2021

An open collaborative system for Digital Stemmatology

Laura Mancini
University of Siena, l.mancini19@student.unisi.it

An open collaborative system for Digital Stemmatology

Laura Mancini[1]

[1] PhD Candidate, University of Siena, Italy
l.mancini19@student.unisi.it

Abstract. The emerging of digital technologies in the past century led to the creation of tools designed to facilitate the approach to existing subjects adapting the new technologies to older methods. In this paper we suggest taking a step forward, applying a modern technology characteristic to Stemmatology: the study of the genealogical relationships between surviving manuscripts presenting the same text. We propose an on open collaborative system, an online platform on which it is possible to share, comment, and edit *stemmata codicum*, based on the design of the latest online platforms. Furthermore, this system will be equipped with an AI tool, designed to help the users to create their *stemmata* and based on an algorithm built for this purpose and based on the latest findings.

Keywords: Stemmatology, Artificial Intelligence, Philology, Web Development, Digital Platform, Digital Humanities

1 Introduction

In order to understand the features of the system we are proposing, we will need to briefly present Stemmatology and its principles. After that, we will explain which notions of Computational Mathematics and Phylogenetics are being applied to this field, leading to the study of the cutting-edge systems and algorithms employed in digital Stemmatology to create a graph from the text tradition and to analyse it. Given these notions, we will present our proposal in detail, starting from its architecture, the AI tool and the algorithm and arriving at its user interface.

2 Stemmatology

When we think about a modern book, we hardly consider it as a collation of multiple elements or a variation of an original work. It can be seen that the invention of printing in the 15th Century represented the beginning of the way we now consider books, but this does not apply to the older texts. Approaching a classic or medieval text, we should bear in mind that the text that we are reading might not be the original work: years of copying changed the text in many ways (by mistake, adapting it to a dialect or to a more modern stage of the language) and as philologists it is our job to recon-

struct the story of the original text. This 'original' text may not be the one chosen for the printed edition: as a matter of fact, the invention of printing made the reconstruction process harder for us. When XV's century editors wanted to create a print edition of an older text, most of the time they took the edition available in their area, or the most common edition at the time, without a proper philological work. This happened for centuries, compelling us to deal with famous printed versions that are much different from the original. Today we are aware of multiple examples showing how this mechanism led to errors, the most prominent one could be the printing of the Bible, still corrected and amended nowadays. These things do not happen anymore, at least since the beginning of Philology, a science that studies ancient texts and how they were transmitted to us, tracing their path, trying to get as close as possible to the old, original text. It is now important to describe this process and how it is possible to construct the genealogical tree of a text, from the initial stages to the most advanced ones. Starting from the very beginning, we notice that for the same text there are different manuscripts, each one different from one another, even slightly, and therefore we should attempt to understand the historical, cultural, dialectal, or diachronic reasons that led to the differences in order to reconstruct the original text. Together with the differences, we need to find the common elements that make a text similar to another one, hence we are able to form a sort of scheme indicating the "genealogical relationship between the surviving manuscripts" [19, p. 62]. At the end of this study, we should be able to reconstruct a critical edition and to do so, we will rely on the teachings of one of the most known personalities of the Nineteenth century: Karl Lachmann (1793-1851), a German Latinist and philologist that created the method still used nowadays, despite minor changes and adjusting.

First, we need to acknowledge all the testimonies of our chosen text, accounting also the new findings. This phase is called recognition and census. After this primary stage, we are going to move to the *recensio*, which is the step where we will determine the authenticity of our manuscripts. This requires notions of Paleography, History, Linguistics, Literature. After having examined the single manuscript, we need to confront it with the others of the tradition it belongs to, operating what it is called a *collation*. After these two stages are accomplished, we should be able to determine which manuscript is the oldest one and in which parts the manuscripts report a different text (called *varia lectio*).

The differences we find are, by definition, *errors*, since they represent an alternative form from the original and correct text. This is the reason why we call them guide-errors: two manuscripts presenting the same *lectio* will have a conjunctive error while two manuscripts presenting different lectio will have one or more separative errors. To put it in another way, the differences between texts are called *lectio* and we are able to individuate the differences between witnesses, bearing in mind that every apograph copy maintains the errors of its antigraph. Of course, it has to be taken into account that not all the errors could serve as a tool to distinguish between different

lectio: any copyist could have made some errors by distraction (and these kinds of errors are called polygenetic). What we obtain at the end is called archetype, despite not being able to say that it is the original text, it represents the closest reconstruction we could obtain with the surviving witnesses [19, p. 71]; another definition offered by Roelli describes the archetype as "the most recent witness from which all extant witnesses of a text derive" [12, p. 221]. In the end, the archetype is collocated right below the original in the *stemma codicum*: the final diagram indicating the genealogical relationship between surviving witnesses. *Stemma* is a Latin word indicating the genealogical tree made by the relationships between manuscripts and the first one has been designed by Bengel in the 18th Century [12, p. 211].

It could be considered pointless to underline how difficult it is to construct a stemma with a *codex unicus*, which is the only testimony of a text tradition. In this case, since we could not have diachronic examples of the same text, or texts with dialectal differences, we would only try to study the language and the history of that text, making hypotheses on the middle steps that separate the original copy from the one we are able to read. There is another case in which this method is nearly useless, and this happens when a copyist used more than a manuscript to create a copy, in this case, called *contaminatio*, we do not have solely a vertical tree anymore, but also some horizontal lines as well.

We are now going to describe the aspect of a stemma, defining the noting method we choose to use in this paper and in the online system as well. It has been already said that at the top of the tree we can find the original, while the archetype is right below it. There could be even an hyperarchetype beneath it, representing texts often lost that are the ancestor of families of similar witnesses [12, p. 221], a sort of subcategorization, and it is noted by a Greek letter [5, p. 32]. Finally, we find the single texts, noted by a capital letter, representing the surviving witnesses we used to create the stemma.

The last stage of this editorial process is the creation of an edition and its definition varies according to the differences between traditions. Following Odd Einar Haugen's analysis [12, pp. 59-64], we can distinguish between a monotypic edition, in which the editor focuses on the text of a single testimony; a diplomatic edition, which is an example of a regularized orthography (partially or completely); the synoptic edition, which is divided into two typologies: the first one is the juxtaposition of different versions of the same work, while the second type is represented by the juxtaposition of witnesses to the same work; the eclectic edition is based on multiple witnesses and, finally, the critical edition is an "edition based on a genealogical recension of the manuscripts and using the information in the establishment of the text" [12, p. 361].

3 Computation Mathematics and Phylogenetics

In order to understand the second and most important part of this paper, we need to explain how Computational mathematics and Information Technologies can be applied within this field and how an algorithm could be created following the Stemmatics steps.

First, it is important to underline that the *stemma codicum* can be described with graph theory. Using its definition, we can assume "a graph is a pair G = (V, E) of sets satisfying E $[V]^2$; thus, the elements of E are element subsets of V" [6, p. 2]. This means that we have a graph, which in our case is the stemma, composed by two kinds of elements: V represents the nodes or points, the former denomination will be preferred in this paper and in the stemma these elements are the manuscripts, the archetype and the hyperarchetype; on the other side E represents the edges or lines between the points, and in our case, these are the branches of the genealogical tree. Furthermore, a graph with the vertex set on a specific element, let us assume it is V, is called a graph on V, consequently in the stemma it will be on the original text (if it is being considered) or on the archetype. A model representing this is the DAG (directed acyclic graph) and it is designed in the image below:

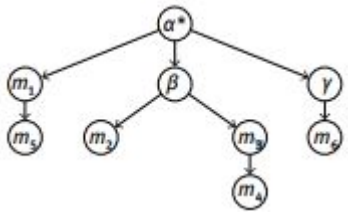

Fig. 1. DAG

Thanks to this model it is possible to introduce another concept about graphs: they can be directed or undirected. A directed graph, as the one we need to study "is a pair (V,E) of disjoint sets (vertices and edges) together with two maps init: E → V and ter: E → V assigning to every edge e an initial vertex init(e) and a terminal vertex ter(e). The edge is said to be directed from init(e) to ter(e)" [6, p. 25]; this means that in a stemma there is a clear orientation from the older testimonies to the more recent ones and also that the edges between the nodes have only one direction.

We need to introduce another concept necessary to understand the following part of this paper: the Phylogenetics. We can define it as "the history of evolution of a species or group, especially in reference to lines of descent and relationships among broad groups of organisms" [9] and it has been used as a useful comparison to Stemmatology, thanks to their similarities: the archetype here is compared to the MRCA,

the most recent common ancestor and many systems of Bioinformatics are now adopted in Computational Stemmatology.

4 Related works

In the past century, all these notions have been used together to construct digital tools that could help philologists in their work. Being aware of the technological limits, Computational Stemmatology and Bioinformatics created algorithms that produced graphs, rather than stemmata, that needed a further study and implementation by a human expert to be regarded as stemmata.

Although scholars began applying Computational methods to Stemmatology in the 1950s, it was only in the 90s that Bioinformatics and Phylogenetics were applied to these studies. Hoenen describes how the textual transmission is conceived as a DNA string, a linear sequence, and how as a string this can also be studied in phylogenetics software [12, p. 298].

Even if nowadays Phylogenetics algorithms are the most used tools, there has been a recent development in these studies: on the one hand Roos, Heikkila ad Myllymaky (2006) invented an algorithm that is not a direct loan from Bioinformatics, on the other the first datasets have been produced in order to study the differences between methods and algorithms. The largest study in digital traditions was made by Roos and Heikkila in 2009 [12, p. 300] and it is still the most complete dissertation upon this topic.

Furthermore, in the past decade, an attempt to create a wider digital platform has been made by a creative research project (CREA), funded by KU Leuven, and it is called Stemmaweb (https://stemmaweb.net/). It is a tool for collation analysis, its algorithm assumes the collation is a graph and that is the way it creates a variant graph. For the first time, it was possible to upload a text (or, better, its tradition) in a TEI encoding format and the algorithm would give back a stemma, with the possibility to examine variants against that stemma, view collation and relationships and to download the tradition. It is also possible to edit the analysis options, such as ignoring orthographic and spelling variations (since it is clear to us that the only important errors in this stage are not polygenetic). Another interesting feature allows us to explore the collation and the resulting graph, visualizing the relations between words and testimonies. Additionally, a colour legend is supposed to distinguish between orthographic, punctuation, spelling, grammatical, repetition or transposition differences between the multiple *lectio*. Unfortunately, Stemmaweb, created in 2010-12 and not implemented since, is now outdated and poorly functioning, as its creators state as well (https://stemmaweb.net/?p=74).

Another step forward has been made by a team of software engineering students at the University of Bern and the Digital Humanities group at the University of Vienna: StemmaRest, a repository for editions. It has been conceived as a backend for Stem-

maweb and its aim is to demonstrate that a scalable database is more effective than the previous relational attempt, since it allows to manipulate larger sets of data, accessing and transforming them according to the algorithm needs [3]. It is in an embryonal stage, but it has already been presented at the European Society for Textual Scholarship conference, and it is expected to have a better algorithm and more functionalities that consider the critical edition of a text more than the manuscript text. Both projects are available on GitHub: a collaborative code repository that allows a code to be open and readable to whom it may be interesting for (Stemmaweb at: https://github.com/tla/stemmatology/; StemmaRest at: https://github.com/DHUniWien/tradition_repo), while an extensive description of the entire process starting from the digitalization of a text and leading to its analysis in StemmaWeb is provided by Seretan [15].

5 Our Proposal

After having presented the systems that have been used until this moment, we are now going to present our proposal for a new supportive system for scholars, researchers, and students, bearing in mind the structure of the Internet of nowadays: this means that the system we have in mind is cooperative and collaborative, a platform where it is simple to be assessed by other experts or where a student can make some tests.

Moreover, we must consider that the Stemmatology is incredibly fragile: a new manuscript can be found at any time, destroying, and making out of date all the hypotheses made until that moment. For example, in Italy some 1350's scrolls of the Divine Comedy have been found in 2021 in the Collegio Ghislieri, and the text contained in them is considered the oldest *lectio* now available.

What if there was a repository and collaborative system in which also the students could make hypotheses and scholars could visualize all the up-to-date stemmata published? This is the kind of system we are building, and we are going to explain it in detail. In this paper we will refer to the system as StemmaTree, even if the name and the domain are still being discussed.

Since we have in mind an open, digital, and collaborative system, we need to build our website thinking about who will use it. As a matter of fact, being designed as a scholar's tool, it will need to be well implemented from a usability point of view. For this reason, its architecture has to be well thought and designed accordingly, making the users' experience as easy as possible, keeping the contents organized and meeting the requirements of the tool we are building. The scheme we have in mind is thought to present a structure along the lines of this one:

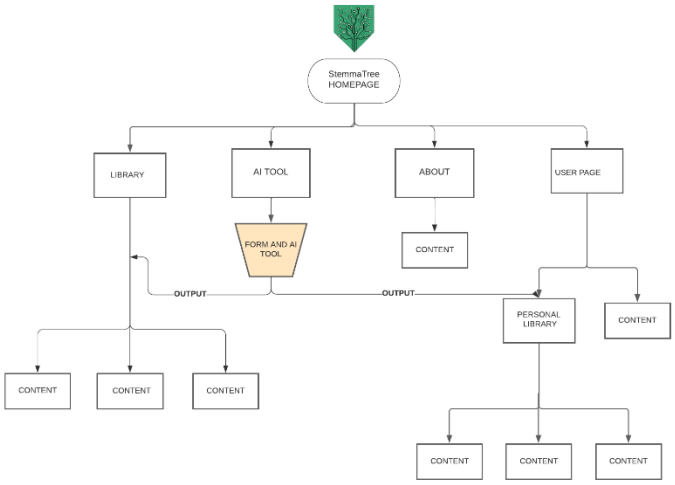

Fig. 2. Architecture

From this scheme we are able to distinguish three elements: the categories that contain other elements, such as the two libraries, the About page and the User page, which means to represent a personal space for whoever will use our AI; the AI tool which will be based on a form and will be described in the final paragraph and, finally, the contents, that could be the about pages or the output stemmata. It can be seen how the output of the process that will involve the mechanism of the artificial intelligence will be stored in both the libraries. In fact, the user will have a personal library as well, in which there will be stored the stemmata they created. In addition to this, there will be hierarchy in the typologies of users and it will come as a distinction between the published stemmata and the ones intended as experiments. As stated before, this structure is meant to help whoever uses it to do it in the easiest way possible, but also to browse it without getting lost in the process.

5.1 Graphical User Interface (GUI)

Designing our GUI, we will need to consider the facility to navigate the website on behalf of the users. This can be achieved providing the right design of the features we implemented, anticipating the users' behaviours and difficulties. In this paragraph we will describe in detail the features and we will present a first prototype of their design

on the website. Starting from the landing page, we decided to create it as a modern homepage, making our tool to look like a product; as such we could describe the behaviour of the users as potential clients, providing clear explanations of how the tool works and how it will help them in their studies. Given that the platform is thought of as a collaborative system, a high engagement would be preferred, so that the users will confront and assess more stemmata.

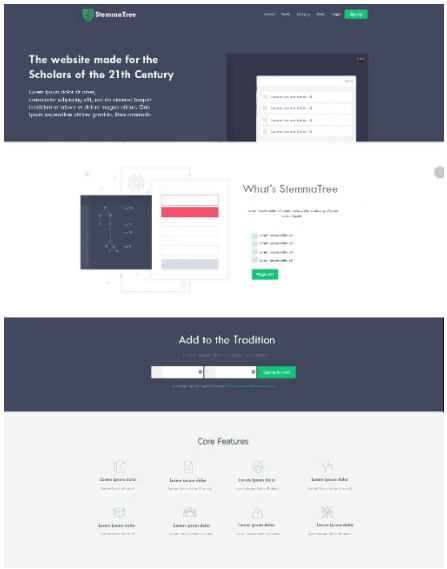

Fig. 3. Landing Page

As pointed out in the image, the landing page will present basic information, a section in which it will be possible to do a quick research into a text tradition, to add a personal contribution to it and even a section describing the core features of the website. From the navigation bar it will be possible to reach every section of the website, which will be now presented into detail.

First, we need to introduce the AI (Artificial Intelligence) tool. In the first part of this paper, we examined how the Computational Mathematics is being applied to Stemmatology and now we are designing one of these algorithms, it will be based on a form that will help the users to upload a personal stemma or to have a digital tool helping them in their studies. In the following paragraph we are going to explain in detail how it will function.

Moreover, how the stemmata will be stored and organized for the users is a much more important aspect to explain: they will be available in a Library, which is described here:

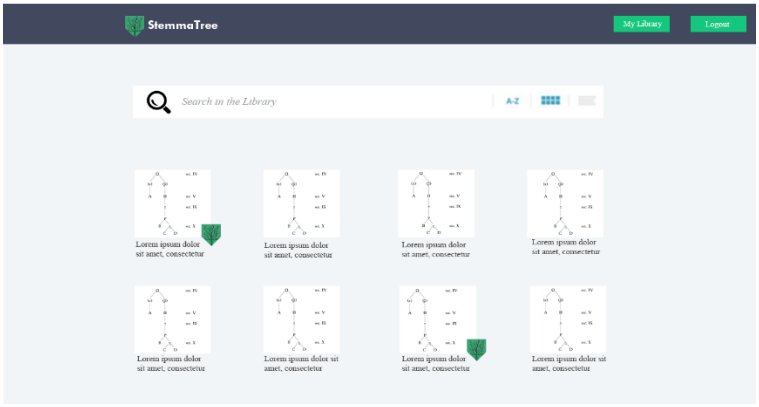

Fig. 4. Library

The Library will be organized taking genres, languages, and traditions into account. It will be possible to do a quick research in the research bar, using as key words either the authors of the critical edition or the authors of the original text. More important to us is how to distinguish between a stemma formulated by a scholar who published a paper or a critical edition, and a stemma made as an exercise from either a scholar or a student. The 'published' stemmata will be noted with a special sign, as it is shown in the picture above. Clicking on a stemma, it will be possible to explore its page and open a comments and notes area, where users will be able to assess other stemmata, leaving a feedback, or replying with their version of the same stemma. This mechanism would not be possible if we would have not provided a hierarchy between users, therefore creating multiple typologies of accounts: Scholar, Student and Researcher. It will be possible for Students to assess other stemmata, in a sort of peer assessment solution; moreover, they will visualize scholars and researcher's stemmata as well. Scholars will be intended as those who already published a critical edition, or an academic paper and their stemmata will be regarded as the official ones in the Library. A Scholar will be able to edit his stemmata, creating alternatives and it will be possible to do it for studying purposes rather than publishing ones, as it is clear to us that it is important to protect ideas and intuitions before they are officially published.

The last typology of account will be the Researcher and it is intended to be an intermediate stage between the other two. Researchers will be able to create their stemmata, assess those of students; their stemmata could also be assessed by other peers and Scholars as well. If they publish a critical edition, their account will be automatically upgraded to a Scholar type.

5.1 Artificial Intelligence (AI)

From the previous sections, it should be clear that the process of building a Stemma from a collection of documents can roughly divided into two main phases:

1. For each couple of documents in the collection, we compare them, trying to find their relationship: we can compute a similarity between the documents (undirected link) or a father/son relation (directed link). In both the cases, there is a value computed (how similar the documents are or how likely one is the father of the other) associated to the link.
2. Given all the (weighted) links in input, the goal is to design the stemma by dropping or adding edges where needed.

The first phase can be easily automated using similarity metrics and meta-information (where available). The weight on each edge is defined by a proper similarity function that, given two documents (i.e., the two nodes), return the similarity between the two documents. We plan to let users choose between several classical similarity functions, such as the Jaccard function; see the text Modern Information Retrieval [4] for more details.

The second phase can be interactive, with the system that can propose a stemma and the user that can check the links and decide, based on his knowledge, whether to keep or to change some of them. We are implementing algorithms for the first phase, while, initially, the system will present all the links, each with its weight, and the user will be able to design the stemma, i.e., no automated stemma design. Our goal is to collect a dataset of collections of documents and related stemmas, and then use this dataset to train a neural network to design the stemmas by itself. The user will always be able to edit the resulting stemma and new data (collections of documents and stemmas) will be used to train again the neural network, hopefully improving it. From the user experience perspective, we designed it as a two-steps process. In the first part it will be asked to the users if they want to create a new tradition or to add their attempt to an existing one:

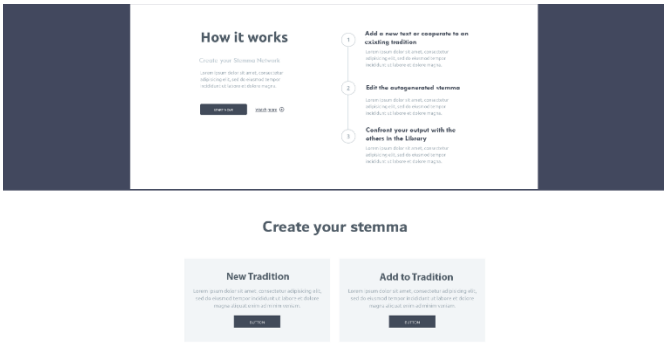

Fig. 5. Algorithm, first step

Starting a new tradition, for example, the user will need to upload all the texts in a TEI (Textual Encrypted Initiative) format and answer several questions about them in the form we created. Once we have decided which one is the best algorithm for our goal, it will be possible to design the right questions to ask, considering all the principles of AI, but also the rules of Stemmatology. After the users have answered the questions and uploaded their texts, the tool will generate an initial output that can be edited and noted further, eventually generating a stemma that could be published in the Library or not, according to the users' intention.

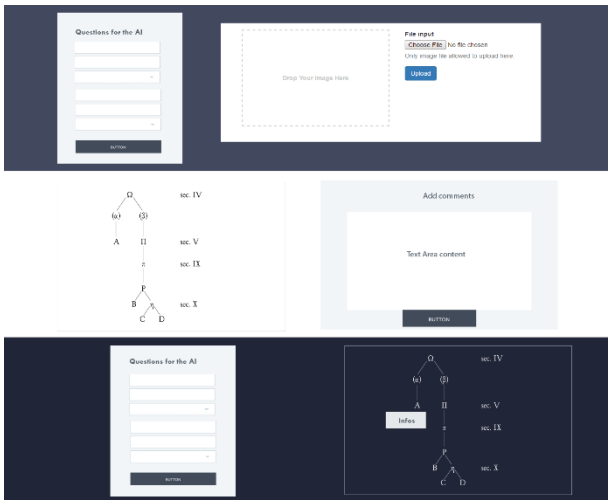

Fig. 6. Algorithm, second step

Finally, this mechanism is studied to avoid the corruption of the dataset from the errors made by users. First of all, creating it, we will only take into consideration the stammata made by 'Scholars' and not by 'Students'. Also, the stemmata published in the public library are subjected to peer reviewing, so that a stemma with mistakes will not be studied by other 'Students' or taken into consideration for the algorithm.

Furthermore, for what concerns the errors and the ones for malicious intent, being the weight of each edge defined by a similarity function, it is simple to consider a certain similarity value, under which it will result in an alarm.

6 Conclusions

To sum up, in this paper we presented our proposal for a new digital platform, applying modern structures such as being open, up to date and collaborative with another science, demonstrating how both scholars and students would benefit from this tool and how it would create a new space of confrontation for scholars. We aim at designing a new algorithm built specifically on the necessities of the AI tool provided on the digital platform. We hope it will be possible to continue this project considering that it would represent a step forward in both fields of IT and Stemmatology.

References

1. M. Agosti and F. Tomasi. Collaborative research practices and shared infrastructures for humanities computing, Cluep, 2014.
2. G. Alvoni. Scienze dell'antichità per via informatica: banche dati, Internet e risorse elettroniche nello studio dell'antichità classica. Clueb, 2002.
3. T. Andrews, I. Gershoni, R. Imhof, S. Kaufmann, J. Schaerer, T. Studer, and S. Zumbrunn. Efficient stemmatology: a graph database application in the digital humanities.
4. R. Baeza-Yates, B. Ribeiro-Neto, et al. Modern information retrieval, volume 463. ACM press New York, 1999.
5. G. Contini. Breviario di ecdotica, Einaudi, 1986.
6. R. Diestel. Graph theory. Springer-Verlag, New York, 2010.
7. M. R. Digilio. Elementi di critica testuale, in M. Battaglia (ed.) Medioevo volgare germanico, pp. 1–251, 2016.
8. D. Fiormonte. Scrittura e filologia nell'era digitale. Bollati Boringhieri, Turin, 2003.
9. L. Gittleman. John. "Phylogeny", 2016. online, accessed 14/05/2021.
10. L. Laura. Breve e universale storia degli algoritmi. Luiss University Press, 2019.
11. G. Milanese. Filologia, letteratura, computer. Idee e strumenti per l'informatica umanistica, Vita e pensiero, 2020.
12. P. Roelli. Handbook of Stemmatology: History, Methodology, Digital Approaches. De Gruyter, 2020.

13. T. Roos and T. Heikkila. Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets, in Literary and Linguistic Computing, 24(4) (2009): pp. 417–433
14. S. J. Russell and P. Norvig. Artificial intelligence-a modern approach, Series in Artificial Intelligence, Englewood Cliffs, NJ. The Knowledge Engineering Review 11, no. 1 (1996): pp. 78-79
15. V. Seretan. Digital critical edition of apocryphal literature: Sharing the pipeline. In Sharing the Experience: Workflows for the Digital Humanities. Proceedings of the DARIAH-CH Workshop 2019 (Neuchatel). Elodie Paupe; Simon Gabay; Sara Schulthess, 2020.
16. M. Szurawitzki. Digital germanic philology? questions, challenges and obstacles for scholars of German, 2009.
17. S. Timpanaro. La genesi del metodo del Lachmann, UTET, 2010.
18. F. Tomasi. Metodologie informatiche e discipline umanistiche. Carocci, 2009.
19. P. Trovato. Everything You Always Wanted to Know about Lachmann's Method. Libreria universitaria.it edizioni. 2014.
20. P. Van Reenen, A. den Hollander, and M. van Mulken. Studies in stemmatology II. John Benjamins Publishing, 2004.