How Many Stemmata with Root Degree k?

Armin Hoenen
CEDIFOR
Goethe University Frankfurt
hoenen@em.uni-frankfurt.de

Steffen Eger
UKP
TU Darmstadt
eger@ukp.informatik.tu-darmstadt.de

Ralf Gehrke
CEDIFOR
Goethe University Frankfurt
gehrke@rz.uni-frankfurt.de

Abstract

We are investigating parts of the mathematical foundations of stemmatology, the science reconstructing the copying history of manuscripts. After Joseph Bédier in 1928 got suspicious about large amounts of root bifurcations he found in reconstructed stemmata, Paul Maas replied in 1937 using a mathematical argument that the proportion of root bifurcating stemmata among all possible stemmata is so large that one should not become suspicious to find them abundant. While Maas' argument was based on one example with a tradition of three surviving manuscripts, we show in this paper that for the whole class of trees corresponding to Maasian reconstructed stemmata and likewise for the class of trees corresponding to complete historical manuscript genealogies, root bifurcations are apriori the most expectable root degree type. We do this by providing a combinatorial formula for the numbers of possible so-called *Greg trees* according to their root degree (Flight, 1990). Additionally, for complete historical manuscript trees (regardless of loss), which coincide mathematically with *rooted labeled trees*, we provide formulas for root degrees and derive the asymptotic degree distribution. We find that root bifurcations are extremely numerous in both kinds of trees. Therefore, while previously other studies have shown that root bifurcations are expectable for true stemmata, we enhance this finding to all three philologically relevant types of trees discussed in breadth until today.

1 Introduction

Stemmatology is the science trying to reestablish the copy history of a text surviving in a number of versions. One of the editors' objectives in stemmatology can be approaching the original authorial wording, which itself is most probably lost, given the body of extant text variants (Cameron, 1987).

In order to do so, the philologist may reconstruct the copy history of the manuscripts so as to better understand which variants are most likely original. Usually, the visual reconstruction is a graph or more precisely a tree where the nodes symbolize manuscripts and the copy processes are depicted by the edges. Such a visual reconstruction is then called a stemma. For an example of a stemma, see Figure 1.

Maybe the biggest and surely most famous problem in philology is an observation that the French philologist Joseph Bédier made editing the medieval French text "Le lai de l'ombre" in 1890, 1913 and 1928 (Bédier, 1890, 1913, 1928). Bédier observed that 105 out of 110 stemmata, the vast majority, in a collection he had made without controlling for root degree patterns had a bifurcation immediately below their root, an observation repeated multiple times thereafter on different collections, compare Table 1.

This observation was worrisome. If there are exactly two texts (nodes) directly below the assumed authorial original (root),[1] the implications for text reconstruction of the urtext are the following. An editor may choose one of the two texts as his/her preferred base text at will and reconstruct the ancestral text from this base text eliciting only in special cases the second or yet another variant.

[1] More precisely, in most cases, a root of such a tree represents a hypothetical intermediary: the latest common ancestor of all survivors. It corresponds to the oldest objectively reconstructible text and is called archetype.

Proceedings of the 15th Meeting on the Mathematics of Language, pages 11–21,
London, UK, July 13–14, 2017.

Collection	root bifurcations	root tri- or multifurcations
Bédier (1928)	95.5%	4.5%
Castellani (1957)	82.5%	17.5%
Haugen (2015) Bibliotheca A.	85.5%	14.5%
Haugen (2015) Editiones A.	80.5%	19.5%

Table 1: Percentages of root bifurcative stemmata in four collections, reported in (Haugen, 2015). Note that extending his collection through stemmata which are not yet viewed as conclusive by the composer, Castellani (1957, p.24) reports only $75-76\%$ root bifurcating trees.

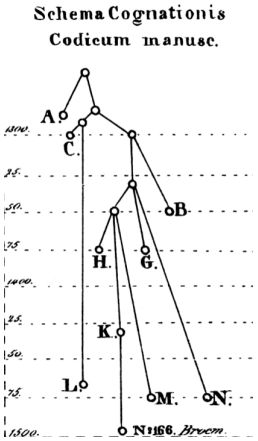

Figure 1: First modern stemma by Schlyter, 1827, from O'Hara (1996).

Bédier was worried about editors consciously or subconsciously choosing a base manuscript for the urtext after their taste and justifying this by postulating root bifurcations in their stemmata. As a second explanation for a large incidence of root bifurcations in reconstructed stemmata he suspected a methodology-inherent tendency for overseparation since editors always look for the *one* authorial in opposition to all other variants (a fallacy of the stemmatic method).

One can easily imagine that the subsequent debate had far-reaching consequences for textual criticism and editing. The community divided into best text editors (or Bédierists) which abandoned stemmatic approaches altogether and based their editions on a good available manuscript and those which continued and continue to produce stemmata (or Lachmannians). More realistically, any modern editor may choose among one of those approaches depending on his/her material and circumstances. Nevertheless, the argument has ever since stimulated much research repeatedly including mathematical argumentation, see for instance, Greg (1931), Maas (1937), Fourquet (1946), Whitehead (1951), Pasquali (1952), Castellani (1957), Hering (1967), Kleinlogel (1968), Weitzman (1982), Weitzman (1987), Grier (1989), Haugen (2002), Timpanaro (2005), Haugen (2010), Haugen (2015), Hoenen (2016). Maas argued that the number of stemmata with a root bifurcation among all possible stemmata which can be reconstructed (thus regarding stemma generation apriori as a random process) would be naturally high. One should thus rather not be too surprised of large proportions in real reconstructed stemmata: those were no good reason to abandon the stemmatic method. Maas numerically based this counter argument on the example of traditions with three surviving manuscripts.[2] Bédierists could have reacted to this and could have tried to seek a generalization of his argument. However, neither Bédierists nor Lachmannians have ever come up with such a generalization. What if Maas' argument would only hold for three surviving manuscripts, but witness completely different proportions for 4, 5, or 60 survivors? Would those numbers reveal justification for being suspicious of the real-world reconstructions?

In fact, Maas himself estimated numbers of possible stemmata for a number of surviving manuscripts of up to 5 according to Flight (1990), who decades later generalized the type of graphs Maas had considered for the modeling of stemmata. Flight (1990) provided a formula to count numbers of these so-called Greg trees, given a certain number of survivors. However, the question of the proportion of root bifurcating stem-

[2]Maas distinguishes two kinds of traditions of medieval texts: texts read by many and texts read by few. He assumes that strict stemmatics fails for texts read by many, which should be characterized by a larger number of survivors. Yet, not all philologists follow this distinction. Pasquali and Pieraccioni (1952) distinguish *open* and *closed* traditions, where the latter are such which are largely free of flaws complicating stemmatic assessment. Closed traditions are not straighforwardly connected with the number of survivors, compare also West (1973), which is why there is no reason to limit the range of surviving manuscripts to very small numbers and surely not to just one or two examples.

mata and how this proportion develops—thus ultimately the generalization of Maas' argument—has not yet been answered. In this paper, we fill this gap and provide a formula for the numbers of possible root k-furcating stemmata given m surviving manuscripts and compute the proportion of root bifurcating stemmata among all stemmata given m survivors.

Our work connects to a tradition both in linguistics and biology to count certain subclasses of graphs. In our case these graphs are trees, whereas other works have counted alignments between two or multiple sequences, that is, certain bi- or multi-partite graphs (Griggs et al., 1990; Covington, 2004; Eger, 2015).

2 Counting Manuscript Trees: Prerequisites

The theoretical entity used to model manuscript genealogy is a *tree*. A tree, as a concept from graph theory, is a set of *nodes* V together with a set of (unordered) *edges* E, with $E \subseteq \{\{u, v\} \,|\, u, v \in V\}$. The two defining properties of trees is that they must be free of cycles (including self-cycles) and connected. General works on counting different types of trees appear early on (Cayley, 1889), and research on trees is comprehensive, compare Moon (1970). The similarity of the three disciplines of historical linguistics, phylogeny and stemmatology has likewise been noticed early and led to various transfers and adaptations between methods of those fields, compare O'Hara (1996). Especially in the domain of phylogeny the understanding of trees is a central issue and consequently much research has focussed on phylogenetic trees, see for instance Felsenstein (1978); Swofford (1990); Huson (1998); Felsenstein (2004). One characteristic of phylogenetic trees is that they are apriori exclusively bifurcating. Thus, the question for a proportion of root bifurcating trees becomes meaningless. Apart from this, the manual reconstruction of a consistent and complete genome or characterome of ancestors is by no means as central an issue as in stemmatics (Platnick and Cameron, 1977; Cameron, 1987).

In the context of manuscript trees, although a number of the above enumerated philological studies count stemmatic trees under certain conditions or elaborate on specific phenomena, Flight (1990) is apparently the first to provide a generalized definition for stemmas. He aims at solving the question, which he attributes to Maas (1958), how many different stemmas may exist for some given number of surviving manuscripts (Flight, 1990, p.122).

To solve this, he counts so called *Greg trees*.[3] Based on Flight (1990), we define a rooted directed Greg tree (which Flight names after the textual critic W. W. Greg) as a tree with a distinguished root, m labeled nodes standing for surviving manuscripts and n unlabeled nodes symbolizing hypothetical manuscripts. The latter must have an outdegree of at least two. There can be neither chains of hypothetical manuscripts (unlabeled nodes) with indegree one and outdegree one nor unlabeled leafs. This restriction corresponds to philological practice (Maas, 1937). A rooted Greg tree therefore symbolizes a reconstructed stemma. With this definition, Flight (1990) recovers the numbers of possible trees for three surviving manuscripts as postulated by Maas (1937), see Figure 2. Flight (1990) gives a recursive formula for the enumeration of unrooted and rooted Greg trees, building on all (four) generalized conditions on how to add a new labeled node and tabulates all possible Greg trees for up to 12 labeled nodes. Thus, he extends values mentioned by Maas as well as corrects Maas' numbers. From the 22 rooted Greg trees for 3 survivors, there are 12 root bifurcating ones, compare again Figure 2. The recursive formula Flight gives for rooted Greg trees $g(m, n)$ on m labeled and n unlabeled nodes is:[4]

$$\begin{aligned} g(m,n) = &(m+n-2)\cdot g(m-1,n-1) \\ &+ (2m+2n-2)\cdot g(m-1,n) \\ &+ (n+1)\cdot g(m-1,n+1). \end{aligned}$$

If we fix m, the number of unlabeled nodes n can vary in the range of $\{0, 1, \ldots, m-1\}$ and the sum, over n, of all such (m, n)-trees for a fixed m is the number $g(m)$ of possible rooted Greg trees for m survivors (Flight, 1990). This gives the number of possible stemmata one can reconstruct for m surviving manuscripts adhering to philological principles.[5]

[3] According to Josuat-Vergès (2015), a similar problem in phylogeny has been described and tackled by Felsenstein (1978) as recognized by Knuth (2005).

[4] Flight refers to these as g^*, but for brevity and since we do not deal with unrooted Greg trees, we denote them simply as g.

[5] The number sequence $g(m)$ is listed as integer sequence A005264 in the On-Line Encyclopedia of Integer Sequences (OEIS), published electronically at `https://oeis.org`.

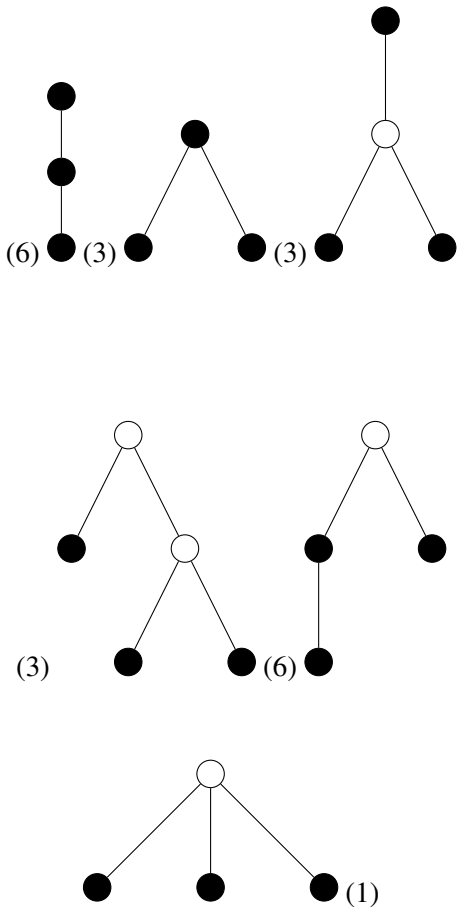

Figure 2: The unlabeled rooted (root topmost node) topologies of possible stemmata for three surviving manuscripts as thought of by Maas (1937). White nodes symbolize reconstructed lost manuscripts (unlabeled) whereas black nodes are survivors (labeled). The number in brackets refers to the number of possible distinct labeled trees (label permutations) for each topology.

While Flight (1990) does not compute numbers of Greg trees according to their root degree, Hering (1967), referring to a colleague of his,[6] tabulates the numbers of root k-furcating Greg trees (and the numbers of rooted Greg trees being the sum over all k) up to $m = 6$. The sums for all k at a fixed m coincides exactly with $g(m)$ calculated by Flight (1990). Alas, there is no formula provided by Hering (1967). Furthermore, he states that a calculation for more than 6 survivors would be difficult. This is demoralizing insofar as surely numbers (much) larger than $m = 6$ are relevant to the philological debate. For instance, according to Weitzman (1987), numbers of survivors in Greek and Latin traditions can range from 1 to "well over 100".

3 Counting Manuscript Trees: New Formulas

3.1 A Meta Formula

First, we present a general formula for counting trees with fixed root degree and two different types of nodes (e.g., black and white), which we use later on to derive our main results. We write $\mathcal{T}$ for a class of trees and T for $|\mathcal{T}|$.

If the root of a rooted tree has degree k and the tree has μ black nodes and ν white nodes, it means that the tree has k subtrees, which we also perceive as rooted. The root node, r, is either black or white. We connect r to the root of each subtree. Each of these subtrees can have some size $s_1 + p_1, \ldots, s_k + p_k$, where s_i is the number of black nodes in branch i and p_i is the number of white nodes in the same branch. The sum of the s_i must equal $\mu - \delta_B$ and the sum of the p_i must equal $\nu - \delta_W$, since there are in total μ black nodes and ν white nodes. Here, δ_B is a binary variable indicating whether r is a black node and analogously for δ_W, where $\delta_B = 1$ iff $\delta_W = 0$. If the black nodes are distinguishable, we can choose the subsets of nodes of sizes $s_1, \ldots, s_k$ from a total of $\mu - \delta_B$ nodes, and analogously for the white nodes. There are $\binom{\mu-\delta_B}{s_1,\ldots,s_k}$ possibilities to do so, where $\binom{m}{k_1,\ldots,k_\ell} = \frac{m!}{k_1!\cdots k_\ell!}$ are the multinomial coefficients.

Now, we specialize. We assume that the black nodes are distinguishable and the white nodes are indistinguishable. Then, for any class of rooted trees $\mathcal{T}_{\mu,\nu}$ with μ such black nodes and ν such

[6] Prof. Dr. Wolfgang Engel, a mathematician from Rostock University.

white nodes, the number $T_{\mu,\nu,k}$ of rooted labeled trees from $\mathcal{T}_{\mu,\nu}$ in which the root has degree k has the form

$$\mu \sum_{(\mathbf{s},\mathbf{p})\in\mathcal{C}((\mu-1,\nu),k)} \binom{\mu-1}{\mathbf{s}} F(\mathbf{s},\mathbf{p}) + \sum_{(\mathbf{s},\mathbf{p})\in\mathcal{C}((\mu,\nu-1),k)} \binom{\mu}{\mathbf{s}} F(\mathbf{s},\mathbf{p}).$$

Here, $\mathcal{C}((a,b),\ell)$ denotes the number of *vector compositions* (Eger, 2017) of the 'vector' $(a,b) \in \mathbb{N}^2$ with ℓ parts; that is,

$$\mathcal{C}((a,b),\ell) = \{(s_1,\dots,s_\ell),(p_1,\dots,p_\ell) \mid s_1+\dots+s_\ell = a,\ p_1+\dots+p_k = b\}.$$

Moreover, by $\mathbf{s}$ and $\mathbf{p}$, we denote tuples $(s_1,\dots,s_k)$ and $(p_1,\dots,p_k)$, respectively. The above sum formula arises because the root node can either be black or white. If it is black, we have the additional factor μ because the black nodes are distinguishable and each of them can be the root.

Finally, F is a function of the sizes $s_1,\dots,s_k,p_1,\dots,p_k$ which will be specified in any particular case.

Now, we have overcounted $T_{\mu,\nu,k}$ since we have counted subtrees as if they were ordered, while in reality different orders of the subtrees do not constitute a distinct tree $t \in \mathcal{T}_{\mu,\nu,k}$. Thus, we have to divide by $k!$ to finally arrive at:

$$T_{\mu,\nu,k} = \frac{\mu}{k!} \sum_{(\mathbf{s},\mathbf{p})\in\mathcal{C}((\mu-1,\nu),k)} \binom{\mu-1}{\mathbf{s}} F(\mathbf{s},\mathbf{p}) + \frac{1}{k!} \sum_{(\mathbf{s},\mathbf{p})\in\mathcal{C}((\mu,\nu-1),k)} \binom{\mu}{\mathbf{s}} F(\mathbf{s},\mathbf{p}). \tag{1}$$

It is possible that $T_{\mu,\nu,k}$ can be expressed simpler—e.g., as a linear combination of the terms $T_{\mu+\tau,\nu+\rho,k+\kappa}$ for integers τ,ρ,κ—for specific choices of F.

3.2 *Root k-furcating Greg Trees*

We are now ready to derive the general formula for the number $g_k(m,n)$ of root k-furcating Greg trees for m survivors (labeled nodes) and n hypothetical (unlabeled) nodes.

The only question remaining from above is how we have to specify the function $F(\mathbf{s},\mathbf{p})$ on the k subtrees. This is very simple, however. Since all branches i are independent of each other, $F(\mathbf{s},\mathbf{p})$ takes the form of a product of individual factors:

$$F(\mathbf{s},\mathbf{p}) = \prod_{i=1}^{k} g(s_i,p_i)$$

where g is the function of Flight (1990). The number $g_k(m,n)$ of root k-furcating Greg trees for m survivors and n hypothetical nodes is hence given by (1) with this specification of F.

We make three additional remarks. The s_i satisfy $s_i \geq 1$, since the specification of Greg trees disallows to have only unlabeled nodes (i.e., $s_i = 0$) in a branch. In contrast, the p_i may take on the value zero and therefore satisfy $p_i \geq 0$. Moreover, the p_i actually satisfy $0 \leq p_i < s_i$ because of the link restrictions on unlabeled nodes in Greg trees. While the constraint on the p_i's is automatically taken care of by the function g of Flight (1990), explicitly accounting for it can speed up computations.[7] Finally, when $k = 1$, we have to exclude the second term in (1) from consideration because, by definition, the root of a Greg tree cannot have degree one when it is unlabeled.

The numbers $g_k(m)$ of root k-furcating Greg trees for m survivors and an arbitrary number of hypothetical manuscripts n is the sum over n of root k-furcating (m,n)-trees. In other words,

$$g_k(m) = \sum_{n\geq 0} g_k(m,n).$$

Table 2 shows the growth of $g_k(m)$ until $m,k = 15$.

We are now interested in the proportions of root bifurcating Greg trees among all Greg trees since this was aluded to in Bédier (1928). That is, we investigate the ratio

$$R_2(m) = \frac{g_2(m)}{\sum_{k\geq 1} g_k(m)}.$$

[7] In order to more efficiently compute the numbers, we also used further simplified formulas for specific k where possible. Root unifurcating Greg trees (here g_1) are especially easily computed. The root can only be labeled, since an unlabeled node as root must have degree at least two. Then, the number of possible root unifurcating Greg trees corresponds to $m \cdot g(m-1)$. Root-$(m-1)$-furcating rooted Greg trees for all $m \neq 2$ coincide with the pentagonal numbers (sequence $A000326$ in the OEIS), whose number is given by $\frac{3m^2-m}{2}$. This is so because there are only three principle architectures of root-$(m-1)$-furcating rooted Greg trees, the individual formulas for the enumeration of which sum to the same as the pentagonal numbers: $m + m(m-1) + \binom{m}{2}$. Finally, for a root m-furcation, there is always only one Greg tree.

m \ k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Σ
1	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
2	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	3
3	9	12	1	-	-	-	-	-	-	-	-	-	-	-	-	22
4	88	151	22	1	-	-	-	-	-	-	-	-	-	-	-	262
5	1 310	2 545	445	35	1	-	-	-	-	-	-	-	-	-	-	4 336
6	26 016	54 466	10 425	1 025	51	1	-	-	-	-	-	-	-	-	-	91 984
7	643 888	1 417 318	286 321	31 780	2 030	70	1	-	-	-	-	-	-	-	-	2 381 408
8	19 051 264	43 472 780	9 102 604	1 090 201	80 360	3 626	92	1	-	-	-	-	-	-	-	72 800 928
9	655 208 352	1 536 228 588	329 980 456	41 636 973	3 368 001	178 290	6 006	117	1	-	-	-	-	-	-	2 566 606 784
10	25 666 067 840	61 466 251 616	13 457 494 060	1 763 775 280	152 280 345	8 964 417	358 890	9 390	145	1	-	-	-	-	-	102 515 201 984
11	$1.13 * 10^{12}$	$2.75 * 10^{12}$	$6.1 * 10^{11}$	$8.23 * 10^{10}$	$7.46 * 10^{9}$	$4.74 * 10^{8}$	$2.13 * 10^{7}$	$6.61 * 10^{5}$	$1.4 * 10^{4}$	176	1	-	-	-	-	$4.58 * 10^{12}$
12	$5.49 * 10^{13}$	$1.36 * 10^{14}$	$3.05 * 10^{13}$	$4.21 * 10^{12}$	$3.96 * 10^{11}$	$2.66 * 10^{10}$	$1.3 * 10^{9}$	$4.64 * 10^{7}$	$1.18 * 10^{6}$	$2.02 * 10^{4}$	210	1	-	-	-	$2.26 * 10^{14}$
13	$2.93 * 10^{15}$	$7.33 * 10^{15}$	$1.66 * 10^{15}$	$2.34 * 10^{14}$	$2.27 * 10^{13}$	$1.59 * 10^{12}$	$8.31 * 10^{10}$	$3.25 * 10^{9}$	$9.41 * 10^{7}$	$1.97 * 10^{6}$	$2.82 * 10^{4}$	247	1	-	-	$1.22 * 10^{16}$
14	$1.71 * 10^{17}$	$4.31 * 10^{17}$	$9.86 * 10^{16}$	$1.41 * 10^{16}$	$1.39 * 10^{15}$	$1.02 * 10^{14}$	$5.58 * 10^{12}$	$2.34 * 10^{11}$	$7.47 * 10^{9}$	$1.71 * 10^{8}$	$3.16 * 10^{6}$	$3.83 * 10^{4}$	287	1	-	$7.15 * 10^{17}$
15	$1.07 * 10^{19}$	$2.73 * 10^{19}$	$6.29 * 10^{18}$	$9.01 * 10^{17}$	$9.2 * 10^{16}$	$6.89 * 10^{15}$	$3.94 * 10^{14}$	$1.75 * 10^{13}$	$6.03 * 10^{11}$	$1.61 * 10^{10}$	$3.27 * 10^{8}$	$4.89 * 10^{6}$	$5.01 * 10^{4}$	330	1	$4.53 * 10^{19}$

Table 2: Numbers of root k-furcations for all possible k for rooted Greg trees with m survivors up to $m = 15$, $g_k(m)$. Note that the first numbers until $m = 6$ occur in Hering (1967). Exact numbers are provided until $m = 10$ and for $(m-1)$-trees and otherwise numbers are in scientific notation. The last column contains the sum $\sum_{k=1}^{m} g_k(m)$ which equals the number of all rooted Greg trees for the current m, compare Flight (1990). Code for computation is available from https://github.com/ArminHoenen/KFurcatingRootedGregTrees. The sequence of numbers for $k = 2$ is now integer sequence A286432 in the OEIS.

At $m = 2$, the proportion is one third, at $m = 3$, Maas' famous example, we witness a proportion of $R_2 = 0.54545$. For $m = 10$, R_2 is already 0.59958 with the increase slowing down. For $m = 20$, we have $R_2 = 0.60351$ and at $m = 100$ survivors the proportion is $R_2 = 0.60599$. Growth is further slowing down and at $m = 200$ the proportion is $R_2 = 0.60626$. While we are not able to prove it, we think it is a very safe conjecture that $R_2(m)$ converges to below 0.607, as m tends toward infinity. Figure 3 plots the proportions of trees with root degree $k = 1$, $k = 2$ and $k > 2$, as m becomes larger. Figure 4 plots the root degree distribution for fixed m.

Root bifurcations thus outweigh all other root degree patterns by far. Maas' argument was therefore generally true as what regards a large expectability of root bifurcations in reconstructible stemmata. Nevertheless, the observed proportions are considerably lower than Bédier's ones. However, a better fit occurs when we exclude all trees with root degree one from consideration. A root degree of one requires root to be labeled and thus surviving, a case which is empirically probably quite rare, although not impossible. In Bédier's collection presumably, there simply had not been any root unifurcating stemma with a surviving root and he does not comprehensively discuss this general possibility. In Castellani's (1957) and Haugen's (2010) collections there have been no counts of root unifurcations. At $m = 200$, the fraction of unifurcating trees is about 21.467%, which means that the fraction of trees with root degree two is

$$\tilde{R}_2(m) = \frac{0.60626}{1 - 0.21467} = 0.7719$$

at $m = 200$, when trees with root degree one are discarded. Comparing this number to those in Table 1, we observe that the empirically reported numbers for actual collections of stemmata are just slightly above this reference point. This would indicate that there seems to be a bias for root bifurcations, but that this bias is rather low.

While Bédier had looked at $R_2(m)$ or $\tilde{R}_2(m)$ (coinciding in his collection), Maas explicitly looked at

$$R_{k>2}(m) = \frac{\sum_{k>2} g_k(m)}{\sum_{k\geq 1} g_k(m)}$$

for $m = 3$, and based his counter argument to Bédier's conclusions on that. This has been criticized variously because $R_2(3)$ corresponds to 12 in 22, the complement of which is not Maas' 1 but 10 in 22, a ratio probably too small to base a counter argument on it. Neither Bédier nor Maas discuss root unifurcating cases extensively, but they could make a crucial difference in the ratios of interest since including root unifurcating trees, *non-root-bifurcating* would no longer be equivalent to *root multifurcating* in meaning. Thus, Maas' shift of focus from root bifurcating to root multifurcating introduces ambiguity. Responding to such ambiguity, we demonstrated a mathematically sound way of looking both at proportions of root degree patterns with ($R_2(m)$) and without root unifurcations ($\tilde{R}_2(m)$).

Hering (1967), probably aware that root degrees of $k = 1$ appear to be somewhat unrealistic in actually observed stemmas, stated that instead of following Maas' focus, one should rather look at

$$R_{\mathrm{HE}}(m) = \frac{\sum_{k>2} g_k(m)}{g_2(m)}$$

which Hering (1967) investigated until $m = 6$ and for which he speculated that it would probably never surpass 0.33 or lie even lower. Looking at the plot of the proportions, see Figure 3, we can see that Hering was right, the asymptote is however rather 0.3. The extraordinary role of root unifurcations is immediately visible, since they are the only k witnessing a decline. This naturally follows from their restrictions—for instance their root can only be labeled, meaning that only the first term in (1) will be relevant, while for all other root degree patterns both add up.

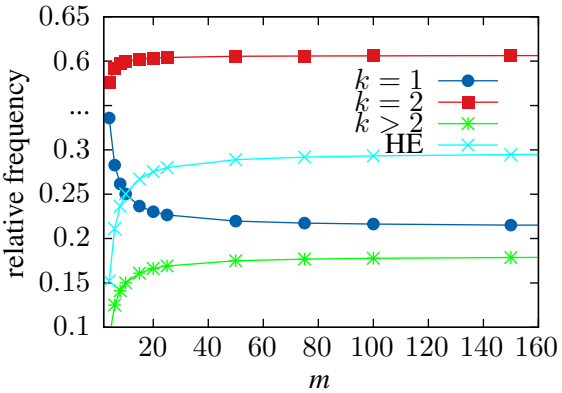

Figure 3: Proportions of root unifurcating and root bifurcating rooted Greg trees among all possible rooted Greg trees for a fixed m as well as $R_{k>2}(m)$ and $R_{\mathrm{HE}}(m)$. Note that the first three proportions add to 1.

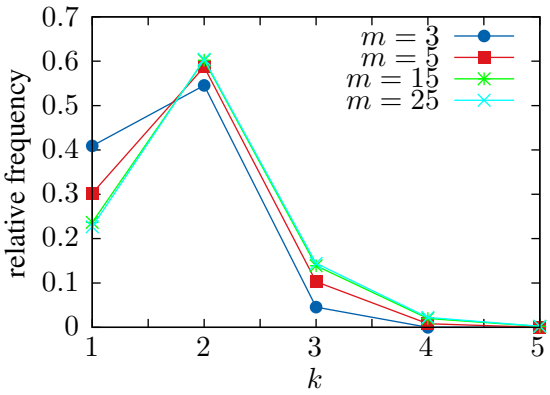

Figure 4: Root degree distribution for trees counted by $g_k(m)$ for fixed $m = 3, 5, 15, 25$.

In order to gain a deeper insight, we are now looking at another type of tree which plays an important role in stemmatology.

4 The Second Type of Manuscript Tree

While Maas had looked at possible trees a philologist can reconstruct, other studies looked at true historical trees and their proportionalities. The underlying process reflected in stemmatological trees is the generation of manuscripts and their copying. There is (in many cases) one original—which we can understand as a root node to a rooted tree—which gets copied a certain number of times (children in first generation). Each manuscript (including root) can be copied a certain number of times again (always including 0 times) and so forth. We assume each node to represent a unique text symbolized through a distinct label. In this way, the copy history can be understood/displayed as a rooted labeled tree. Since copying is a process from a vorlage[8] to a copy, the edges can be understood as directed.

Such a tree depicts the complete copy history of a text—and not as a stemma does, the reconstructible portion of it. It ignores loss of manuscripts (does not assume or know any unlabeled node) and extends to the entire copying history of a text. In order to avoid terminological confusion, the class of trees depicting this complete copy history of a tradition has been called an *arbre réel* in philology, a term coined by Fourquet (1946)—for convenience referred to as *arbre* in the rest of the paper.[9] Arbres were usually used as hypothetical units of argumentation for outlining general scenarios of copying and proliferation in philological discourse, see for instance Castellani (1957). However, recently, they have gained actuality through artificial traditions, that is, complete copied sets with known ground truth (Spencer et al., 2004; Baret et al., 2006; Roos and Heikkilä, 2009; Hoenen, 2015), where arbres are used for evaluation, comparing them to computationally reconstructed stemmata.

In the following, we are looking at arbres themselves and provide an answer to the question how prevalent root bifurcation is in arbres. This may be useful for future research on the general effects of loss induced tree transformations (turning an arbre into a stemma), as has been exemplarily done for a restricted set of topologies by Trovato and Guidi (2004). Greg (1927) had already hypothesized that deformations arbres undergo through historical manuscript loss may be a reason for expectable root bifurcations in stemmata.[10]

We note that the following is a special case of our already derived results. In other words, we now evaluate $g_k(m, 0)$, in our above notation. However, this special case admits simpler closed-form formulas as well as a derivation of the asymptotic degree distribution.

5 Rooted Labeled Trees

By Cayley's formula (Cayley, 1889), the number T'_m of labeled trees on m nodes is given by m^{m-2}. The number T_m of rooted labeled trees is then given by m^{m-1} since each of the m nodes can be the root. Now, let's assume that the root has degree $k = 1, \ldots, m-1$. How many such trees are there, $T_{m,k}$?

To answer this, we invoke our meta formula, Formula (1), with the following specification of $F(\mathbf{s}, \mathbf{p})$:

$$F(\mathbf{s}, \mathbf{p}) = g(s_1, 0) \cdots g(s_k, 0)$$

since $\mathbf{p} = (0, \ldots, 0)$, as we have no unlabeled nodes in this case. We have $g(s, 0) = T_s$ since $g(s, 0)$ retrieves the number of rooted labeled trees with s nodes.

[8] Vorlage is a loaned term for *original of a copy, not of a tradition* deriving from German used in philology.

[9] Although in French terminology the same term is used for so-called R-trees, there is no conceptual overlap whatsoever.

[10] The kind of stemma we are talking about here is not a reconstructed stemma for any number of surviving manuscripts but rather the one single "true" stemma or *stemma reale* as termed by Timpanaro (2005).

Thus, combining this insight with the formula of Cayley, we find that there are exactly

$$\frac{m}{k!} \sum_{\mathbf{s} \in \mathcal{C}(m-1,k)} \binom{m-1}{\mathbf{s}} s_1^{s_1-1} \cdots s_k^{s_k-1} \qquad (2)$$

rooted labeled trees on m nodes with root degree k, where we let $\mathcal{C}(m-1,k)$ stand for $\mathcal{C}((m-1,0),k)$. An alternative, simpler formula for $T_{m,k}$ is given by:

$$T_{m,k} = m \cdot \binom{m-2}{k-1} \cdot (m-1)^{m-1-k}. \qquad (3)$$

For $k = 1$ this formula is not difficult to show. For $k = 2$ it has the following combinatorial interpretation. A rooted labeled tree has a root, for which we may choose any of the m nodes. Then there are $(m-1)$ vertices left. There are $(m-1)^{m-3}$ possible labeled trees on them. Since the $(m-1)$ vertices form a tree, there are $(m-2)$ edges connecting them. We may take any of these, and replace it by connections of their endpoints to the root. This yields all the rooted labeled trees in which the root node has degree 2. For $k > 2$ a similar, but more involved argument applies (Moon, 1970, Theorem 3.2).

Next, we ask for the probability $P_m[k]$ that a randomly chosen rooted labeled tree from $\mathcal{T}_m$ has root degree $k = 1, 2, \ldots$. We find

$$P_m[k] = \frac{T_{m,k}}{T_m} = \frac{\binom{m-2}{k-1}}{(m-1)^{k-1}} \cdot \left(\frac{m-1}{m}\right)^{m-2}. \qquad (4)$$

The second factor in this product equals $(1 - \frac{1}{m})^{m-2}$ and thus converges to $\exp(-1)$ as $m \to \infty$. For the first factor $A = \frac{\binom{m-2}{k-1}}{(m-1)^{k-1}}$, we find

- for $k = 1$: $A = 1 \longrightarrow 1$,
- for $k = 2$: $A = \frac{(m-2)}{(m-1)} \longrightarrow 1$,
- for $k = 3$: $A = \frac{(m-2)(m-3)}{2} \frac{1}{(m-1)(m-1)} \longrightarrow \frac{1}{2}$

as $m \to \infty$. In general, we have for A:

$$A = \frac{(m-2)(m-3)\cdots(m-k)}{(m-1)(m-1)\cdots(m-1)} \frac{1}{(k-1)!}$$

When k is fixed and $m \to \infty$, then this converges to $\frac{1}{(k-1)!}$. Hence, the asymptotic distribution $P[k]$ of $P_m[k]$ is

$$P[k] = \frac{\exp(-1)}{(k-1)!}$$

which is a Poisson distribution with parameter $\lambda = 1$, denoted as Poisson(λ).

Figure 5 compares the asymptotic Poisson $P[k]$ distribution to the actual finite distributions $P_m[k]$. We see that convergence is rapid. For $m = 40$, $P_m[k]$ is visually already extremely close to Poisson($\lambda = 1$).

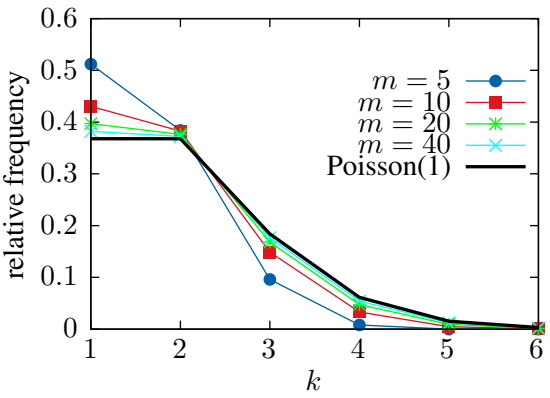

Figure 5: Asymptotic distribution Poisson($\lambda = 1$) and finite distributions $P_m[k]$ for $m = 5, 10, 20, 40$.

From $P[k]$, we infer that root bifurcations are asymptotically twice as likely as trifurcations but exactly as likely as unifurcations, and have a probability of roughly 0.37. Moreover, the larger k gets, the smaller the probability of root k-furcating trees—and this probability is rapidly decaying in k. As a side note, we emphasize that the asymptotic probability for bifurcations has a particularly beautiful mathematical form, namely, the inverse of Leonhard Euler's constant e.

These mathematical derivations, if they are based on a plausible description of reality, suggest that in history many original manuscripts may have been copied only once, the same number has been copied twice, half as many three times and a third of that number four times, a fourth of that number (for four) five times and so on. That is, if indeed a random process that selects each arbre for a fixed number of trees on m nodes with equal likelihood is a good model of true copy history. On this, any more sophisticated model can operate.

If root bifurcations are already very numerous, then an immediately related question would be what consequences this could have for a stemma when thinking about the transformations an arbre undergoes through historical loss. To this end, Weitzman (1982; 1987) has shown, and Trovato

and Guidi (2004) come to a similar conclusion, that historically realistic scenarios of loss would imply a large quantity of bifurcations and root bifurcations in stemmata based on transformed arbres. Those do exceed $\frac{1}{e}$ and thus a possible effect of historical loss is to increase the percentage of root bifurcations, in which case $\frac{1}{e}$ would rather operate as a lower bound.

6 Conclusion

We have counted root k-furcating rooted labeled trees and root k-furcating rooted Greg trees. For the former, the asymptotic root degree distribution has been derived mathematically. For the latter, we have provided exact formulas that allow to approximate the asymptotic root degree distribution. From this, we (very strongly) conjecture that root bifurcating Greg trees have an asymptotic probability of above (and close to) 0.606.

In both cases, relating to a model of representation of arbres (true and complete historical manuscript genealogies) and stemmata (reconstructed genealogies from surviving nodes), the proportions of root bifurcating trees for historically relevant tradition sizes is the largest in respect to the other root degrees. Therefore, while previously other studies have shown that root bifurcations are expectable for true stemmata, we enhance this finding to reconstructible stemmata and arbres so that this statement now covers the three philologically relevant general types of trees discussed until today. Concerning stemmata, we have argued that the proportions of root bifurcating stemmata observed in real collections of genealogies is close to what is mathematically predicted, with a seemingly small bias for root bifurcations.

In the philological debate, where numerical arguments have been pursued since the very beginning, the formulas presented here contribute to clarify the basic combinatorial nature of the entities involved in the modeling of manuscript evolution. We believe that in an ever more computational stemmatological endeavour cultivating the mathematical foundations can only have positive effects.

While our findings with respect to root degrees of rooted labeled trees are certainly far from novel to the mathematics community, our formulas for Greg trees, which generalize rooted labeled trees, are, to our best knowledge, original.

References

P. Baret, C. Macé, and P. Robinson (eds.). 2006. Testing methods on an artificially created textual tradition. In *Linguistica Computationale XXIV-XXV*. Instituti Editoriali e Poligrafici Internationali, Pisa-Roma, volume XXIV-XXV, pages 255–281.

J. Bédier. 1890. *Jean Renart: Le lai de l'Ombre*. Saint-Paul. Reprint, New York: Johnson, 1968.

J. Bédier. 1913. *Jean Renart: Le lai de l'Ombre*. Firmin-Didot.

J. Bédier. 1928. La tradition manuscrite du 'Lai de l'Ombre': Réflexions sur l'Art d'Éditer les Anciens Textes. *Romania* 394:161–196, 321–356. (Rpt. Paris: Champion, 1970).

H. D. Cameron. 1987. The upside-down cladogram: problems in manuscript affiliation. In *Biological Metaphor and Cladistic Classification: an Interdiscplinary Approach*, University of Pennsylvania, pages 227–242.

A. E. Castellani. 1957. *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*. Number 20 in Discours universitaires. Éditions universitaires.

A. Cayley. 1889. A theorem on trees. *Quarterly Journal of Mathematics* 23:376–378.

M. A. Covington. 2004. The number of distinct alignments of two strings. *Journal of Quantitative Linguistics* 11(3):173–182. https://doi.org/10.1080/0929617042000314921.

S. Eger. 2015. On the number of many-to-many alignments of multiple sequences. *Journal of Automata, Languages and Combinatorics* 20(1):53–65.

S. Eger. 2017. The combinatorics of weighted vector compositions. *ArXiv preprint* https://arxiv.org/abs/1704.04964.

J. Felsenstein. 1978. The number of evolutionary trees. *Systematic Zoology* 27(1):27–33.

J. Felsenstein. 2004. *Inferring phylogenies*. Sinauer Associates Sunderland.

C. Flight. 1990. How many stemmata? *Manuscripta* 34(2):122–128.

J. Fourquet. 1946. Le paradoxe de Bédier. *Mélanges* 1945(II):1–46.

W. W. Greg. 1927. *The calculus of variants: an essay on textual criticism*. Clarendon Press.

W. W. Greg. 1931. Recent theories of textual criticism. *Modern Philology* 28(4):401–404.

J. Grier. 1989. Lachmann, Bédier and the bipartite stemma: towards a responsible application of the common-error method. *Revue d'histoire des textes* 18(1988):263–278.

J. R. Griggs, P. Hanlon, A. M. Odlyzko, and M. S. Waterman. 1990. On the number of alignments of k sequences. *Graph. Comb.* 6(2):133–146. https://doi.org/10.1007/BF01787724.

O. E. Haugen. 2002. The Spirit of Lachmann, the Spirit of Bédier: Old Norse Textual Editing in the Electronic Age. In *Annual Meeting of The Viking Society, University College London*. volume 8.

O. E. Haugen. 2010. Is stemmatology inherently dichotomous? On the silva portentosa of Old Norse stemmata. *Studia Stemmatologica* .

O. E. Haugen. 2015. The silva portentosa of stemmatology Bifurcation in the recension of Old Norse manuscripts. *Digital Scholarship in the Humanities* 30(2).

W. Hering. 1967. Zweispaltige Stemmata. *Philologus-Zeitschrift für antike Literatur und ihre Rezeption* 111(1-2):170–185.

A. Hoenen. 2015. Das artifizielle Manuskriptkorpus TASCFE. In *DHd 2015 - Von Daten zu Erkenntnissen - Book of abstracts*, DHd. http://gams.uni-graz.at/o:dhd2015.abstracts-gesamt.

A. Hoenen. 2016. Silva Portentosissima Computer-Assisted Reflections on Bifurcativity in Stemmas. In *Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University*. pages 557–560. http://dh2016.adho.org/abstracts/311.

D. H. Huson. 1998. Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics* 14(1):68–73.

M. Josuat-Vergès. 2015. Derivatives of the tree function. *The Ramanujan Journal* 38(1):1–15.

A. Kleinlogel. 1968. Das Stemmaproblem. *Philologus-Zeitschrift für antike Literatur und ihre Rezeption* 112(1-2):63–82.

D. E. Knuth. 2005. The art of computer programming, volume 4: Generating all combinations and partitions, fascicle 3.

P. Maas. 1937. Leitfehler und Stemmatische Typen. *Byzantinische Zeitschrift* 37(2):289–294.

P. Maas. 1958. *Textual Criticism*. Clarendon Press.

J. W. Moon. 1970. *Counting labelled trees*. Canadian Mathematical Congress.

R. J. O'Hara. 1996. Trees of history in systematics and philology. *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano* 27(1):81–88.

G. Pasquali and D. Pieraccioni. 1952. *Storia della tradizione e critica del testo*. Le Monnier.

N. I. Platnick and H. D. Cameron. 1977. Cladistic methods in textual, linguistic, and phylogenetic analysis. *Systematic Zoology* 26(4):380–385.

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

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

D. L. Swofford. 1990. *PAUP: Phylogenetic Analysis Using Parsimony Version 3.0, May 1990*. Illinois Natural History Survey.

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

P. Trovato and V. Guidi. 2004. Sugli stemmi bipartiti - decimazione, asimmetria e calcolo delle probabilità. *Filologia Italiana* 1:9–48.

M. P. Weitzman. 1982. Computer simulation of the development of manuscript traditions. *ALLC Bulletin. Association for Library and Linguistic Computing Bangor* 10(2):55–59.

M. P. Weitzman. 1987. The evolution of manuscript traditions. *Journal of the Royal Statistical Society. Series A (General)* pages 287–308.

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

F. Whitehead and C. E. Pickford. 1951. The two-branch stemma. *Bulletin Bibliographique de la Société Internationale Arthurienne* 3:83–90.