Lost Manuscripts and Extinct Texts: A Dynamic Model of Cultural Transmission

Jean-Baptiste Camps[1,2,*,†], Julien Randon-Furling[3,4,*,†]

[1]Venice Center for Digital and Public Humanities, Univ. Ca' Foscari, Fondamenta Malcanton 5449, Venezia, 30123, Italy
[2]Centre Jean-Mabillon, École nationale des chartes, Paris Sciences & Lettres, 65 rue de Richelieu, Paris, 75002, France
[3]Centre Borelli, Univ. Paris-Saclay, ENS Paris-Saclay, CNRS, SSA, INSERM, 91190, Gif-sur-Yvette, France
[4]SAMM, FP2M (FR2036), Université Paris-1 Panthéon-Sorbonne, CNRS, Paris, 75013, France

Abstract

How did written works evolve, disappear or survive down through the ages? In this paper, we propose a unified, formal framework for two fundamental questions in the study of the transmission of texts: how much was lost or preserved from all works of the past, and why do their genealogies (their "phylogenetic trees") present the very peculiar shapes that we observe or, more precisely, reconstruct? We argue here that these questions share similarities to those encountered in evolutionary biology, and can be described in terms of "genetic" drift and "natural" selection. Through agent-based models, we show that such properties as have been observed by philologists since the 1800s can be simulated, and confronted to data gathered for ancient and medieval texts across Europe, in order to obtain plausible estimations of the number of works and manuscripts that existed and were lost.

Keywords

agent-based models, stochastic models, loss of cultural artefacts, text transmission, stemmatology

1. Introduction

How much do we preserve of the written knowledge, science [18] and culture of the past? And how representative is what we know compared to what existed? Such fundamental questions depend on the process through which texts were distributed materially.

Before the advent of the printing press, written texts were circulated in manuscript form. In order to make the text available, the author would dictate it to a secretary, or write a draft on wax tablets, papyrus, parchment or, eventually, paper, and this original, authorial manuscript would then have to be copied manually by a scribe in the form of a new manuscript, and then circulated. Copies could then be used to create more manuscripts, again by manual copying, perhaps by other scribes in other regions at a later date. During this process, successive modifications were introduced in the text, either by error or intentionally, to make the text more

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium
*Corresponding author.
†These authors contributed equally.
jean-baptiste.camps@chartes.psl.eu (J. Camps); julien.randon-furling@cantab.net (J. Randon-Furling)
https://github.com/Jean-Baptiste-Camps/ (J. Camps)
0000-0003-0385-7037 (J. Camps); 0000-0001-9497-2297 (J. Randon-Furling)

suited to its intended audience. These alterations in the written sequence that forms the text could then be transmitted by a given manuscript to its "descendants". Wear and tear, accidents, fashions caused the destruction of some manuscripts, while others enjoyed the long life of library preservation. In the end, knowledge was lost and some texts went extinct, while other texts gained traction or were eventually preserved for future generations.

In textual studies and philology, since the development of the "common errors" methods from the 19th century onward [48, 24], the analysis of text alterations allows philologists to reconstruct relations between the surviving copies of a given work (so-called witnesses), and to represent them as a tree-like graph, called a *stemma codicum* (fig. 1). This reconstructive process is not entirely different from the methods used by biologists to reconstruct the links between existing or extinct species, based on their shared characteristics that are supposed to derive from a common ancestry, and to represent it as a phylogenetic tree. Methods from this biological subfield, known as cladistics, have even sometimes been directly applied to texts, with controversial results [3, 31, 47, 36, 26]. These trees represent the result of an enquiry into the relationships between the surviving witnesses, together with the hypothetical lost nodes that can be deduced from them. To arrive at it, researchers have to study the variants — more precisely, common innovations or errors (i.e., mutations) — observed in the text of the surviving witnesses (fig. 1, A). The reconstructed tree will show only what can be deduced from surviving witnesses: the witnesses themselves, and as many hypothetical nodes as are needed to explain their relationships (fig. 1, B). While it may be the case with more recent works that all nodes are known and the graph represents the full transmission of the text — for instance with the genealogy of printed editions (fig. 1, C) —, most of the time the tree represents only a (potentially very small) subset of what existed (fig. 1, D).

A long standing observation, not yet fully understood, about the structure of the reconstructed trees was made by the French philologist Joseph Bédier, almost a century ago, in 1928 [6]. He observed that most trees reconstructed by philologists show a root bifurcation (a root with outdegree 2): in most reconstructions, the original (or the lowest common ancestor, called the *archetype*) had two, and only two, direct descendants, preventing an accurate reconstruction of the original text by majority principle (e.g., two witnesses vs one).[1] Instead of searching an explanation for such data in the dynamics of text transmission, he interpreted this "forest of bifid trees" as the result of a methodological flaw or an unconscious bias. Indeed, the main goal of establishing textual genealogies was, at that time, perceived to be the "mechanical" reconstruction of the original text (or, at least, of the archetype), and the elimination of the personal judgement of the philologist in the choice of the original variant in the places were several readings were in competition. But, for this to be possible at the top of the genealogy, it is necessary to be able to proceed by majority principle (e.g., for three direct descendants,

[1] It is to be noted that this bias for bifurcation in stemmata is different from the systematic bifurcations that appear in many linguistic trees or biological phylogenies, in which the process of divergence of two lineages at a given point in time is represented as a bifurcation. For true multifurcations (or *pitchforks*) to happen, it would need the simultaneous divergence of three lineages. These true multifurcations, called *hard polytomies*, are difficult to spot or verify, and subsequently rare in phylogenetic trees. Moreover, they can only appear in very specific evolutionary contexts, notably in cases of rapid simultaneous radiations of lineages: for instance, tectonic uplift may have abruptly fragmented the habitat of the lizard genus *Liolaemus* between simultaneously isolated populations, and, in combination with rapid climate change in temperature, caused the rapid divergence of several lineages of the cold blooded and temperature-sensitive lizards [39].

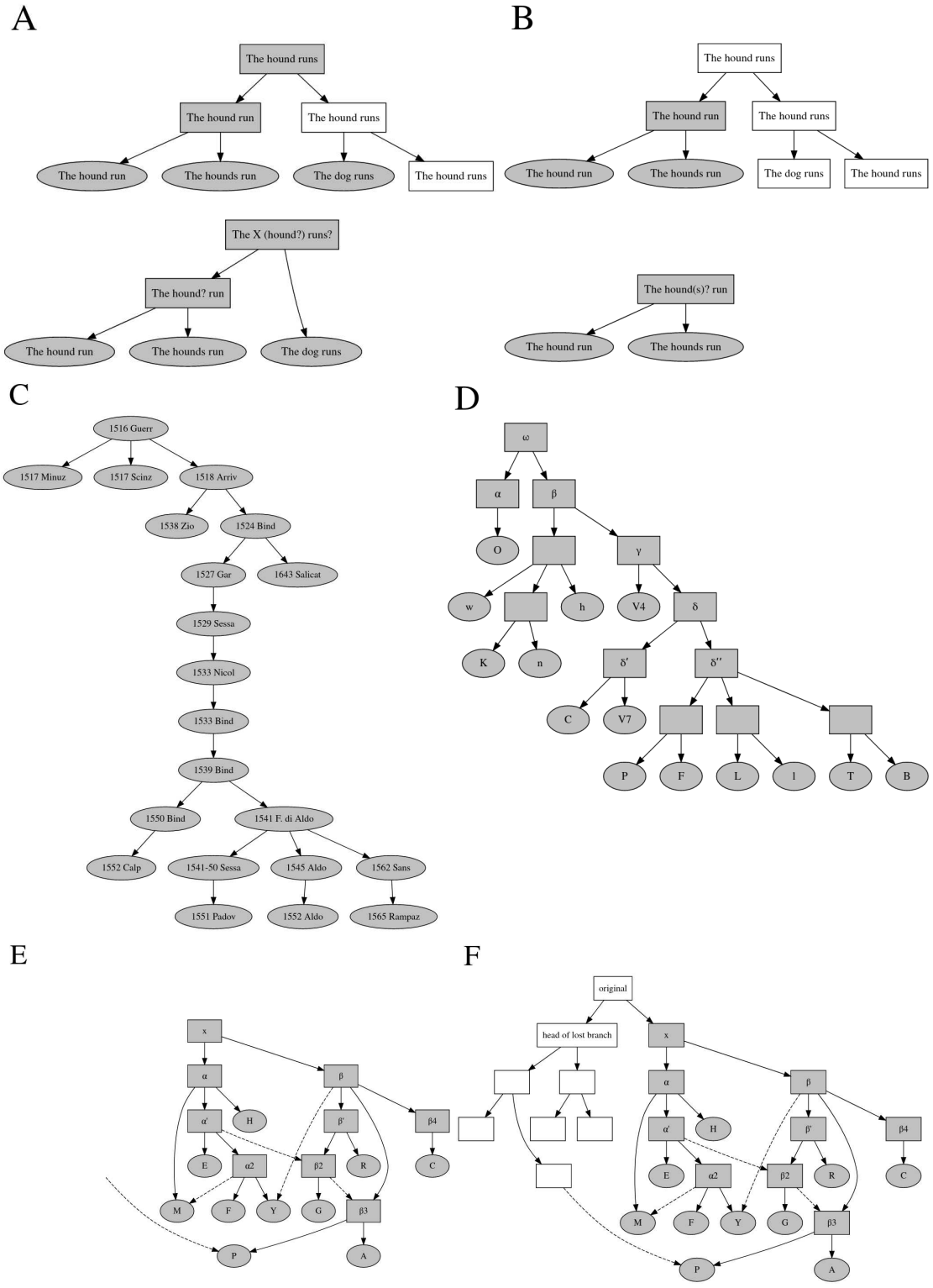

Figure 1: From the real tree to the reconstruction. **A** an artificial example of the transmission of a very short text and its plausible reconstruction, where circles depict extant manuscripts, rectangles lost ones), and gray, nodes that are recoverable when reconstructing the tree; **B** another artificial example, where the distribution of lost manuscripts causes the loss of a whole branch, and where, consequently, the last common ancestor is not the original (root), but a lower witness (archetype); **C** the observed phylogeny of the successive editions of Fortunio, *Regole grammaticali della volgar lingua* (1516) [44, 23]; **D** the reconstructed phylogeny (*stemma*) of the Old French *Song of Roland* [45]; **E** the reconstructed phylogeny of the Anglo-French *Guy de Warewic* (Ewert, 1932), showing many cases of lateral transmission ("contamination") in dashed lines, and even one instance of lateral transmission from outside the tree; **F** the same tree, but including (box shape) the outline of a lost branch (cf. [28]).

the variant present in two of them will be selected over the one in single occurrence). Yet, Bédier observed, philologists tended to establish genealogies with only two direct descendants (or even, to revise existing genealogies with three direct descendants to reduce them to two), regaining control over the selection of reading and abolishing the 'mechanical' choice. For this reason, he advocated to stop reconstructing the original, and use only the genealogy to select the 'best' witness (the closest to the original) and transcribe it in a conservative fashion, limiting modifications or corrections. This spurred a century long debate and caused, between 'bédierists' and supporters of the genealogical method, a long lasting methodological schism that remains to be fully resolved [1, 20].

Bédier's initial observation has been replicated, with estimates of the proportion of root bifurcation varying, from Bédier's 95.5%, to somewhat lower estimates ranging from 70% to 83% [46, 16, 25]. Yet, some have argued that the prevalence of root bifurcation could be an explainable feature of manuscript transmission of texts as it reached us. Tentative explanations include combinatorial estimations of the proportion of root bifurcation for a given number of witnesses, under the assumption that all configurations are equally likely [34, 16, 29], or consider the effects of decimation (i.e., manuscript loss) [22], for instance by applying a uniform loss probability to static preexisting trees [23] or by calculating a node specific loss probability to simulated trees [27]. Even if it generated little follow-ups, there have also been rare attempts of using birth and death process for exploring the dynamics of manuscript transmission [51, 52].

The abundance of roots – and more generally nodes [25] – with out-degree 2, is not the only property that can be observed in many stemmas. The asymmetry between branches (cf. fig. 1, C and D), the presence or not of lateral transmission (generally called "contamination"; fig. 1, E) are other properties worthy of investigations, as well as those that can indicate that the tree made from extant witnesses represents only a small portion of the original tradition (i.e., lateral transmission from outside the tree; root identifiable not with the original but with a later manuscript; fig. 1, F). It is reasonable to assume that some of these properties reflect the dynamics of manuscript transmission, while others keep trace of important destruction, decimating manuscripts and removing even full branches. For the texts, this can be seen as an evolutionary process, where two antagonist tendencies are at work: the apparition of textual variants in individuals, causing the increase of diversity in the tradition, and the extinction of full branches, causing some variants to prevail upon others and so reducing diversity.

Here again, these observations can be put in perspective with problems occurring in evolutionary biology, where, too, two antagonists tendencies are at work, mutation and fixation, either by drift or natural selection, in a context where processes of speciation and extinction are strongly linked and where extant species represent only a small subset of the species that have existed [55]. In both cases, survival might be the exception and extinction the rule, be it by "bad genes or bad luck", a process that can be seen in terms of gambler's paradox [43, 14]: a gambler starts playing a game, in which, at each discrete step, he or she has a chance of loosing, let say 50%; even if the game is fair, after a sufficient number of random steps, the inevitable and only possible final outcome is ruin (extinction). If the gambler has a winning streak at the beginning, ruin might be significantly delayed, but, very often, ruin will happen immediately.

Basic processes of reproduction and destruction create complex shapes in the trees, from which one might want to deduce whether they can be fully explained by random process akin to

genetic drift, or if differences in selective values are to be suspected. In other terms, going back to textual traditions, if cultural context, through literary taste, canon or fashion for instance, creates a form of evolutionary pressure on textual traditions. Any insights gained on this question would have an applicability beyond the question of the transmission of antique and medieval texts, because it seems that similar dynamics are at work in the diffusion of content in other medium, including print [23] or even the web [2], and have been observed in areas such as the cognitive evolution of scientific fields and the dynamics of scientific memes [7, 17].

Data on cases as different as the songs of the Medieval Occitan troubadours from southern France or the incunabula editions printed in Renaissance Italy outline the same Pareto-like world, where a large number of texts are kept only in a single or handful of documents, while a limited number of "successful" texts are kept in a large number of copies (fig. 2, **A** and **B**), where most authors are known only for one or two texts, while a very limited number of writers can have dozens of texts preserved (fig. 2, **C** and **D**)... Such a process is also apparent in the constitution of a literary canon of a limited number of authors and texts. This 'canonization' can be seen has a progressive loss of diversity, where an ever shrinking number of authors and texts take on an ever growing share of the circulated documents (fig. 2, **E** and **F**). But is this due to chance or to a selective process?

In fact, some properties as were just mentioned for textual traditions have some pendants in evolutionary models, concerning for instance the very unequal distribution of descendants [19], the varying patterns of biodiversity varying in time and space, studied in macroecology and biogeography [10, 42] and the dynamics of speciation and extinction that manifest themselves in the shape of phylogenies and the loss of branches from the tree of life [35, 55]. For this reason, there are inspiration and resources to be found in the study of mathematical properties of evolutionary trees, regarding the establishment of a null-model [8].

Can we gain some insight on what existed, what was lost, and the driving forces between extinction or survival of texts — by drift or selection? In order to do so, we need a better understanding of the dynamical processes of manuscript transmission.

2. An agent-based model of a stochastic process

In this paper we present a selection of the first results obtained with a stochastic model for the transmission of manuscripts in the Middle Ages. Following Weitzman [51, 52], we use so-called *birth-and-death* processes. These are random processes introduced in probability theory to describe, among other things, simple population dynamics and genealogies. For the simplest versions of theses processes, it is possible to derive analytically (i.e. with mathematical formulae) certain quantities of interest: the expected number of individuals (here for us, manuscripts) still present at a time t, the extinction probability, the survival probability,... But also, quantities of particular interest in the context of manuscript genealogies: the probability that the latest common ancestor (lca) be an archetype rather than the original, or the probability of root bifidity for the reconstructed stemma. Here we favour a numerical approach, that allows us to explore more complicated variants of a birth-and-death process through an agent-based computer simulation. The agents correspond to manuscripts, and during each time step of the simulation each agent has a probability λ of being copied (i.e., 'giving birth' to one copy) and

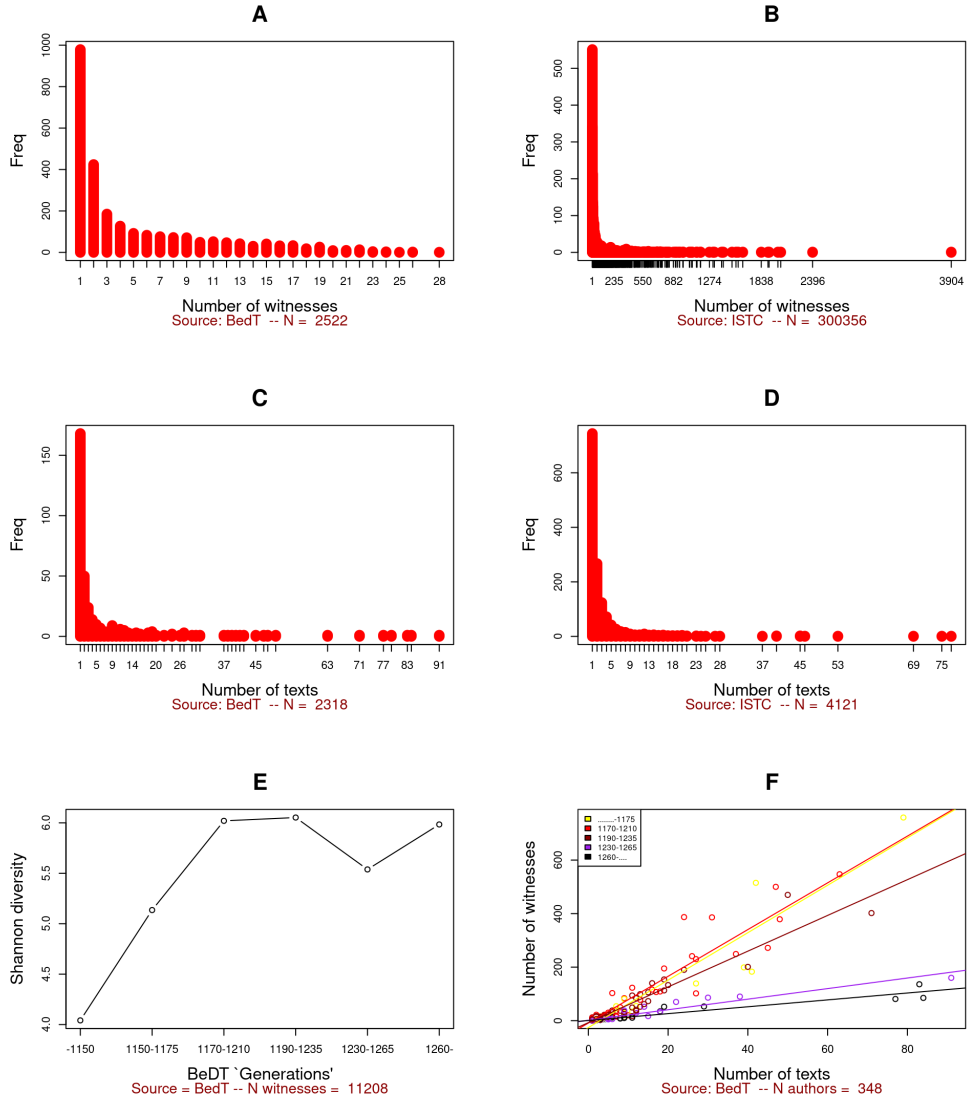

Figure 2: A Pareto-like world with diminishing diversity. **A** distribution of witnesses per troubadour text; **B** distribution of the extant copies per work for incunabula printed in Italy; **C** distribution of texts per author for troubadours and **D** incunabula; **E** Shannon diversity, Generations as sites, texts as species, witnesses as individuals; **F** number of texts and witnesses per author.

a probability μ of disappearing. Following Cisne, we limit the increase rate of the population by letting the birth rate λ depend on the size of the population at the previous step k_{t-1}:

$$\lambda_t = \lambda \frac{K}{k_{t-1}}, \tag{1}$$

where K is a theoretical upper bound on the maximum possible size of the manuscript population extant at any given time during the period under consideration.

Each simulation starts with a single agent (the original manuscript), at $t = 0$. At the first step, this agent has a probability of being copied (birth rate λ) and a probability of being destroyed (death rate μ), both between 0 and 1. If an agent is destroyed, it ceases to be able to give birth, but, as long as this hasn't happened, it can still be copied at each time step (according to λ). If all agents are destroyed, and the tradition extinct, the simulation stops. Otherwise, it keeps going for a fixed number of active steps (e.g. 1000) and, optionally, a fixed number of inactive steps (e.g. 1000), were manuscripts can no longer be copied (λ becomes 0) but can still be destroyed, reflecting the long period where, after the Middle Ages, the texts were no longer copied, but manuscripts were still subject to destruction.

In the Cisne-type model, where a population limit K is used, λ is adjusted at step t according to the total active population and the value of K (eq. 1). This limit K is the theoretical maximum number of copies of a given work that could exist simultaneously at any time step, and it reflects the maximum capacity of the book market, before being saturated by copies of a text. It is similar to the notion of the carrying capacity of an ecosystem, i.e. the maximum number of individuals from a given species that an ecosystem can support, given the availability of food, water or habitat.

Once enough simulations are run for a given set of parameter values, the resulting trees can be analysed to compute properties such as the rate of survival of traditions, the rate of survival of agents (manuscripts), the average age of surviving manuscripts, the ratio of bifidity in the genealogies (once the genealogy is simplified, e.g. by removing destroyed manuscripts without descent and more generally destroyed manuscripts that would not appear in a stemma due to reconstruction rules), the generation of the lowest common ancestor, etc. (Fig. 3).

3. Phase diagrams obtained through computer simulations

In our simulations and our choice of parameter values, we focus here on the transmission of medieval texts. We ran agent-based simulations of a Cisne-type tradition with a total time frame of 500 pseudo-years, of which 250 active (with manuscripts being copied and destroyed) and 250 inactive (with manuscripts being only destroyed). This duration is chosen to match the time between the development of medieval Western vernacular literatures in the 13th century and the Renaissance, and the Renaissance and the progressive advent of modern cultural heritage curation, from the second half of the 18th century. Each pseudo-year is equivalent to four time-steps— that is, a time-step in the simulation corresponds roughly to 3 months. This was derived using an estimate for the time taken to produce an average 200 page manuscript (though the speed of scribes is known to vary a lot, from 1 to 10 leaves a day) [41]. We chose $K = 100\ 000 = 10^5$, as an order of magnitude (rather than 10^4 or 10^6) for the total number of manuscripts in a given tradition that could be extant simultaneously at any one time. This order of magnitude is a very rough estimate based on human population and our own assumptions: the medieval population of countries such as France or Italy was in the 10^7 range; we estimate that the maximum saturation of this market for a given book would be reached if around 1% of the population were to own a copy.

As for the remaining free parameters, namely the base "birth" or copy rate, λ, and the "death" rate μ, we explored all possible pairs of values of these parameters within their range (from 0

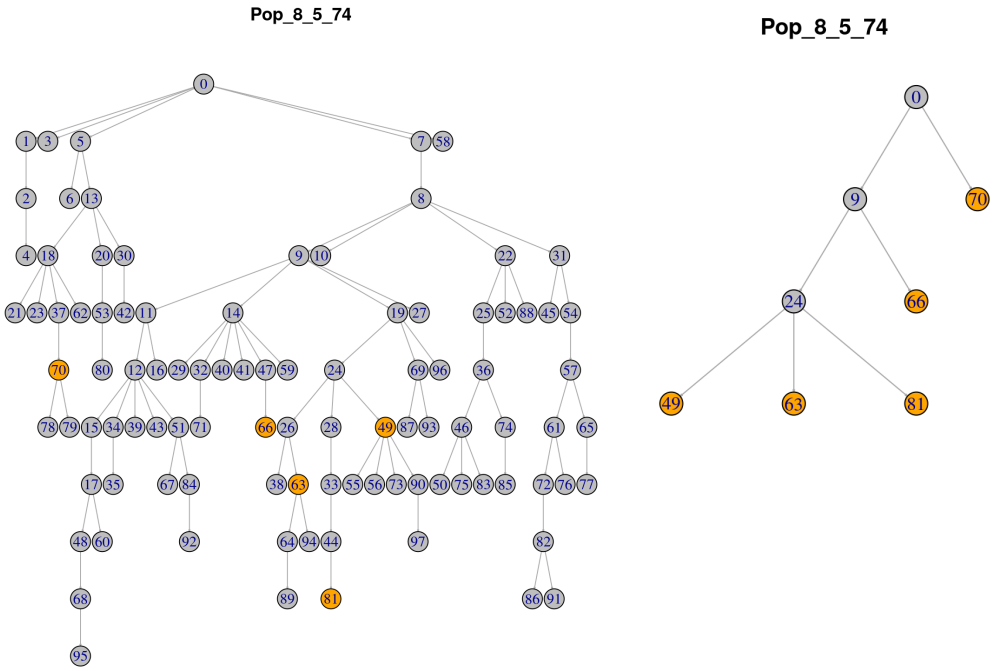

Figure 3: Final state of a simulation (one of 100, for $\lambda = 0.0008$, $\mu = 0.0005$ and $K = 10^5$), with destroyed manuscripts in grey, and surviving in orange (left); stemma-like simplification, with only the necessary nodes to express the relationships between the surviving witnesses (right). In this case, the generation of the LCA is 1 (original and LCA are the same), the root is bifid, the survival rate of manuscripts is approximately 5% ($\frac{5}{98}$). It is to be noted that manuscript 70 is the sole witness of an otherwise fully lost branch, and that all other are descended from lost manuscript 9, and even, with the exception of 66, from lost manuscript 24 (6th generation). This type of configuration is also encountered in many real traditions (see for instance above, fig. 1).

to 1 in theory, but reduced here to 10^{-4} to 10^{-3} by field expertise, i.e. rough estimates from philological knowledge).

It is to be noted, that historical and philological knowledge of loss rates is very scarce and elusive, but it can still be approached from various angles, such as the collection of data from ancient library catalogues, inventories, wills, as well as allusions and intertextuality [54, 4, 9]. Buringh [9] provides estimates for the Latin West, with a geometric mean of loss around -25% per century, with variations from -11% in the 9th to -32% in the 14th and 15th centuries (with local variations between medieval institutions from −3% to −71% per century). The global loss rate for non-illustrated manuscripts of several well known collections have been estimated around 93-97% [32, 38, 53, 40]. But there is a potential bias in accounting only for well known institutional collections, from which some manuscripts are known to have survived: trying to account for fully lost libraries, Buringh [9] is compelled to revise his estimates higher, to -25% by century until the 12th, up to -43% in the 15th. For incunabula, using editions whose original number of print made is known, it is possible to gather loss estimates by counting known sur-

viving exemplars in public or private collections: doing so for Venetian incunabula, Trovato [49] finds very variable loss rates according to textual and material typology, from 73% for the *Decretales* printed on parchment to 99.3% for more popular chivalrous literature (*Orlando furioso* for instance). This shows the importance both of variation in time and space, and of textual contents and material typology. In some extreme cases, loss can be very close to 100%, for reasons that may combine the fragility of the document form, lack of consideration for the documents or large scale historical events such as political instability, invasions or major cultural changes; examples are provided by cases as different as the Merowingian royal diplomas on papyrus or the Lombard royal charters [21], the Mayan (pre-colombian) manuscripts or medieval notarial acts [30]. Production estimates have also been attempted on the basis of the quantity of sealing wax acquired by a given producer (a chancellery for instance [5]). More founded loss estimates have also been gathered by counting how many of the acts mentioned in imperial or royal registers are kept in original or consigned in the archives of the recipients: this gives a loss rate of originals varying from 80% (acts from the emperor Charles IV in 1360-1361) to 90% for the acts from Louis X of France, increasing to 99% for the judgements rendered by his Parliament, suggesting here as well a massive effect of typological variation [30, 15], resulting in very strong biases in the body of documents available to us.

For our simulation needs, if we start from Trovato's estimates (potentially more reliable, because based on editions whose original number of copies is known), we get a survival rate whose order of magnitude is between 0.1 and 0.01 (between 10^{-2} and 10^{-3}) in 500 years (2000 steps in our model), in similar ranges as Buringh's 0.75 in 100 years and those reported by Holtz and Canteaut; from this we can deduce a step loss rate for a given total survival rate. For instance, for 1% survival rate, $(1-\mu)^{2000} = 0.01$, which simplifies to $\mu = 0.002$. So we retain values of μ between 10^{-4} to 10^{-3}. Given that books could not have been produced order of magnitudes higher or slower than they were destroyed (or we would be either drown in medieval manuscripts or keep none), we explore the same range for λ. Of course, fixed rates are a limitation, and do not yet account for substantial variations in time (such as massive extinction events, like, e.g., the fall of the Roman empire, the fire in Alexandria library, the shift from *volumen* to *codex* or from caroline to gothic script, etc.).

The space of all possible values is usually called the phase space in physics and other mathematical sciences, and a representation of the value taken by a given observable quantity (eg the extinction probability) when parameters are varied across the phase space is called the phase diagram of this quantity. In our simulations, due to limitations in computing power, we were not yet able to explore full parameter spaces for the models, and limited ourselves to these plausible values, thus not producing complete phase diagrams but instead heat maps representing a portion of the parameter space.[2]

We thus produced heat maps for a number of relevant observables, based on the variation of parameter λ and μ between 10^{-4} and 10^{-3} , with $K = 10^5$ (Fig. 4). Note that, since there is a stochastic component in the model, each heatmap is computed by averaging over the results of a relatively large number of simulations, here 100 — this means that we produced 100 artificial

[2]In the future, we plan to extend the explored parameter space for λ and μ, using exact computations whenever analytical solutions are available, and by augmenting the number of simulations, using more computing power and time. In particular, we will need to explore the impact of the variation of parameter K for which, for now, we only used a fixed initial value.

manuscript traditions for each pair $(\lambda, \mu) \in [10^{-4}, 10^{-3}]$, varying values by increments of 10^{-4}; hence each heatmap required 10 000 simulations.

The approach then consists in identifying, within the heat maps, regions in which the values for the observables are consistent either with measured quantities (as is done in the natural sciences) or with estimates for these quantities coming from other, independent and altogether different, models. We have circled in red such regions on the heatmaps in Fig. 4.

The features selected for this comparison reflect three different aspects of the traditions. The first group (fig. 4, first row) deals with the survival rate of works of traditions, that are not directly observable in historical data, but that can be compared with estimates based on secondary information or on other models [33]; the median final population of surviving traditions can, on the other hand, be directly observed by counting (known) surviving manuscripts of real-world texts. The second row deals with age properties of individual manuscripts, that, in real-world data, are sometimes known (dated manuscripts) or estimated (based on features of writing style, support, ink, language, etc.). Finally, the third row deals with the structural properties of the resulting trees, such as the distance between the original and the lowest common ancestor (archetype), a feature of much interest for existing traditions, as it cannot be directly observed, but gives an idea on how distant the text accessible to us is from the original, and how much of its history is lost). Information concerning the outdegree of the LCA matches the initial observation of Bédier on the prevalence of bifidity (root bifurcation), while the Shannon (biodiversity) index gives an insight into the asymetry of the branches, observable on real-world stemmata.

These heat maps show that the results obtained through these simulations are internally consistent in terms of not only population size and survival rates, but also in terms of structural properties of the resulting trees. In particular, the results obtained for a ratio $\frac{\mu}{\lambda}$ between $\frac{5}{8}$ and $\frac{6}{7}$ are surprisingly consistent with the observed properties of some medieval traditions, in particular those from chivalric narratives in Old French. In particular, values of 0.55 for the survival of works and 0.05 for the survival of manuscripts (fig. 4), **A** and **B**, red squared area, bottom-right tile) are identical to those provided by Kestemont et al. for Old French chivalric romances, using unrelated methods from ecodiversity [33]. Yet, for what regards specifically Old French epics, known as *chanson de geste* – a genre predating the later form of the *roman*, and whose circulation and reception considerably differs for a long time –, the median final population of 2 and the third quartile of LCA outdegree of 2 (though, median LCA Shannon index is 0.69).[3] This would lead us to revise Kestemont et al. estimates to 0.22 (instead of 0.55) for the specific survival of Old French epics (*chansons de geste*) and 0.01 (instead of 0.5) for the survival of epic manuscripts (a figure closer to that observed by Trovato [49] for their later Italian successors). On the other hand, specific values, this time, for (Arthurian and Antique matters) *romans* yield a third quartile of LCA outdegree of 3 more coherent with Kestemont et al. general estimate (median tradition size, according to Martina [37], is 2 – and mean 4.8 – for the sole *romans en vers*, similarly to *chansons de geste*, but is expected to be higher for later *romans en prose*).

[3] Data about the traditions of the *chansons de geste* follow Vitale-Brovarone's [50] and Camps' [11]. Information on the shape of stemmata have been computed based on a restriction to *chansons de geste* and deduplicating of the collection provided by OpenStemmata [13, 12].

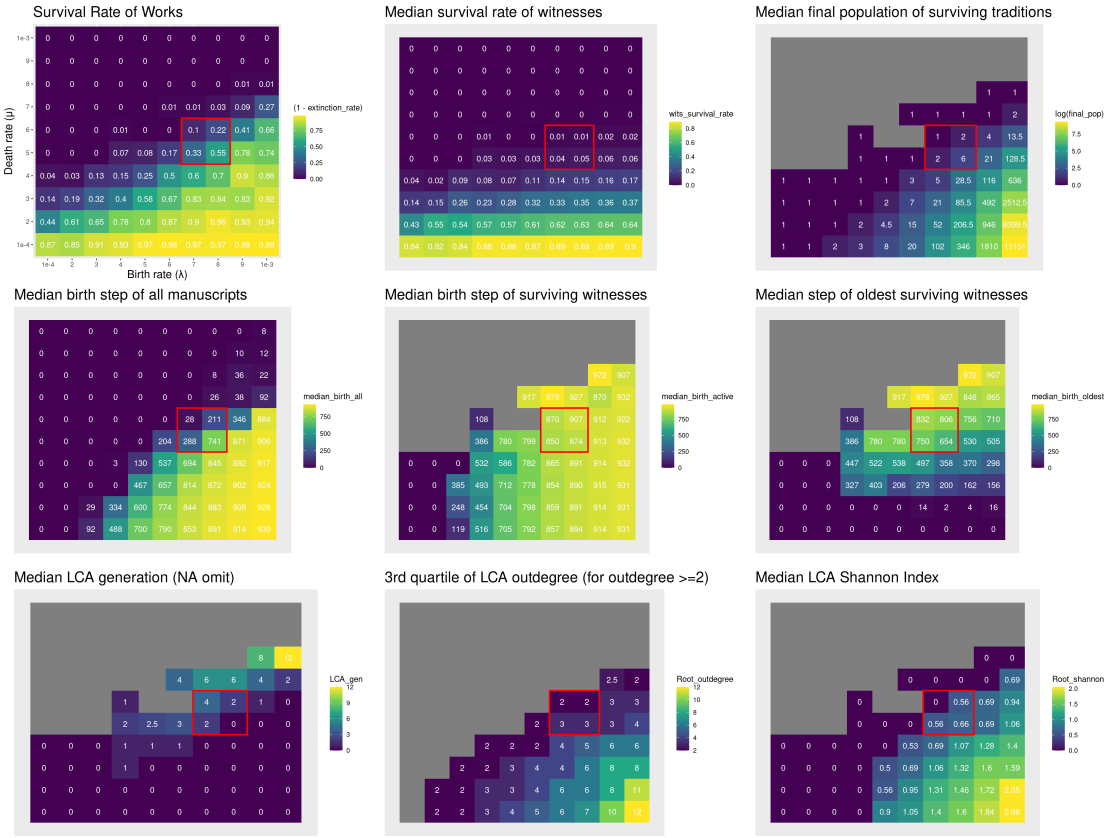

Figure 4: Heat maps (phase diagrams) for the simulation of Cisne-like models; first row contains population size and survival properties, namely **A** Survival rate of traditions (i.e., trees); **B** survival rate of manuscripts (i.e., nodes); **C** the median final population of surviving witnesses for traditions with at least one; second row contains data about the age of manuscripts in the simulation, for **D** all manuscripts ever created; **E** extant ones; as well as **F** the date of the oldest extant manuscript; the third row concerns structural properties of the trees themselves, **F** the median distance between the lowest common ancestor of the surviving manuscripts and the actual root of the original tree; **G** the third quartile of the LCA (archetype) out-degree and **H** and the median Shannon index for the families (the main branches stemming from the LCA). For each pair of parameter values between 0.0001 and 0.001, 100 simulations were run for 1000 active and 1000 inactive steps. Grey areas correspond to irrelevant values of the parameters, or unstable regions. The red square shows parameter regions where the observables computed on the simulated manuscript populations are consistent with observations made for Medieval French epics or with plausible estimates made using different methods.

The situation is, for the moment (and until further data is acquired) consistent yet deserving of further inquiry concerning the distribution of age of surviving manuscripts: in the simulation's red-squared area, the median date of surviving manuscripts would be in the 800's step (around 200 years after the original) and the median date of the oldest for each tradition in the 150-200 years range. For *chansons de geste*, the median date of surviving manuscripts would

be between 1250 and 1300 [50, 11] – 150 to 200 years after the first documents of the genre itself (the end of the 11th century for the composition of the oldest version of the *Roland*). For *romans en vers*, it is, like the genre itself, slightly offset in time, with a peak between 1275 and 1325 [37].

4. Discussion

Combining previous inquiries by Weitzman and Cisne with the power of computer simulations and the methodology of statistical physics, we are able to reproduce the evolutionary process that underlies the observable data for, at least, some textual traditions such as those from medieval French epics and romances. The results obtained can even corroborate or refine results obtained by unrelated methodologies, such as those recently published by Kestemont et al. [33], indicating that these relatively simple birth-and-death process have relevancy in philology as well as they have in Evolutionary Biology for instance. This method then provides us a way to account both for population dynamics in time, loss or production estimates, as well as the shape of the stemmata (the phylogenies) of manuscripts, answering the century long Bédier observation [6], whose lack of solution until now has been at the core of a lasting schism in philological studies.

Indeed, concerning the problem of bifidity, initially raised by Bédier, our simulations tend to show that a ratio of root bifidity of at least 75% can be coherent with other observable properties of the textual traditions of Old French texts, such as the final population or even the date of surviving manuscripts. According to our simulations, it is not necessary to hypothesise any flaw or bias in the method. In fact, it seems that bifidity is one of the measurable properties resulting from the transmission dynamics of manuscript texts.

The range of further investigations opened by this research is considerably large. Models using individual variable rates of λ and μ could be used to account for phenomenon such as efforts of preservation of old and venerable artefacts, or higher selective values of some copies, or accelerated destruction due to small scale (e.g., burnt libraries), larger scale (e.g., the Dissolution of English monasteries, French Wars of Religion,...) or global events (e.g., shift in book types such as from *volumen* to *codex* or caroline to *gothic* scripts, major cultural changes like the Renaissance, ...). Modelling should also include the actual variation of the texts, to reflect the introduction of variants (mutations) in some families, and processes of transmission of inherited variants, as well as lateral transmission.

More generally, once having established this 'null model', deviations due to different factors should be explored, such as higher selective values for some mutations (*variants*) or individuals, fluctuations in time and space and the existence of different 'ecological niches' (e.g., the Anglo-Norman public versus the readers of Franco-Italian epics), typological variation in books or texts, chocks and bottlenecks, etc. The question of the age of surviving manuscripts should also be explored and accounted for, especially in the light of potential variations of λ and μ in time. For instance, the demand and rate of copy for a given text could be expected to be highest shortly after its initial release, when it is most fitted to the taste and fashion of the time, perhaps reinforce itself if the text gets a quick breakthrough, and then decrease with the passing of years. Similarly, the rate of destruction could vary at a global or local level, as some

shocks lead to peaks of destruction or canonicalisation and conservation efforts lead to lower rates. Last but not least, the model should account for non standard transmission, in particular lateral transmission (contamination), a process not uncommon in textual transmission but that is also encountered in the natural world (e.g., lateral gene transfer).

Finally, coming back to the question of the relative importance of drift versus selection, our current results show that a purely stochastic process can account for many observable properties of textual transmission, without having to model differences in selective value for the agents. Yet, to fully answer this question, other type of models have to be experimented, implementing different scenario for selection, and then systematically compared to the results obtained with the current model.

The generality of the models considered here makes them applicable not only to medieval texts, but to any type of cultural transmission, at least in written form, from manuscript circulation to the elaboration of a canon of works. Further investigations should try to verify it on the broadest possible range of cases, starting with Western Medieval and Antique texts, but preferably encompassing cultural productions from very different time periods and continents.

References

[1] C. A. Baker, M. Barbato, M. Cavagna, and Y. Greub, eds. *L'ombre de Joseph Bédier: théorie et pratiques éditoriales au XXe siècle*. Strasbourg: ÉLiPhi, 2018.

[2] A.-L. Barabási and R. Albert. "Emergence of Scaling in Random Networks". In: *Science* 286.5439 (1999), pp. 509–512. DOI: 10.1126/science.286.5439.509. URL: https://science.sciencemag.org/content/286/5439/509.

[3] A. C. Barbrook, C. J. Howe, N. Blake, and P. Robinson. "The phylogeny of the Canterbury Tales". In: *Nature* 394.6696 (1998), pp. 839–839.

[4] H. Bardon. *La littérature latine inconnue*. Paris: C. Klincksieck, 1952.

[5] R.-H. Bautier. "Introduction". In: *Les notaires et secrétaires du roi sous les règnes de Louis XI, Charles VIII et Louis XII, 1461-1515*. Ed. by A. Lapeyre and R. Scheurer. Vol. 1. Paris: Bibliothèque nationale, 1978, pp. Ix–xxxix.

[6] J. Bédier. "La tradition manuscrite du Lai de l'Ombre. Réflexions sur l'art d'éditer les anciens textes (premier article)". In: *Romania* 54.214 (1928), pp. 161–196.

[7] R. A. Bentley. "Random Drift versus Selection in Academic Vocabulary: An Evolutionary Analysis of Published Keywords". In: *Plos One* 3.8 (2008), e3057. DOI: 10.1371/journal.pone.0003057. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003057.

[8] F. Bienvenu, F. Débarre, and A. Lambert. "The split-and-drift random graph, a null model for speciation". In: *Stochastic Processes and their Applications* 129.6 (2019), pp. 2010–2048. DOI: 10.1016/j.spa.2018.06.009. URL: https://www.sciencedirect.com/science/article/pii/S0304414918303004.

[9] E. Buringh. *Medieval Manuscript Production in the Latin West*. Brill, 2010-11-19. DOI: 10.1163/9789047428640. URL: http://booksandjournals.brillonline.com/content/books/9789047428640.

[10] J. S. Cabral, L. Valente, and F. Hartig. "Mechanistic simulation models in macroecology and biogeography: state-of-art and prospects". In: *Ecography* 40.2 (2017), pp. 267–280. DOI: 10.1111/ecog.02480.

[11] J.-B. Camps. "La 'Chanson d'Otinel': édition complète du corpus manuscrit et prolégomènes à l'édition critique". thèse de doctorat, dir. Dominique Boutet. Paris: Paris-Sorbonne, 2016. DOI: 10.5281/zenodo.1116735. URL: https://halshs.archives-ouvertes.fr/tel-01664932.

[12] J.-B. Camps, G. Fernandez Riva, and S. Gabay. *Open Stemmata: Database*. 2021. URL: https://github.com/OpenStemmata/database/.

[13] J.-B. Camps, S. Gabay, and G. F. Riva. "Open Stemmata: A Digital Collection of Textual Genealogies". In: *EADH2021: Interdisciplinary Perspectives on Data, 2nd International Conference of the European Association for Digital Humanities*. Krasnoyarsk, 2021. URL: https://halshs.archives-ouvertes.fr/halshs-03260086.

[14] P. Canettieri, V. Loreto, M. Rovetta, and G. Santini. "Philology and Information Theory". In: *Cognitive Philology* 1 (2008). URL: http://ojs.uniroma1.it/index.php/cogphil/article/view/8816.

[15] O. Canteaut. "Quantifier l'activité des chancelleries à l'aune de la tradition des actes : l'exemple de la chancellerie des derniers Capétiens (1314-1328)". In: *Actes royaux et princiers à l'ère du numérique (Moyen Âge-Temps modernes)*. Ed. by O. Canteaut, O. Guyotjeannin, and O. Poncet. Pau, 2020, pp. 103–114.

[16] A. Castellani. *Bédier avait-il raison?: La méthode de Lachmann dans les éditions de textes du Moyen Age. Leçon inaugurale donnée à l'université de Fribourg le 2 juin 1954*. Discours universitaires, Nouvelle série = Freiburger Universitätsreden, Neue Folge 20. Fribourg (Suisse): Éditions Universitaires, 1957.

[17] D. Chavalarias and J.-P. Cointet. "Phylomemetic Patterns in Science Evolution–The Rise and Fall of Scientific Fields". In: *Plos One* 8.2 (2013), e54847. DOI: 10.1371/journal.pone.0054847. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0054847.

[18] J. L. Cisne. "How Science Survived: Medieval Manuscripts' "Demography" and Classic Texts' Extinction". In: *Science* 307.5713 (2005-02-25), pp. 1305–1307. DOI: 10.1126/science.1104718. URL: http://science.sciencemag.org/content/307/5713/1305.

[19] J. Cosette, A. Moussy, F. Onodi, A. Auffret-Cariou, T. M. A. Neildez-Nguyen, A. Paldi, and D. Stockholm. "Single cell dynamics causes Pareto-like effect in stimulated T cell populations". In: *Scientific reports* 5.1 (2015), pp. 1–10.

[20] F. Duval. *La 'Tradition manuscrite du Lai de l'Ombre' de Joseph Bédier ou la critique textuelle en question*. Paris: Honoré Champion, 2021.

[21] D. Ganz and W. Goffart. "Charters Earlier than 800 from French Collections". In: *Speculum* 65.4 (1990), pp. 906–932. DOI: 10.2307/2863567.

[22] W. W. Greg. “Recent theories of textual criticism”. In: *Modern Philology* 28.4 (1931), pp. 401–404.

[23] V. Guidi and P. Trovato. “Sugli stemmi bipartiti. Decimazione, asimmetria e calcolo delle probabilità”. In: *Filologia italiana* 1 (2004), pp. 9–48.

[24] O. E. Haugen. “2 The genealogical method”. In: *Handbook of Stemmatology*. Ed. by R. Philipp. De Gruyter, 2020, pp. 57–138. DOI: 10.1515/9783110684384-003.

[25] O. E. Haugen. “The silva portentosa of stemmatology: Bifurcation in the recension of Old Norse manuscripts”. In: *Digital Scholarship in the Humanities* 31.3 (2015), pp. 594–610. DOI: 10.1093/llc/fqv002. URL: https://academic.oup.com/dsh/article/31/3/594/2340338.

[26] A. Hoenen. “History of computer-assisted stemmatology”. In: *Handbook of Stemmatology*. Ed. by R. Philipp. De Gruyter, 2020, pp. 294–303. DOI: 10.1515/9783110684384.

[27] A. Hoenen. “Silva Portentosissima – Computer-Assisted Reflections on Bifurcativity in Stemmas”. In: *Digital Humanities 2016 (DH2016): Conference Abstracts. Jagiellonian University & Pedagogical University*. Kraków, 2016, pp. 557–560. URL: http://dh2016.adho.org/abstracts/311.

[28] A. Hoenen. “The stemma as a computational model”. In: *Handbook of Stemmatology*. Ed. by R. Philipp. De Gruyter, 2020, pp. 226–241. URL: 10.1515/9783110684384.

[29] A. Hoenen, S. Eger, and R. Gehrke. “How Many Stemmata with Root Degree k?” In: *Proceedings of the 15th Meeting on the Mathematics of Language*. 2017, pp. 11–21.

[30] E. Holtz. “Überlieferungs- und Verlustquoten spätmittelalterlicher Herrscherurkunden”. In: *Turbata per aequora mundi: Dankesgabe an Eckhard Müller-Mertens*. Ed. by M. Lawo and O. B. Rader. Hanover: Harrassowitz, 2001, pp. 67–80.

[31] C. J. Howe, A. C. Barbrook, M. Spencer, P. Robinson, B. Bordalejo, and L. R. Mooney. “Manuscript evolution”. In: *Trends in Genetics* 17.3 (2001), pp. 147–152. DOI: 10.1016/s0168-9525(00)02210-1. URL: https://www.sciencedirect.com/science/article/pii/S0168952500022101.

[32] M. Kestemont and F. Karsdorp. “Estimating the Loss of Medieval Literature with an Unseen Species Model from Ecodiversity”. In: *Proceedings of the Workshop on Computational Humanities Research*. Vol. 2723. Ceur. 2020, pp. 44–55. URL: http://ceur-ws.org/Vol-2723/short10.pdf.

[33] M. Kestemont, F. Karsdorp, E. de Bruijn, M. Driscoll, K. A. Kapitan, P. Ó Macháin, D. Sawyer, R. Sleiderink, and A. Chao. “Forgotten books: The application of unseen species models to the survival of culture”. In: *Science* 375.6582 (2022), pp. 765–769. DOI: 10.1126/science.abl7655.

[34] P. Maas. “Leitfehler und stemmatische Typen”. In: *Byzantinische Zeitschrift* 37.2 (1937), pp. 289–294. DOI: 10.1515/byzs.1937.37.2.289.

[35] G. M. Mace, J. L. Gittleman, and A. Purvis. “Preserving the Tree of Life”. In: *Science* 300.5626 (2003), pp. 1707–1709. DOI: 10.1126/science.1085510. URL: https://science.sciencemag.org/content/300/5626/1707.

[36] C. Macé, ed. *The Evolution of Texts: confronting stemmatological and genetical methods. Proceedings of the international workshop held in Louvain-la Neuve on September 1-2, 2004.* Pisa: Istituti editoriali e poligrafici internazionali, 2006.

[37] P. A. Martina. "La produzione manoscritta del romanzo francese in versi : modelli materiali e modelli di cultura". These de doctorat. Sorbonne université, 2018. URL: https://www.theses.fr/2018SORUL051.

[38] U. Neddermeyer. "Von der Handschrift zum gedruckten Buch: Schriftlichkeit und Leseinteresse im Mittelalter und in der frühen Neuzeit quantitative und qualitative Aspekte". PhD thesis. Wiesbaden: Harrassowitz, 1998.

[39] M. Olave, L. J. Avila, J. W. Sites Jr, and M. Morando. "Model-based approach to test hard polytomies in the Eulaemus clade of the most diverse South American lizard genus Liolaemus (Liolaemini, Squamata)". In: *Zoological Journal of the Linnean Society* 174.1 (2015), pp. 169–184.

[40] F. v. Oostrom. *Stemmen op schrift: geschiedenis van de Nederlandse literatuur vanaf het begin tot 1300.* Amsterdam: Bert Bakker, 2013.

[41] E. Overgaauw. "Fast or slow, professional or monastic. The writing speed of some late-medieval scribes". In: *Scriptorium* 49.2 (1995), pp. 211–227. DOI: 10.3406/scrip.1995.1726. URL: https://www.persee.fr/doc/scrip%5C%5F0036-9772%5C%5F1995%5C%5Fnum%5C%5F49%5C%5F2%5C%5F1726.

[42] T. F. Rangel, N. R. Edwards, P. B. Holden, J. A. F. Diniz-Filho, W. D. Gosling, M. T. P. Coelho, F. A. S. Cassemiro, C. Rahbek, and R. K. Colwell. "Modeling the ecology and evolution of biodiversity: Biogeographical cradles, museums, and graves". In: *Science* 361.6399 (2018). DOI: 10.1126/science.aar5452. URL: https://science.sciencemag.org/content/361/6399/eaar5452.

[43] D. M. Raup. *Extinction: bad genes or bad luck?* WW Norton & Company, 1992.

[44] B. Richardson, ed. *Giovan Francesco Fortunio: Regole grammaticali della volgar lingua.* Roma: Antenore, 2001.

[45] C. Segre, ed. *La chanson de Roland.* Documenti di filologia 16. Milan et Naples: R. Ricciardi, 1971.

[46] W. P. Shepard. "Recent theories of textual criticism". In: *Modern Philology* 28.2 (1930), pp. 129–141.

[47] M. Spencer, E. A. Davidson, A. C. Barbrook, and C. J. Howe. "Phylogenetics of artificial manuscripts". In: *Journal of Theoretical Biology* 227.4 (2004), pp. 503–511. DOI: 10.1016/j.jtbi.2003.11.022. URL: https://www.sciencedirect.com/science/article/pii/S0022519303004442.

[48] S. Timpanaro. *La genesi del metodo del Lachmann.* 4th. Torino: UTET Libreria, 2003.

[49] P. Trovato. *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.* Limena: Libreriauniversitaria.it edizioni, 2014.

[50] A. Vitale-Brovarone. “La diffusion manuscrite des chansons de geste: une vue d’ensemble”. In: *Tra Italia e Francia. Entre France et Italie. In honorem Elina Suomela-Härmä*. Ed. by E. Garavelli, M. Helkkula, and O. Välikangas. Mémoire de la Société Néophilologique de Helsinki 69. Helsinki, 2006, pp. 473–488.

[51] M. P. Weitzman. “Computer simulation of the development of manuscript traditions.” In: *Allc Bull.* 10.2 (1982), pp. 55–59.

[52] M. P. Weitzman. “The Evolution of Manuscript Traditions”. In: *Journal of the Royal Statistical Society. Series A (General)* 150.4 (1987), pp. 287–308. DOI: 10.2307/2982040. URL: http://www.jstor.org/stable/2982040.

[53] H. Wijsman. *Luxury Bound: Illustrated Manuscript Production and Noble and Princely Book Ownership in the Burgundian Netherlands (1400-1550)*. Vol. 16. Burgundica. Turnhout: Brepols Publishers, 2010. DOI: 10.1484/m.burg-eb.5.105851.

[54] Wilson R M. *The Lost Literature Of Medieval England*. Methuen, 1952. URL: http://archive.org/details/in.ernet.dli.2015.86593.

[55] K. Yessoufou and T. J. Davies. “Reconsidering the Loss of Evolutionary History: How Does Non-random Extinction Prune the Tree-of-Life?” In: *Biodiversity Conservation and Phylogenetic Systematics: Preserving our evolutionary heritage in an extinction crisis*. Ed. by R. Pellens and P. Grandcolas. Topics in Biodiversity and Conservation. Cham: Springer International Publishing, 2016, pp. 57–80. DOI: 10.1007/978-3-319-22461-9_4.