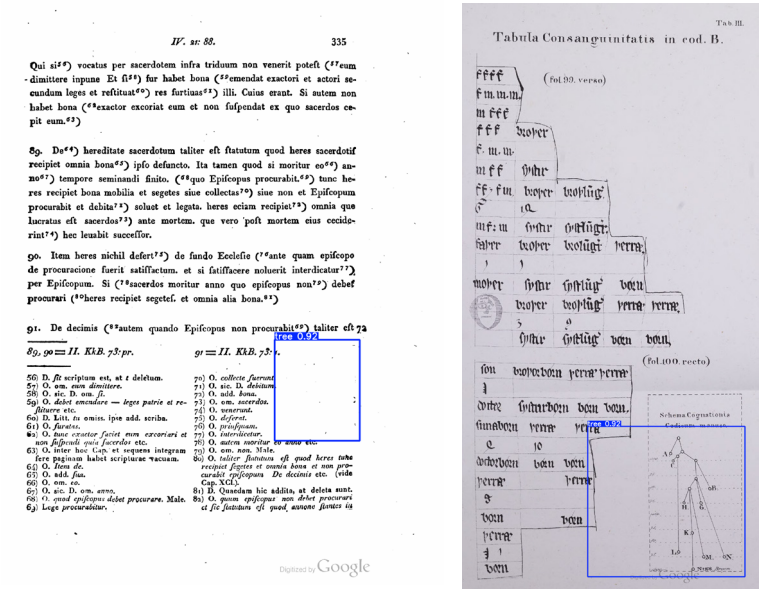


Figure 3 A synthetically generated stemma codicum on a pseudo text page



Figures 4a-b Yolo object recognition, left, an artifact (p. 441) and right, the original true stemma (p. 703)

The YOLO model on the other hand had learned some artefacts of the synthetic data leading to the recognition of rectangles in the margins. This was partly because the graphs with a white box background had been placed onto the texts leaving some text to all sides, which is not matched in authentic texts. Yolo computes the probability of its objects and it is both easy to filter for the margins by the coordinates of the recognized objects and by the probability. Excluding empty pages, objects in the margins and objects below 0.9 probability, the model recognized the true stemma, see Figure 4 plus 3 false positives. The recognition boundaries of the original stemma were a bit distorted and include the unusual lines of the adjacent table. Double checking recognition on all 694 pages from the Handbook of stemmatology, which contains many different kinds of stemmata and graphs, the model does truly recognize stemmatic structures even if not strictly trees and such which are not extremely similar to its input, see Figure 5. At 0.9 probability threshold, on the Persee dataset, the performance dropped to 0.35 precision and a mere 0.1 recall. Many stemmata were not recognized, false positives outnumbered true positives, but true negatives were correctly matched. The model itself being only trained with synthetic data has of course utter limitations, especially since almost none of the persee stemmata looked like the ones, the model had seen during training. The results still point to an applicability because Yolo is known to be a powerful model and

e-ISSN 2724-3923

because it distinguished true negatives well. The same problem as with the LLM might occur, namely that the distinction from other diagrams must be well trained towards and even a combination with text might not be working. This is a challenge if one targets somewhat creative stemmata of the past, the appearance of which is unclear, if they exist.

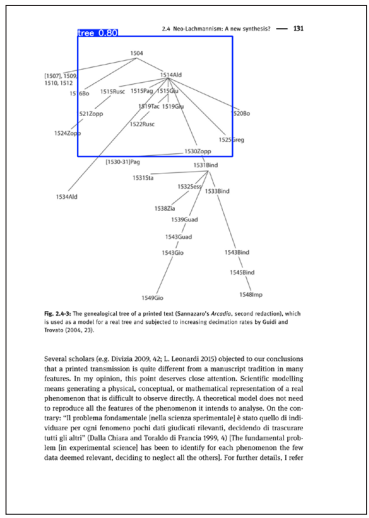
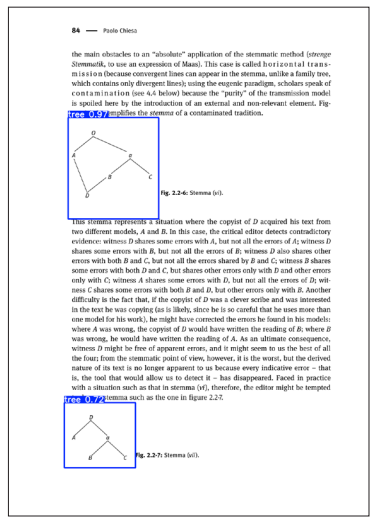


Fig. 2.4-3: The genealogical tree of a printed text (Sannazaro's *Arcadia*, second redaction), which is used as a model for a real tree and subjected to increasing decimation rates by Guidi and Trovato (2004, 23).

Several scholars (e.g. Divizia 2009, 42; L. Leonardi 2015) objected to our conclusions that a printed transmission is quite different from a manuscript tradition in many features. In my opinion, this point deserves close attention. Scientific modelling means generating a physical, conceptual, or mathematical representation of a real phenomenon that is difficult to observe directly. A theoretical model does not need to reproduce all the features of the phenomenon it intends to analyse. On the contrary: "Il problema fondamentale [nella scienza sperimentale] è stato quello di individuare per ogni fenomeno pochi dati giudicati rilevanti, decidendo di trascurare tutti gli altri" (Dalla Chiara and Toraldo di Francia 1999, 4) [The fundamental problem [in experimental science] has been to identify for each phenomenon the few data deemed relevant, deciding to neglect all the others]. For further details, I refer



the main obstacles to an "absolute" application of the stemmatic method (*stronge Stemmatik*, to use an expression of Maas). This case is called horizontal transmission (because convergent lines can appear in the stemma, unlike a family tree, which contains only divergent lines); using the eugenic paradigm, scholars speak of contamination (see 4.6 below) because the "purity" of the transmission model is spoiled here by the introduction of an external and non-relevant element. Figure 2.2-6 simplifies the *stemma* of a contaminated tradition.

This stemma represents a situation where the copyist of *D* acquired his text from two different models, *A* and *B*. In this case, the critical editor detects contradictory evidence: witness *D* shares some errors with *A*, but not all the errors of *A*; witness *D* shares some errors with *B*, but not all the errors of *B*; witness *D* also shares other errors with both *B* and *C*, but not all the errors shared by *B* and *C*; witness *B* shares some errors with both *D* and *C*, but shares other errors only with *D* and other errors only with *C*; witness *A* shares some errors with *D*, but not all the errors of *D*; witness *C* shares some errors with both *B* and *D*, but other errors only with *B*. Another difficulty is the fact that, if the copyist of *D* was a clever scribe and was interested in the text he was copying (as is likely, since he is so careful that he uses more than one model for his work), he might have corrected the errors he found in his models: where *A* was wrong, the copyist of *D* would have written the reading of *B*; where *B* was wrong, he would have written the reading of *A*. As an ultimate consequence, witness *D* might be free of apparent errors, and it might seem to us the best of all the four; from the stemmatic point of view, however, it is the worst, but the derived nature of its text is no longer apparent to us because every indicative error – that is, the tool that would allow us to detect it – has disappeared. Faced in practice with a situation such as that in stemma (vi), therefore, the editor might be tempted to adopt a stemma such as the one in figure 2.2-7.

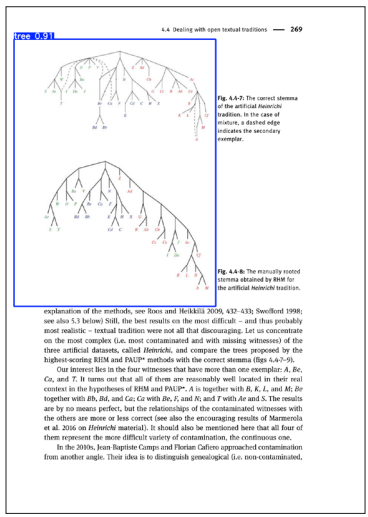


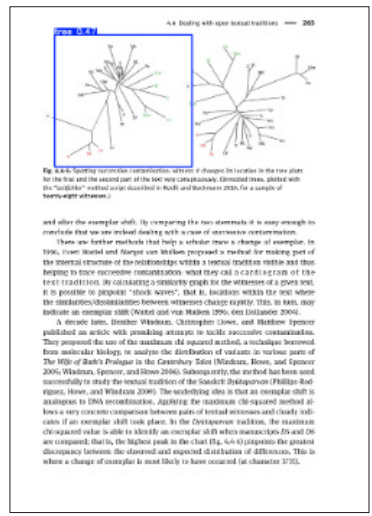
Fig. 4.4-7: The correct stemma of the artificial *Heinrichi* tradition. In the case of mixture, a dashed edge indicates the secondary exemplar.

Fig. 4.4-8: The manually rooted stemma obtained by RHM for the artificial *Heinrichi* tradition.

explanation of the methods, see Roos and Heikkilä 2009, 432–433; Swofford 1998; see also 5.3 below). Still, the best results on the most difficult – and thus probably most realistic – textual tradition were not all that discouraging. Let us concentrate on the most complex (i.e. most contaminated and with missing witnesses) of the three artificial datasets, called *Heinrichi*, and compare the trees proposed by the highest-scoring RHM and PAUP* methods with the correct stemma (figs 4.4-7–9).

Our interest lies in the four witnesses that have more than one exemplar: *A*, *Be*, *Co*, and *T*. It turns out that all of them are reasonably well located in their real context in the hypotheses of RHM and PAUP*. *A* is together with *B*, *K*, *L*, and *M*; *Be* together with *Bb*, *Bd*, and *Ge*; *Co* with *Be*, *F*, and *N*; and *T* with *Ae* and *S*. The results are by no means perfect, but the relationships of the contaminated witnesses with the others are more or less correct (see also the encouraging results of Marmerola et al. 2016 on *Heinrichi* material). It should also be mentioned here that all four of them represent the more difficult variety of contamination, the continuous one.

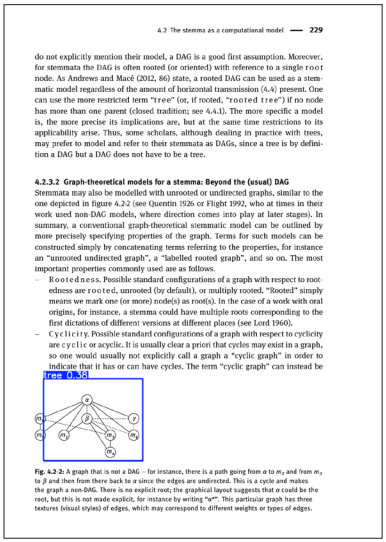
In the 2010s, Jean-Baptiste Camps and Florian Cafiero approached contamination from another angle. Their idea is to distinguish genealogical (i.e. non-contaminated,



and after the exemplar shift. By comparing the two stemmata it is easy enough to conclude that we are indeed dealing with a case of successive contamination.

There are further methods that help a scholar trace a change of exemplar. In 1936, Isaac Boekel and Margot van Mulken proposed a method for making part of the inverted structure of the relationships within a textual tradition visible and thus helping to trace successive contamination: what they call a cardiogram of the text tradition. By calculating a similarity graph for the witnesses of a given text, it is possible to pinpoint "shock waves", that is, locations within the text where the similarities/dissimilarities between witnesses change rapidly. This, in turn, may indicate an exemplar shift (Waitzel and van Mulken 1996, den Uitlander 2004).

A decade later, Heather Windram, Christopher Howe, and Matthew Spencer published an article with promising attempts to tackle successive contamination. They proposed the use of the maximum chi squared method, a technique borrowed from molecular biology, to analyse the distribution of variants in various parts of *The Wife of Bath's Prologue* in the Canterbury Tales (Windram, Howe, and Spencer 2006; Windram, Spencer, and Howe 2006). Subsequently, the method has been used successfully to study the textual tradition of the Sanskrit *Śyāmaparvan* (Phillips Rodriguez, Howe, and Windram 2008). The underlying idea is that an exemplar shift is analogous to DNA recombination. Applying the maximum chi-squared method allows a very concrete comparison between pairs of textual witnesses and clearly indicates if an exemplar shift took place. In the *Śyāmaparvan* tradition, the maximum chi-squared value is able to identify an exemplar shift when manuscripts D5 and D6 are compared; that is, the highest peak in the chart (fig. 4.4-6) pinpoints the greatest discrepancy between the observed and expected distribution of differences. This is where a change of exemplar is most likely to have occurred (at character 3735).



4.2 The stemma as a computational model — 229

do not explicitly mention their model, a DAG is a good first assumption. Moreover, for stemmata the DAG is often rooted (or oriented) with reference to a single `root` node. As Andrews and Macé (2012, 86) state, a rooted DAG can be used as a stemmatic model regardless of the amount of horizontal transmission (4.4) present. One can use the more restricted term "`tree`" (or, if rooted, "`rooted tree`") if no node has more than one parent (closed tradition; see 4.4.1). The more specific a model is, the more precise its implications are, but at the same time restrictions to its applicability arise. Thus, some scholars, although dealing in practice with trees, may prefer to model and refer to their stemmata as DAGs, since a tree is by definition a DAG but a DAG does not have to be a tree.

4.2.3.2 Graph-theoretical models for a stemma: Beyond the (usual) DAG

Stemmata may also be modelled with unrooted or undirected graphs, similar to the one depicted in figure 4.2-2 (see Quentin 1926 or Flight 1992, who at times in their work used non-DAG models, where direction comes into play at later stages). In summary, a conventional graph-theoretical stemmatic model can be outlined by more precisely specifying properties of the graph. Terms for such models can be constructed simply by concatenating terms referring to the properties, for instance an "unrooted undirected graph", a "labelled rooted graph", and so on. The most important properties commonly used are as follows.

- `Rootedness`. Possible standard configurations of a graph with respect to rootedness are `rooted`, unrooted (by default), or multiply rooted. "Rooted" simply means we mark one (or more) node(s) as root(s). In the case of a work with oral origins, for instance, a stemma could have multiple roots corresponding to the first dictations of different versions at different places (see Lord 1960).
- `Cyclicity`. Possible standard configurations of a graph with respect to cyclicity are `cyclic` or acyclic. It is usually clear a priori that cycles may exist in a graph, so one would usually not explicitly call a graph a "cyclic graph" in order to indicate that it has or can have cycles. The term "cyclic graph" can instead be

`tree (4.3)`

Fig. 4.2-2: A graph that is not a DAG – for instance, there is a path going from α to m_3 and from m_3 to β and then from there back to α since the edges are undirected. This is a cycle and makes the graph a non-DAG. There is no explicit root; the graphical layout suggests that α could be the root, but this is not made explicit, for instance by writing "α^{root}". This particular graph has three textures (visual styles) of edges, which may correspond to different weights or types of edges.

Figures 5a-e
Recognition of different types of stemmata from the Handbook of stemmatology through the model ranging from similar to the training to quite different

All in all, the poc has shown that a larger scale object recognition for larger collection seems feasible.

5 Recapitulation of Historical Processes towards Text Mining

What is similar among the three sciences using trees presumably is a larger overall increase in amounts of data that they had to analyse. The reasons for these increasing amounts of data were presumably at least partly different for the three. Colonialism expanded numbers of known species and the knowledge of languages considerably. For philology, however, the main reason for an increase in data should primarily have been the invention of the printing press.

As soon as an editor had to print an ancient work, transmitted in handwriting, he could ask which version of the slightly differing versions at his disposal he should use. The key question and probably the birth helper of stemmatology. Readers would naturally prefer and read that edition which could plausibly offer the best possible version exerting a certain pressure on early print age editors. For choosing, of course, an editor would have to have access to multiple versions – prerequisite 1 – and the philological insight that some versions might be more authentic than others – prerequisite 2.

As for the availability of versions, it should have depended on many factors such as an increased mobility, superior cataloguing and less circulation of the manuscripts themselves. This was only

gradually happening after printing had been invented. In the early days of printing, it is logical that handwritten manuscripts remained in circulation and thus harder to access for printing and throughout the sixteenth century books were still rather rare as compared to today, consider Pettegree (2010). After all, one needed people able to read them and broader literacy through schooling only slowly but steadily advanced, compare for instance Eskelson (2021) on literacy development. At some point in time, printed books must then have become the norm for private and public reading, whilst handwritten manuscripts, codices etc. were becoming less common. The time frame for these processes can roughly be estimated to be during the seventeenth and eighteenth centuries. For an in depth analysis based on the outputs of printing presses, see Buringh, van Zanden 2009. Improved cataloguing in libraries (compare e.g. the first printed catalogue of the Bodleian in 1605; Bodleian 1605/1986) falls in the same range. Thus it is safe to assume that there was a steady increase in available sources for editors.

At the same time, the awareness of variation and how to deal with it in philological discourse increased and concepts such as the shared innovation or error were elaborated upon. In fact, many mechanisms of the stemmatic method and of emendation have been understood since antiquity, compare Haverling (2020). However, putting them together systematically into a rigorous method which is known, learned, practised and taught at least in part of the field ever since falls into the nineteenth century (Haugen 2020, 57) clearly coinciding with the development of the stemmatological method.

These prerequisites could be very different in linguistics and biology. Another important peculiarity for philology is that a single tradition, a text, is a rather isolatable unit. There is no such thing as a tree of all texts. In linguistics and biology however scientists engaging in the analysis of whatever evolutionary entities (clams, canines, felines, … or Indo-European languages etc.) would additionally have to deal with the question of how to accommodate their units into a tree of life or of all languages (if one believes in one language origin). The 'laboratories' of editors are smaller improving chances for an earlier holistic graphical approach.

If it were only for the awareness of change and the availability of versions, one could argue for holy texts evidence has been tantamount ever since, even before printing. However, in that case, two thoughts might help to explain why holy texts witnessed an independent development within philology/stemmatology. On the one hand many other aspects than only micro-variation on a linguistic level would play a role when going towards an urtext, exegesis, the implications of the text. Emendation would be difficult to explain. On the other, for holy texts there was so much evidence that despite the idea of a genealogical tree for the New Testament being expressed by

Bengel (1763) the setting was so complicated that it simply took much longer than in the classics. Here, available evidence was increasing just as much as that for certain works stemmatic relations became too complex to be easily comprehensible by words alone and not complex enough to refrain from attempts to approach the entirety of the evidence graphically.

Given these assumptions, the relatively similar time frame of occurrence of the tree-drawing branches of the three sciences of biology, linguistics and philology could be at least in part incidental. All three saw an increasing amount of data and discourse, whilst the reasons for the increases might have been different. In order to validate such hypotheses, quantitative analyses would be needed. One way could be to use text mining to measure the amounts of witnesses editors used over time and how their relations have been analysed.

6 Text Mining and Information Extraction

In this experiment, relation extraction from secondary literature, especially from philological literature for the time before the invention of stemmata is being investigated. First, a small artificial corpus of editorial descriptions of the same stemma-like structure is generated, then GPT4-o is used with an appropriate prompt. The task more precisely is, from differing textual descriptions of the same tree-like structure to retrieve that structure and display it in an unambiguous format. As a target format, we choose the Newick format.[1] Previous experiments to make the LLM generate a graph directly had led to less usable results.

First, we choose some stemma-like structure, seen in Figure 1. Then, we generate textual descriptions for this structure. We do this by first defining 4 base sentences for each parent. '*O has been copied twice, once into C, once into B*.' could be one such sentence. For each of the sentences, we manually create 5 or 7 alternative formulations. Each text shall feature a variant of each of the four sentences. In this way the entire structure is described. The sequence however may wildly differ, as is normal for human descriptive text. We generate 100 distinct sequence permutations of the four sentences and randomly fill each with a variant. To round up the text, we add some introductory and final text without structural implications. Finally, we insert so-called distractor sentences, that is sentences which do not bare information on the direct relationships, we insert them via

1 See the full definition here: https://phylipweb.github.io/phylip/newicktree.html.

a randomizer. Such a sentence may read *'O was not copied from e.'* and can be inserted anywhere in the text. 44 times such a distractor was inserted. A full example of a generated text would be:

> This text treats the tradition of Rabanus Testus Textus. The text has been transmitted in handwriting. We have located 5 extant copies in various libraries. e was copied, the copy is D. O and A are closely related, probably they have been copied from the same lost manuscript a. O was not copied from e. The archetype r was copied into e and a. O has been copied into B and C. The tradition is thus a limited one in size and scope but the relations are quite clear leading to a wonderful stemma albeit with descriptii and chains of hypothetical nodes.

The expected target structure should be 'r(e(D),a(O(C,B),A)' or any equivalent.

As for the prompt engineering, we chose different approaches. We started with a basic zero-shot approach asking GPT4-o to 'Extract the Newick tree from this text:'. Then we tried a one-shot prompt with a smaller example, we tried chain-of-thought (cot) prompting (Wei et al. 2022), a technique where the task is broken down into subsequent smaller steps, the LLM solves step-by-step. Here, we asked GPT4-o to first extract relations (edges) and then from the relations to build a stemma. Finally, we tried a two-shot scenario. We varied one-shot scenarios as to whether the example was given within the prompt or whether we simulated a conversation as user-assistant interaction (interactive). Finally, we prompted for non wordiness, that is prompted to provide only the tree, no explanation or anything else. For results see Table 1.

The results suggest that zero-shot and cot alone are not enough. This might be so, because given GPT's training data, the task is relatively unusual. However, more shots improve the result significantly consistent with state-of-the-art research on LLMs. With two shots more than 90% of texts lead to the entirely correct extraction. Given that we did not optimize prompt engineering as to wording etc., this is a very good result. Interestingly, the interactive one-shot scenario performed noticeably worse than the non-interactive example. Also cot alone achieved a better result than zero-shot, but adding an example, this was turned around. In order to investigate these effects more data would be required. The format requirement to provide only the tree, not a wordy answer was adhered to.

Table 1 Stemma extraction performance of different prompting approaches

Method	Hits	Misses	Accuracy (%)
zero-shot	6	94	0.06
one-shot	89	11	0.89
cot	9	81	0.09
one-shot interactive	67	33	0.67
two-shot	94	6	0.94
cot with one-shot	82	18	0.82

The distractors did not affect the results, there was no meaningful difference between the accuracies for texts with as opposed to texts without distractors. In a certain sense, the introductory and final phrases were also distractors. Their differing position however suggests that at least placement of such a distractor has few influence on extraction.

The implication of the experiment is that LLMs in stemmatology could be used in the future to compare older but also more recent philological literature and extract stemmata from texts even where the editorial approach may be opposed to stemmata. The number of nodes is the number of witnesses. Especially for historical descriptions of traditions which may also mention and describe witnesses which have later been lost and for older literature at scale this approach of information extraction could lead to new insights. The good results also would point to possibly related tasks such as a binary classification if two texts are equivalent in the tree-structure they describe or not and a task where from a tree, a textual description can be generated for instance for visually impaired readers. Technically, the experiment belongs to the field of text mining or more precisely information extraction. An overview of applications in biology can be found in Farrell et al. 2022. Fine-tuning, dataset simulation and so forth are ultimately other pathways for research in this direction.

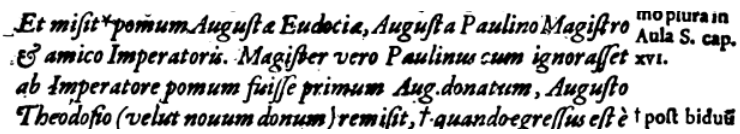
6.1 Application to Real Text

Finally, the previous experiment is again only operating on synthetic data and whilst image recognition already showed that with synthetic data alone, good results can be achieved, here applicability to true data should at least be tested. As a case study we use the *Chronicon Alexandrinum*, a Greek chronicle spanning Creation to Byzantine Emperor Heraclius published in Latin by Matthaeus Raderus (Munich, 1615). Although it is not perfect for our purposes as chonicles have transmission peculiarities, the time range is the right one. The meta-text includes sources like Eusebius, Africanus, Epiphanius, etc.

The images of the Google Book were loaded into GPT4o-mini

alongside a prompt: '*Correcting small OCR inconsistencies, analyze this text and extract from the text any relations between manuscripts or versions in a structured machine readable way.*' An example input page on which this operated, in order to demonstrate how distorted Latin OCR for these texts can be, which is rather the norm than the exception.

Original from p.742 in the pdf (excerpt):



Underlying OCR:

> Etmifit*pomumAugustaEudocia,AugustaPaulinoMagiftroAula S.cap.
> &amico Imperatoris. Magiſter vero Paulinus cum ignoraffet xv1.
> ab Imperatorepomumfuiſſeprimum Aug.donatum,Augusto Theodofio (velut nouumdonum)remifit,&quandoegreſſus estètpoſt biduű

Normalized Version by prompt (showing the underlying ability of the LLM to master such inconsistencies, the note in the margin was however not recognized as such):

> Et misit pomum Augusta Eudocia, Augusta Paulino Magistro Aulae S. cap.
> & amico Imperatoris. Magister vero Paulinus cum ignorasset XVI ab Imperatore pomum fuisse primum Aug. donatum, Augusto Theodosio (velut novum donum) remisit,, quando egressus est et post biduum

The experiment showed that LLMs of the size of GPT4o-mini are able to handle distorted OCR and Latin when extracting copy or citation relations which is an important addition to the first text extraction experiment based only on synthetic English.

7 Conclusion

Poc-level experiments have demonstrated the potential of LLMs and other Machine Learning Models in analysing large collections of digitized data in order to elucidate the pre-history of stemmatology. Object recognition could find and analyse graphical precursors and

possibly even earlier stemmata, whilst text mining methods such as stemmatic relation extraction could trace witness availability and amounts of discourse on witness relations. This could help understand the processes at work in the appearance of tree-drawing in philology across languages and time periods. Data, scripts including prompts and a yolo model have been released on https://github.com/ArminHoenen/prehistorical_stemmata.

Bibliography

Aouinti, F.; Eyharabide, V.; Fresquet, X.; Billiet, F. (2022). "Illumination detection in IIIF medieval manuscripts using deep learning". *Digital Scholarship in the Humanities*. https://academic.oup.com/dsh.

Auroux, S. (1990). "Representation and the Place of Linguistic Change Before Comparative Grammar". de Mauzo, T.; Formigari, L. (eds), *Leipzig, Humboldt, and the Origins of Comparativism*. Amsterdam: John Benjamins, 213-38.

Barrett, T.H. (2001). "Woodblock dyeing and printing technology in China, c. 700 A.D.: The innovations of Ms. Liu, and other evidence". *Bulletin of the School of Oriental and African Studies*, 64(1), 85-9.

Bengel, J.A. (1763) *D. Io. Alberti Bengelii Apparatus criticus ad Novum Testamentum*. 2nd ed. Tubingae (Tübingen): Sumtibus Io. Georgii Cottae.

Bodleian Library (1605/1986). *The First Printed Catalogue of the Bodleian Library, 1605: A Facsimile*. Comp. by T. James. Oxford: Clarendon Press.

Buringh, E.; van Zanden, J.L. (2009). "Charting the 'rise of the West': Manuscripts and Printed Books in Europe, a Long-term Perspective from the Sixth Through the Eighteenth Centuries". *Journal of Economic History*, 69(2), 409-45. https://doi.org/10.1017/S0022050709000837.

Camps, J.-B.; Gabay, S.; Fernández Riva, G. (2021). *Open Stemmata: A Digital Collection of Textual Genealogies = EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities* (Krasnoyarsk, 21-25 September 2021).

Collin, H.S.; Schlyter, C.J. (1827). *Corpus Iuris Sueo-Gotorum Antiqui 1*. Stockholm: Z. Haggestrom.

Eskelson, T.C. (2021). "States, institutions, and literacy rates in early-modern Western Europe". *Journal of Education and Learning*, 10(2), 83-92.

Fant, E.M. (ed.) (1818). *Scriptores rerum Svecicarum medii aevi ex schedis praecipue Nordinianis collectos, dispositos ac emendatos*, Tomus I. Upsaliae: Zeipel et Palmblad (Reg. Acad. Typographi).

Farrell, M.J.; Brierley, L.; Willoughby, A.; Yates, A.; Mideo, N. (2022). "Past and Future Uses of Text Mining in Ecology and Evolution". *Proceedings of the Royal Society B*, 289(1975), 20212721.

Fölster, M.J.; Staack, T. (2021). "Collation in Early Imperial China: From Administrative Procedure to Philological Tool". Quenzer, J.B. (ed.), *Exploring Written Artefacts*, 889-912. Berlin: De Gruyter.

Haugen, O.E. (2020). "2 The genealogical method". Roelli, P. (ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches*. Berlin: De Gruyter, 57-138. https://doi.org/10.1515/9783110684384-003.

Haverling, G.V.M. (2020). "2.1 Background and early developments". Roelli, P. (ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches*. Berlin: De Gruyter, 59-80. https://doi.org/10.1515/9783110684384-003.

Hoenen, A. (2020). "8 Evolutionary models in other disciplines". Roelli, P. (ed.), *Handbook of Stemmatology: History, Methodology, Digital Approaches*. Berlin: De Gruyter, 534-86. https://doi.org/10.1515/9783110684384-009.

Jerome (ca. 384). *Praefatio Hieronymi in Quatuor Evangelia* [Latin text, Migne PL 29, col. 525-528]. Early Church Texts. https://earlychurchtexts.com/main/jerome/preface_to_four_gospels.shtml.

Keyser, P.T. (1993). "The Purpose of the Parthian Galvanic Cells: A First-Century A.D. Electric Battery Used for Analgesia". Journal of Near Eastern Studies, 52(2), 81-98.

Migne, J.-P. (ed.) (1844-65/1864-65). *Patrologiae Cursus Completus: Series Latina*. 221 vols. Paris: Migne; indices 1862-65. Repr.: Turnhout: Brepols, 1982-93.

Migne, J.-P. (ed.) (1857-66). *Patrologiae Cursus Completus: Series Graeca*. 161 vols. Paris: Migne. Repr.: Athens: Centre for Patristic Publications, 1997-98.

Nagy, G. (2004). *Homer's Text and Language*. Urbana: University of Illinois Press.

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), 81-8.

Pettegree, A. (2010). *The Book in the Renaissance*. New Haven: Yale University Press.

Rader, Ma. (ed., trans.) (1615). *Chronicon Alexandrinum idemque astronomicum et ecclesiasticum (vulgò Siculum seu Fasti Siculi)*. Monachii: Ex formis Annae Bergiae viduae.

Roby, C.A. (2023). *The Mechanical Tradition of Hero of Alexandria*. Cambridge: Cambridge University Press.

Tov, E. (2018). *The Essence and History of the Masoretic Text* (lecture paper). Jerusalem: Hebrew University of Jerusalem, 6-8.

Trovato, P. (2017). *Everything You Always Wanted to Know about Lachmann's Method*. Limena (PD): Libreriauniversitaria edizioni.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. (2022). "Chain-of-thought Prompting Elicits Reasoning in Large Language Models". *Advances in Neural Information Processing Systems*, 35, 24824-37.