

1 Analysis of Variation Significance in

2 Artificial Traditions Using Stemmaweb

3 *Tara L Andrews, Universität Bern¹*

4
5 The role of the scholar's intuition in textual scholarship is a subject that has occasioned
6 impassioned debate at times over the last century or more. Is textual criticism a science,
7 or an art—should it be pursued with methodical rigor or with intellectual inspiration?
8 Nowhere is this conflict more pointed than in the sub-field of text stemmatology. While
9 nearly all textual scholars agree that, particularly in the era before the printing press,
10 texts were copied and changed in both intentional and unintentional ways, not all of
11 them admit both the possibility and the utility of deriving a stemma of its transmission.
12 Those who would do so, either for the purposes of text reconstruction or simply to study
13 its history, must align themselves on an ideological spectrum that ranges from the
14 superiority of human intellect and judgment represented by the method of Lachmann, to
15 the wholehearted embrace of empirics and statistics represented by phylogenetic
16 methods.

17 Since the nineteenth century, the process of stemma construction has been more
18 or less codified and methodical. For all the formalization it has undergone, however, at
19 the core of stemmatics there still lies the question of what role, precisely, philological
20 judgment should play. While modern computational methods allow philologists to delay

¹ Email: firstname.lastname@kps.unibe.ch

21 judgment until most of the analysis is done (in the case of neo-Lachmannian binary tree
22 construction) or even to suspend it altogether (in the case of purely phylogenetic trees
23 presented as stemmata), there has been little assessment of the positive difference that
24 philological intuition makes to the recovery of the transmission history of a text.

25 Here we report on an experiment designed to assess the weight that can be given
26 to philological judgment in three cases, all artificial traditions in which the true stemma
27 of the text is known. We shall give an overview of each of these traditions, discuss the
28 methods and tools used for experimentation, examine the results that were obtained,
29 and draw some general conclusions. ²

30 **Background**

31 In his recent study of the development of humanistic method, Rens Bod (Bod, 2013)
32 writes approvingly that ‘stemmatic philology appears to be the only humanities
33 discipline to have become a “normal science”’. This statement might come as something
34 of a surprise to stemmatologists, many of whom are embroiled in an on-going conflict
35 between the desire for empiricism and falsifiability in stemmatic method on the one
36 hand, and the belief on the other hand that mechanical process simply cannot replace
37 human intuition as a means to divine the ‘signal’ in textual variation from the ‘noise’.

38 The history of textual criticism since roughly the time of Lachmann can certainly
39 be understood as a story of attempts to create Bod’s “normal science”—to formalize and
40 generalize the restoration of a text into something approaching a scientific method (e.g.
41 Greg, 1927)— and reactions against these attempts by scholars who believed that no
42 mechanistic approach could ever rival the work produced by the intuition that a genuine

² I am very grateful to the reviewers of this article for their numerous helpful comments, and in particular to Matthew Spencer for his suggestions concerning statistical analysis of the results. Their feedback has vastly improved this paper.

43 master of textual scholarship should possess (e.g. Housman, 1921) or indeed who
44 believed that stemmatic methods tend to produce specious nonsense (e.g. Bédier, 1928).
45 The middle ground after over a century of these debates is perhaps stated most
46 succinctly by West (1973), who explains how a stemma should be created:
47 The investigator will not put off the question of the interrelationships of the manuscripts till he
48 has finished collating them: he will be considering it while he collates them, forming and
49 modifying hypotheses all the time. This will not only make the work considerably more
50 interesting to do (which will make him more alert and accurate while doing it), it will also
51 shorten it, as will be explained presently.

52 As the use of cladistic and other phylogenetic methods accelerated in the last
53 decades of the twentieth century, and as software for automatic collation began to be
54 available, the prevailing attitude changed again: many scholars today (Andrews, 2012a;
55 Robinson, 2004; Wattel, 2004) have advocated best-practice methods in which the
56 collation is produced before any analytical judgment is made concerning the
57 relationships between the texts, on the basis of all available textual information, with as
58 little human interference as possible (although opinion remains divided as to whether
59 the collations should be normalized for orthography, punctuation, and so forth.) Only
60 when the collation is finished should the analysis begin. This attitude is itself represents
61 a shift in textual criticism back in the direction of 'science' from 'art', insofar as
62 interpretation is separated from that which can be done in a mechanical way with
63 reasonable and undisputed accuracy. Even so, while some scholars have wholeheartedly
64 embraced cladistics to such a degree that they no longer attempt even the orientation of
65 a phylogenetic tree into a more traditional stemma, most others prefer a 'happy
66 marriage of our human philological judgment with the computing power of our
67 algorithm' (Roelli and Bachmann, 2010). Cladistic methods do not make any inherent

68 distinction or judgment concerning the significance of a variant; while arbitrary
69 weightings can certainly be supplied by scholars to be used in the algorithm (Howe,
70 Connolly, and Windram, 2012), at present these weightings tend to arise from
71 philological judgment rather than any computable property of the text.

72 Rather than simply the increasing separability of *collatio* and *recensio*, however,
73 Bod seems to draw his impression of stemmatology-as-a-science from multiple studies
74 that appeared in the late 1990s and early 2000s (e.g. Salemans, 1996, 2000; Schøsler,
75 2004; Smelik, 2004) in which attempts were made to derive formal categories of text
76 variation and assign relative text-genealogical weights to different categories.

77 The most well-known of these is the work of Salemans (2000), who proposed a
78 strict set of formal guidelines for the categorization of textual variation and the selection
79 of those variants that should be deemed ‘text-genealogical’, that is, significant enough to
80 form the basis for construction of a text-stemmatic tree. Salemans is straightforward
81 about how he constructed these guidelines. Some of them are drawn from his own
82 philological intuition, informed by the common wisdom of philologists who came before
83 him, for identifying those sorts of variants that are unlikely to occur by chance; others,
84 which appear more strangely restrictive, are meant to ensure that the algorithm he uses
85 can draw up a neat binary tree, as free of contradiction as possible. A few examples of
86 these rules are listed here:

- 87 • A place of variation in the text occurs where there are two or more
88 ‘competing’ readings of the text, while the surrounding readings agree in
89 all text versions; these places should be as small as possible.
- 90 • A place of variation suitable for the construction of a stemma is one that
91 contains exactly two competing variants, each attested by at least two
92 witnesses.

- 93 • Reordering of words (assuming the reordering is grammatically correct)
94 may be used as a text-genealogical variation, so long as there are at least
95 three words being reordered, none of which are adverbs.
- 96 • Nouns and verbs are the most suitable types of readings for creation of a
97 stemma.

98 The primary concern of Salemans was to exclude the possibility (so far as it can
99 be done) that the scholar might compromise his or her stemma by inadvertently
100 assigning text-genealogical significance to a variant that in fact arose coincidentally in
101 parallel in unrelated manuscripts; in order to avoid this possibility, the method tends to
102 discard the vast majority of observed variation from consideration.

103 Cautious as it is, does the method of Salemans work? He used it to produce a
104 plausible stemma for the text of *Lanseloet van Denemerken*, but as Salemans himself
105 affirms in a long discussion of the merits of deductive reasoning, he has used his own
106 textual intuition and prejudices to build up a set of rules for avoiding those very textual
107 prejudices. As Schmid (2004) points out, this has produced a result that conforms very
108 nicely to the intuition by which it is shaped. It is an interesting deductive experiment but
109 there is little in the way of falsifiability in the result.

110 In the same article, Schmid observes that Salemans ‘certainly pinned down [the
111 types of variant readings] that are predominantly suspect of accidental variation’. In
112 other words, Salemans has done an excellent job of codifying the shared philological
113 common wisdom of his time; he has not provided additional evidence that the common
114 wisdom is actually justified. Schmid goes on to demonstrate not only that ‘suspected
115 accidental’ variation is not always coincidental, but also that variation that ought to be
116 safely genealogical by the standard of Salemans is not necessarily so! This has called into
117 sharp question the reliability of philological common sense in the first place.

118 Schmid's findings on the potential significance of 'insignificant' variance have
119 been corroborated elsewhere (Blake and Thaisen, 2004; Spencer, Mooney, et al., 2004);
120 it is clear that, if we discount these entirely, we are losing potentially valuable
121 information. What has not so far been tested in any real way is the philological judgment
122 that is at the heart of all the classification systems that have been proposed.

123 Between 2010 and 2012 a computational object model was developed,
124 implemented as a Perl library, to represent a given tradition together with the variation
125 in its witnesses as an interlinked graph; a companion model was developed, again based
126 conceptually on a graph, to represent arbitrarily complex manuscript transmission. Use
127 of these models made it possible to perform empirical analysis on a variety of stemmata
128 produced using different methods (Andrews and Macé, 2013). The models also provide
129 the underlying framework for a set of software tools that were used to perform the
130 analysis and subsequently made available to other textual scholars for their own use
131 (Andrews, 2012b). One tool allows the categorization and annotation of the way in
132 which individual variant readings are related, another allows the specification of one or
133 more stemma hypotheses, and a third performs an analysis and cross-correlation of
134 reading variants with their consequences for any of the existing stemma hypotheses.
135 The initial experiments conducted using these tools also corroborated the findings that
136 'insignificant' variation was surprisingly likely to follow text-genealogical transmission
137 patterns in both artificial text traditions and genuine traditions for which reasonable
138 certainty of the stemma can be had; we concluded that the application of syntactically-
139 based categories of the sort that are relatively straightforward to identify automatically
140 using linguistic analysis parsers (e.g. spelling variation, grammatical variants of the
141 same word, variants that involve different words fulfilling the same grammatical

142 function, which were termed ‘lexical’ variants in the tools) does not tend to pick out the
143 sorts of variation that are more or less likely to indicate the copying history of the text.

144 With these tools in place, however, and with a set of texts for which the stemma is
145 known (such as the corpus of artificial text traditions), we can instead attempt a much
146 simpler categorization: to indicate those variants which, in the scholarly judgment of a
147 philologist, are likely to be stemmatically significant. From there we can assess the
148 results: how often was the philologist correct, and how often did the copyist produce an
149 unexpected surprise?

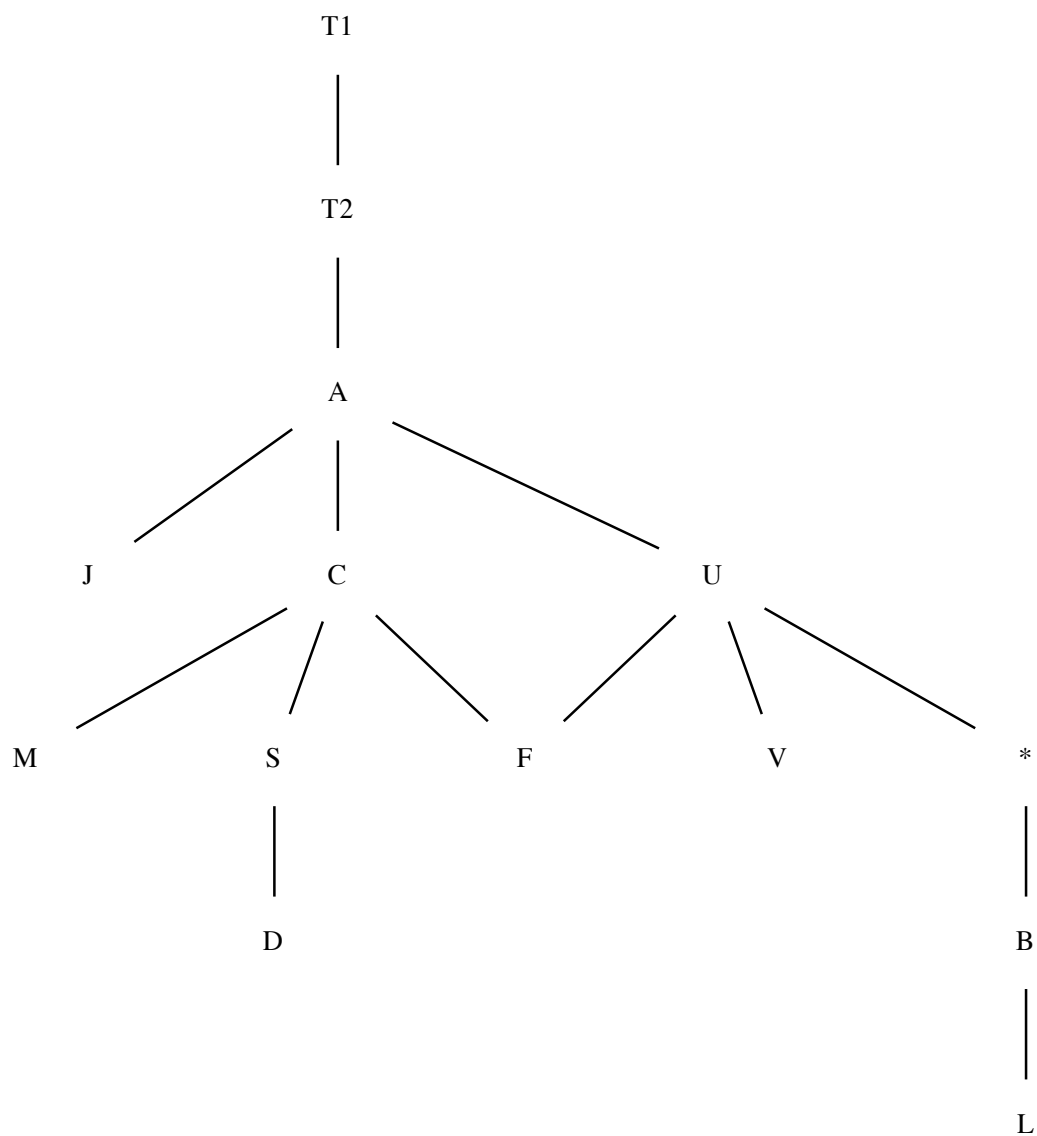
150 **The artificial traditions**

151 In roughly the last decade there have been a number of ‘artificial traditions’ made for the
152 purposes of stemmatological experimentation; these are texts that were copied by
153 volunteers, so that the actual order of transmission is known and a true stemma can be
154 drawn. Three of these were used in the experiment described here.

155 The first is a French translation of a Swedish work, *Notre besoin de consolation est*
156 *impossible à rassasier*. The archetype text, first dictated to a non-native French speaker
157 and then corrected by a native speaker without reference to the printed edition, is 1015
158 words long; it has been made available in 13 copies from 11 different hands (see Fig. 1
159 for the stemma). One of the texts was copied both before and after being mutilated; the
160 first of these copies was itself copied before being ‘lost’, and the second used a different
161 exemplar to replace the missing text. This was done to simulate both the loss of texts in
162 a copying history and the phenomena of ‘contamination’ of the stemma.

163 This tradition was created for the comparison of several different methods for
164 computational stemmatology (Baret, Macé, and Robinson, 2006); this experiment is the
165 only one to date for which the results of ‘classical’, non-computational methods of
166 stemma creation were included alongside the computational versions. In the published

167 experiment, one of the two non-computational methods came closest to reproducing the
168 true stemma, although the computational methods (none of which are able to infer the
169 sort of contamination that was present in the true stemma) were assessed on the basis
170 of the raw output of the algorithm, without any interpretative intervention. The authors
171 note that ‘most philologists’ were easily able to observe the shift of exemplar from the
172 collation alone, which suggests that, had the computational methods been subject to
173 interpretation, the outcome may well have been different.



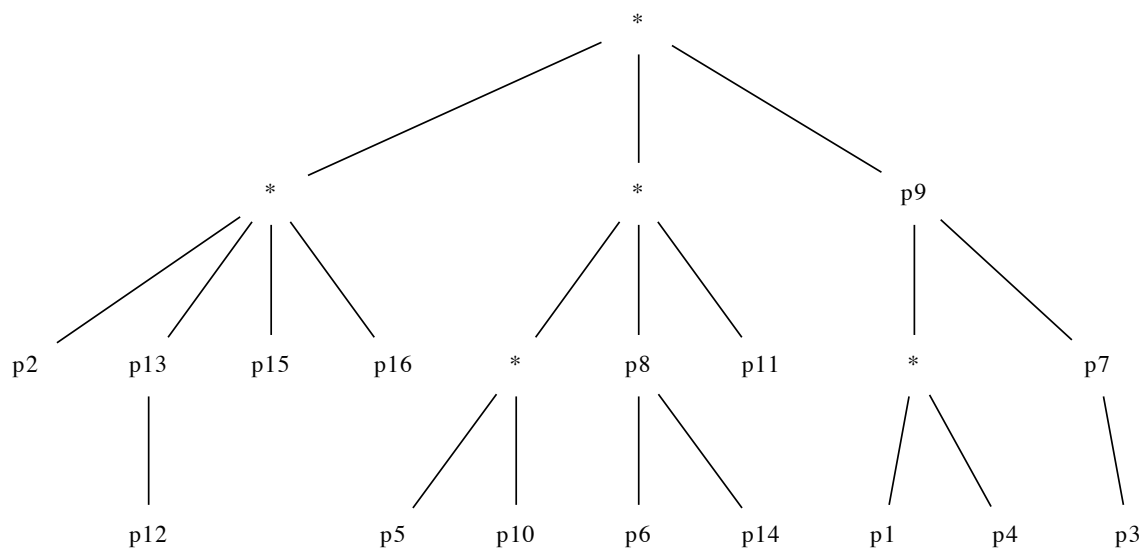
174

175

Fig. 1: Stemma for the *Notre besoin* artificial tradition

176

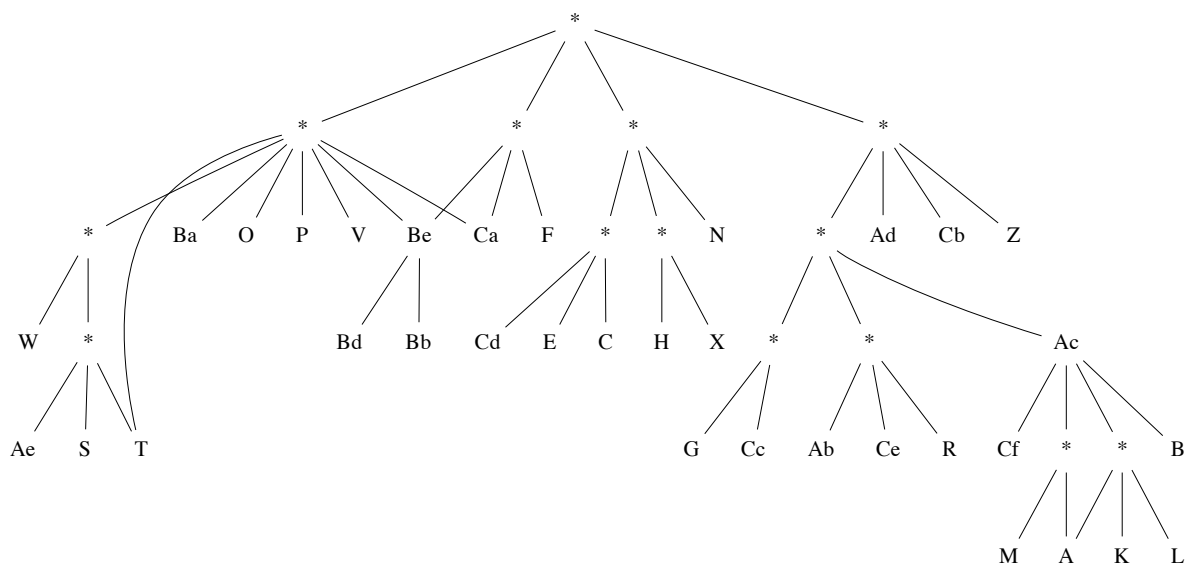
177 The second artificial tradition is an English translation of a portion of the
 178 medieval German epic poem *Parzival*. This text is 834 words long, copied by an
 179 unknown number of volunteer scribes, and is available in 16 versions (see Fig. 2 for the
 180 stemma). Although the text is a little shorter than *Notre besoin*, the somewhat archaic
 181 language gave rise to more frequent variation within copies. The *Parzival* artificial text
 182 was used to test the applicability of phylogenetic methods from evolutionary biology on
 183 textual data (Spencer, Davidson, Barbrook, and Howe, 2004). No attempt to reconstruct
 184 the stemma by hand was reported for this experiment.



185
 186 **Fig. 2: Stemma for the *Parzival* artificial tradition**

187
 188 The third artificial tradition is a text in Old Finnish, *Piispa Henrikin Surmavirsi*.
 189 This text, also known as the “Heinrichi” tradition, is roughly 1200 words long and was
 190 copied by 17 volunteer scribes. 67 copies were made, of which 47 were made available
 191 for analysis (see Fig. 3 for the stemma). The creators of this tradition wished to simulate
 192 medieval copying conditions as far as possible in the modern era; in service to that goal
 193 they chose a text in an archaic language that was only imperfectly known to most of
 194 their scribes (speakers of the modern language), they produced a far larger set of

195 manuscript texts, they had some of the volunteers make two or three copies from
 196 different exemplars, and several of the copies were mutilated after the volunteer work
 197 of copying had finished to simulate damage to manuscripts that tends to occur over
 198 time. This tradition was the primary data set used in a 'computer-assisted stemmatology
 199 challenge' run in 2007 (Roos and Heikkilä, 2009); both the *Notre besoin* and the *Parzival*
 200 artificial traditions were also provided to challenge entrants. No attempt at a stemma
 201 reconstruction by hand of the *Heinrichi* text was reported during the challenge.



202

203 **Fig. 3: Stemma for the available texts of the *Heinrichi* artificial tradition**

204

205 **The experiment**

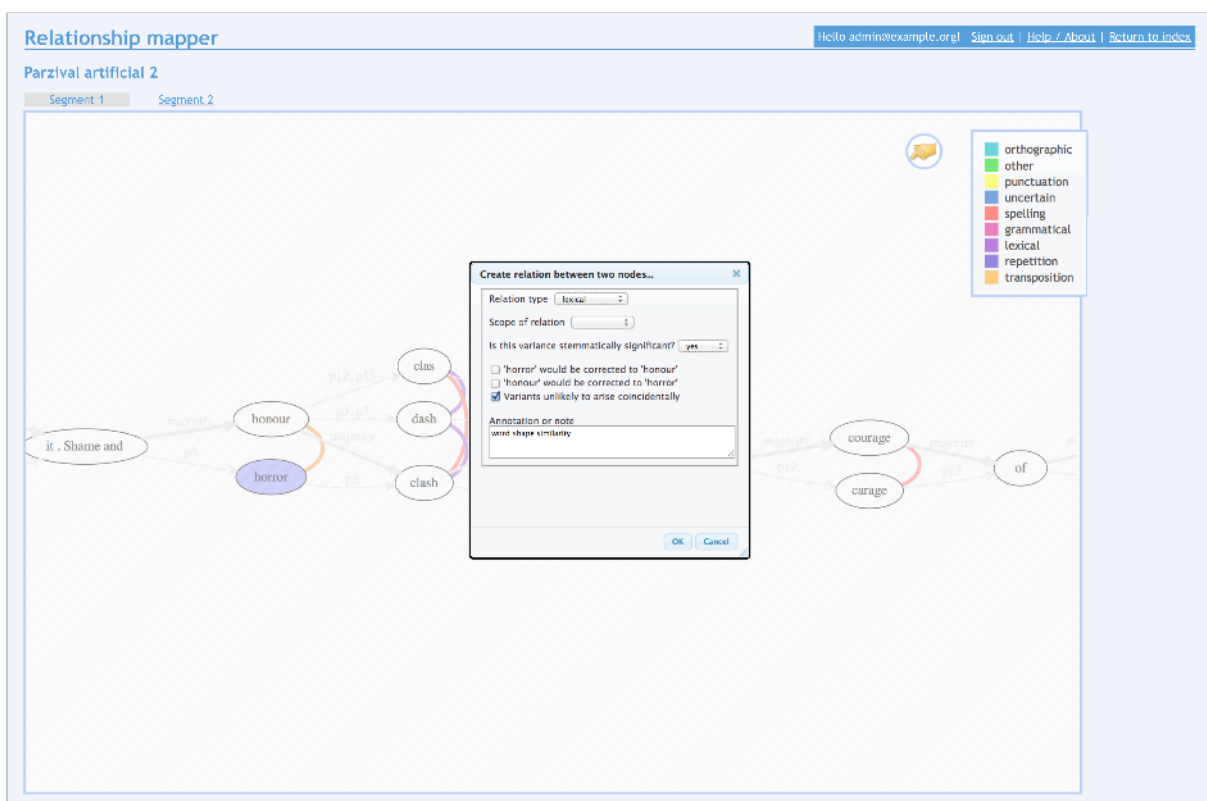
206 For each of the artificial traditions, a volunteer philologist agreed to use the Stemmaweb
 207 software (Andrews, 2012b) to categorize the textual variants according to whether, in
 208 his or her opinion, the variation was stemmatically significant; in the case of the *Parzival*
 209 text, two volunteers were found. The volunteers were chosen both for their experience
 210 in the practice of philological reconstruction of medieval texts and for their native or
 211 near-native familiarity with the language of the text. If there were more than two

212 readings in a variant location, then the determination had to be made for each pair of
213 readings with respect to each other at that location. Since the philologist did not consult
214 the stemma, it was impossible to have any external verification of which reading in a set
215 of variant readings came from the archetype, and which were derivative readings.
216 The premise to be tested is this: a trained philologist should be able to choose variants
217 as ‘significant’ that do, in fact, genealogically follow the true stemma. The converse is not
218 true; the philologist should not be expected to choose with any certainty those variants
219 that positively contradict the stemma; to call a variant ‘insignificant’ merely means that
220 it cannot be relied upon to provide text-genealogical information. A great many so-called
221 ‘insignificant’ variations happen to follow the stemma in all three of the texts.
222 The *Notre besoin* and *Parzival* texts were not normalized in any way; the *Heinrichi* text,
223 due to its sheer size and complexity, was normalized for spelling. Since spelling variation
224 is almost universally considered not to be stemmatically significant, it was felt that this
225 normalization would not harm the philologist’s chances of choosing ‘significant’
226 variation.

227 The Stemmaweb text annotation interface presents the variant texts as a unified
228 ‘variant graph’, in which textual alternatives are represented relative to each other in a
229 continuous presentation of the entire text (c.f. Andrews and Macé, 2013; Dekker, Hulle,
230 Middell, Neyt, and Zundert, 2014; Schmidt and Colomb, 2009). The user may create a
231 relationship between two analogous reading nodes, and define several properties of the
232 relationship (see Fig. 4). In this case the philologist had the option of providing any or all
233 of the following information:

- 234 • How the readings were related syntactically (e.g. whether it was a spelling,
235 grammatical, or some other sort of variation; whether the readings were variant

- 236 grammatical forms; whether they were different words filling the same grammatical
 237 role in the sentence).
- 238 • Whether the variation was significant (possible answers were “yes”, “maybe”, and
 239 “no”).
 - 240 • Whether the variation was unlikely to have occurred coincidentally.
 - 241 • Whether a scribe, upon seeing reading A, might ‘correct’ it to match reading B
 242 without reference to another exemplar (or vice versa).



243
 244 **Fig. 4: Variant classification interface for Stemmaweb: creating a relationship between the parallel**
 245 **readings ‘honour’ and ‘horror’.**

246 There is currently a deficiency in the Stemmaweb software, so that there is no
 247 way to indicate whether a gap (or addition) in the text is stemmatically significant. The
 248 volunteer philologists were made aware of this deficiency at the outset of the
 249 experiment, and each of them was asked to keep a list of which addition/omission
 250 variants might be significant. Two such lists were received, both for the Parzival text; for
 251 the other texts, the philologists working on the texts simply stated guidelines to be

252 applied for these variants. In both cases they advised that they were likely to be
253 significant, unless it was purely a question of easily-replaceable readings such as
254 punctuation.

255 Once annotated, the text variation was compared against the true stemmas for
256 each tradition. For this, the text is subdivided into *variant locations*—these are places in
257 the text where variation occurs, and in terms of the graph a variant location occurs
258 wherever more than one readings occurs at the same rank (that is, the same number of
259 readings distant from the nearest shared prior reading) in the graph. In order to avoid
260 artificially inflating the number of variants, each graph was compressed before analysis,
261 so that individual sequences of readings that did not vary between witnesses, and for
262 which no individual relationships had been made to parallel readings, were treated as a
263 single reading. Three examples of a graph with compression rules applied are given in
264 Fig. 5. In the example marked A, the relationship between βλασφημίας and βλασφημία
265 prevents compression, so that βλασφημία[ς] is treated as one reading, and the omission
266 of ἀπορία in witness Q is treated as a separate reading. In example B, on the other hand,
267 the entire phrase ὡς οὐκ οἶδε is treated as a single omission in witness P(a.c.), and in
268 example C the two words καθαίρει αὐτὸν are treated as a single reading with the
269 alternative καθεαυτὸν in witness S.

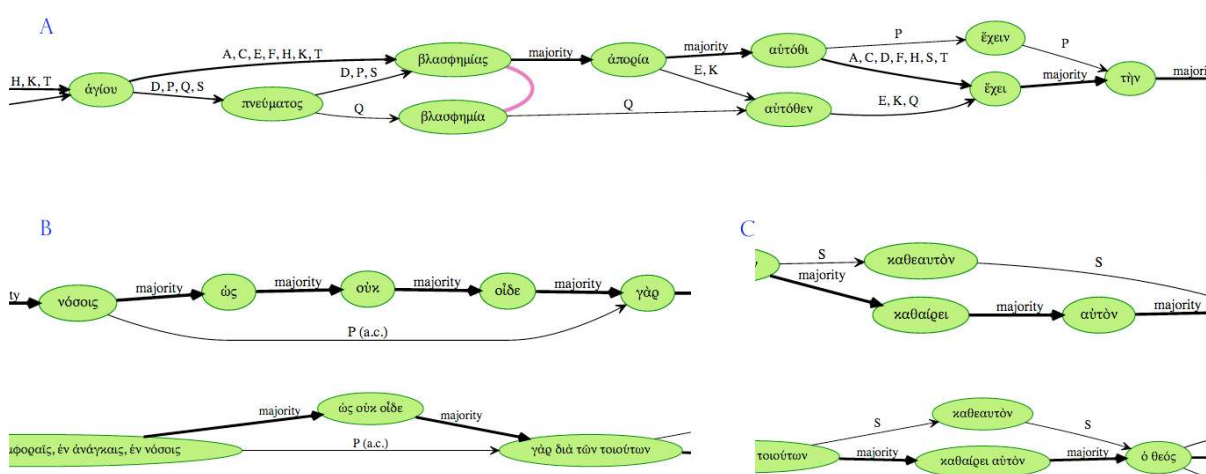


Fig. 5: Examples of reading compression before analysis.

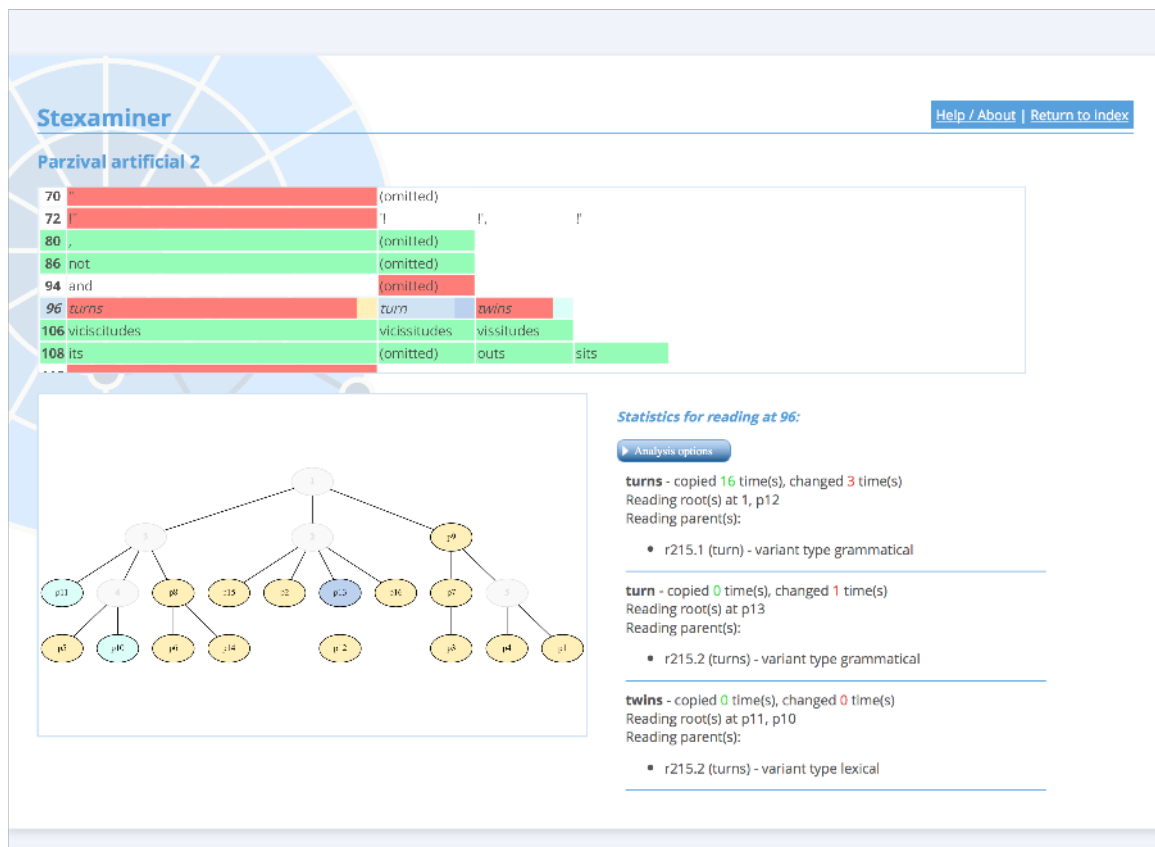
For each distinct variant location within the text, an individual instance of variation was counted when one reading was changed by one or more copyists into a different reading. In the example given in Fig. 6 for a set of non-genealogical variants, the original reading *turns* has been modified two different ways: in witnesses p10 and p11 it became *twins*, and in witness p13 it became *turn*. The reading *turn* itself was modified again, reverting to *turns* in witness p12. Three instances of variation are thus counted: *turns* -> *twins*, *turns* -> *turn*, and *turn* -> *turns*. As a result, coincidental variation is counted as a single instance of variation (*turns* -> *twins*), but the phenomenon of reading reversion, wherein a scribe uses his or her intuition to correct the reading of the exemplar to match an ancestral reading that the scribe did not personally see, is counted as two instances of variation (*turns* -> *turn* and *turn* -> *turns*).

The analysis of variant locations against the stemma is done using a pair of graph calculation programmes that were developed for the purpose (Andrews et al., 2012); the programmes first determine whether the specific occurrence of readings can be explained by genealogical adherence to a given stemma, and then calculate the minimum set of manuscripts (the 'roots') in which each reading could have independently arisen (that is, without having been copied directly from the exemplar.) In the calculation, a

290 particular reading is classified as ‘genealogical’ if and only if there is a single ‘root’ for
 291 the reading in the stemma; for archetypal readings, the ‘root’ will always be the
 292 archetype. No attempt was made to detect potential reading reversions; these were
 293 treated simply as separate variants.

294 Since the philologists were working without reference to a stemma, there are
 295 several pairs of variants that were categorized in the interface but did not occur in the
 296 final analysis, because there was no instance of variation between the readings that
 297 formed the pair. In our example above, any categorization of the pair *turn* – *twins* would
 298 be thus disregarded, although the philologist may well have expressed an opinion,
 299 because according to the stemma no copyist read ‘turn’ and wrote ‘twins’ or vice versa.

300



301

302 **Fig. 6: Analysis of a variant location in *Parzival*. Three instances of variation are recorded: turns ->**
 303 **turn by witness p13, turn -> turns by witness p12, and turns -> twins by witnesses p11 and p10.**

304

305 Results

306 How, then, did our scholarly intuition fare? Taking into account the difficulty with
307 recording significance of addition/omission variants, the traditions were analysed
308 according to three different scenarios:

- 309 1. Addition/omission variants were excluded from the analysis.
- 310 2. Additions were treated as significant unless the added readings were punctuation-
311 only, in which case they were treated as insignificant. Deletions were treated as
312 possibly-significant, unless they were punctuation-only. In the case of the *Parzival*
313 text, the addition/deletion significance information that was provided directly by the
314 philologist was used instead.
- 315 3. Additions were treated as significant (except for the *Parzival* text), and deletions
316 were excluded from analysis.

317

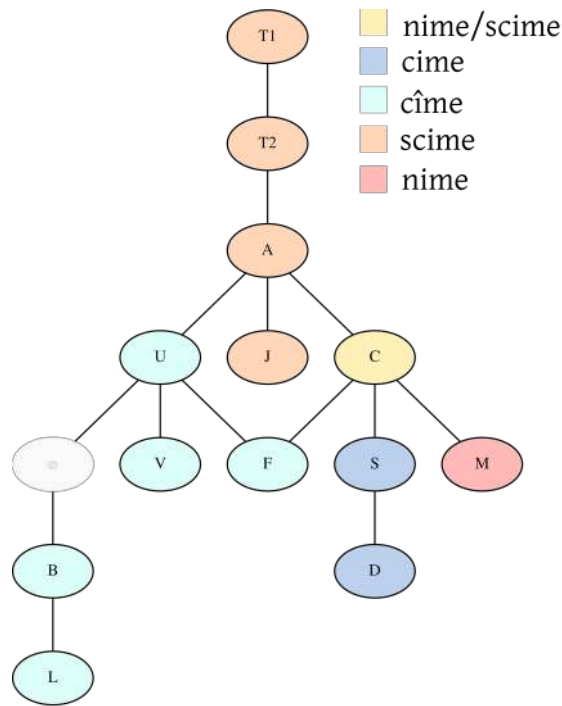
318 As well as the question of additions and deletions, there was the question of
319 orthographic normalization of the text. Due to the sheer size of the *Heinrichi* tradition,
320 the text was normalized for spelling and punctuation before the experiment began; the
321 other two traditions were not normalized beforehand. In order to provide an adequate
322 basis for comparison, the analysis for these two texts was run both with and without
323 normalization in the relevant scenarios.

324 Table 1 shows the aggregate results. For each text (normalized or not) in each
325 scenario, the number of total variants assigned to each of the significance values “yes”,
326 “maybe”, and “no” is given, as well as the number of variants in each category that were
327 found to follow the stemma in a genealogical fashion. Reading the table, for instance, we
328 can see that within the non-normalized *Parzival* tradition there were 211 variants in

329 total, of which 20 were deemed significant and 51 were deemed potentially-significant.
330 13/20 (65%) of the readings deemed significant were in fact genealogical according to
331 the stemma; 31/51 (60.8%) of the readings deemed potentially-significant were
332 genealogical.

333 A list of those variants marked significant for each text is given in Tables 2–5. We
334 have omitted additions and deletions from the list, as well as “type-1” variation—this is
335 a term for variant locations in which only a single manuscript, copied by no others,
336 differed from the rest in its reading. For each relationship link the exemplar and copy
337 reading is listed, along with whether the variation conforms genealogically to the
338 stemma or is an instance of parallel/coincidental variation.

339 There was a somewhat surprising situation to be found within the *Notre besoin*
340 data—when the text was normalized, the number of variants counted went up and the
341 accuracy went down. This was due to the set of readings at rank 47 in the graph (see Fig.
342 7): the potential variants included the words “nime”, “cime”, “cîme”, “scime”, and an
343 illegible word that was either “nime” or “scime”. If the two readings “cime” and “cîme”
344 were treated as separate variants, then the variants could be arranged genealogically on
345 the stemma so that each spelling arose from the reading in witness C; if, however, they
346 were treated as spelling variants of the same word, then it was a parallel variation, in
347 which witnesses U and S independently read ‘cime’ from their exemplars (A and C
348 respectively)! This was an interesting specific counter-example to the prevailing wisdom
349 that texts should be normalized for orthography before analysis.



350

351

Fig. 7: A variant location that is genealogical only before normalization

352

353

Table 1: Aggregate results of variant analysis for the three texts

354

Including addition/deletion assumptions

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

	Parzival 1	Parzival 1 normalized	Parzival 2	Parzival 2 normalized	Notre besoin	Notre besoin normalized	Heinrichi normalized
Total yes	20	19	10	10	22	23	194
Total maybe	51	43	19	17	22	20	420
Total no	140	73	185	107	74	43	749
Genealogical yes	13	12	6	6	16	16	103
Genealogical maybe	31	24	7	6	16	14	115
Genealogical no	73	34	103	54	55	32	382

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

Excluding addition/deletion assumptions

	Parzival 1	Parzival 1 normalized	Parzival 2	Parzival 2 normalized	Notre besoin	Notre besoin normalized	Heinrichi normalized
Total yes	13	12	9	9	20	21	83
Total maybe	32	27	10	8	16	14	0
Total no	98	32	123	50	43	14	557
Genealogical yes	9	8	5	5	15	15	68
Genealogical maybe	18	14	6	5	11	9	0

Genealogical no	59	21	74	30	31	10	371
-----------------	----	----	----	----	----	----	-----

357

358

Excluding only deletion assumptions

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

Table 2: List of significant variants in *Notre besoin* (excluding addition/deletion)

Text position	Genealogical?	Exemplar reading	Copy reading	Note
7	yes	Je n'ai	Jai	
24	yes	minspire	m'inspirent	
24	no	m'inspirent	minspire	Reverted reading
47	yes	nime or scime	cime	
51	yes	arche	arc	
56	yes	abandées	à bander	
68	yes	au deu dieu	odieux	
102	yes	avides	arides	
102	no	arides	avides	Reverted reading
107	yes	la cèse	l'ascèse	
117	yes	Perds	Prends	
121	no	joie	jour	Reverted reading
121	yes	jour	joie	
135	no	du	au	
135	no	au	du	
146	yes	tout	tour	
148	yes	coup	tour	
205	yes	des	pour	
215	yes	être humain	lézard	
217	yes	lézard	être humain	

Table 3: List of significant variants in *Parzival 1*

Text position	Genealogical?	Exemplar reading	Copy reading	Note
9	yes	rue	use	
9	yes	rue	see	

12	yes	clash	dash	
13	yes	where	with	
45	yes	hare	horse	
53	no	reveal	several	
71	yes	Oh	OK	
124	yes	Its	His	
205	yes	rate	note	
205	no	note	rate	
343	no	odd	old	
403	no	cum	and	
403	yes	cum	over	

365

366

Table 4: List of significant variants in *Parzival 2*

Text position	Genealogical?	Exemplar reading	Copy reading	Note
6	yes	heart	heat	
6	no	heat	heart	Reverted reading
9	yes	rue	see	
45	yes	hare	horse	
53	no	reveal	several	
176	yes	is	in	
205	yes	rate	note	
205	no	note	rate	Reverted reading
407	no	cum	and	

367

368

Table 5: List of significant variants in *Heinrichi*

Text position	Genealogical?	Exemplar reading	Copy reading	Note
247	no	wainen	nainen	
303	yes	carcot	carkuhun	
304	yes	gongarita	gangista	
304	yes	gongarita	gangistu	
304	yes	gongarita	amvanta	
463	yes	suin	nin	
471	yes	paljon	tuhansia	
477	yes	enämbi	erämki	
506	yes	cotiani	cariani	
508	yes	pane	pahe	
511	no	ohjat	olijat	
512	yes	suoniset	puaniset	
517	yes	harman	harwan	
522	yes	orhilda	ahtialda	
524	yes	iduilta	iavialta	
526	yes	lihainen	likainen	
531	yes	luocka	kuokka	
533	yes	harjallen	haijuillen	
534	yes	hyvän	kywän	
534	yes	hyvän	luocka	
540	no	aiella	siellä	
545	yes	wiritti	wintti	
547	yes	juoxemahan	juotemahan	
551	yes	laulajtta	kaulojtta	

551	yes	laulajtta	laukijitta	
556	yes	wirguttamahan	weigottamahan	
556	yes	wirguttamahan	wingottumahan	
559	yes	rauta-cahlehisä	routa-cahlehisä	
561	yes	rautainen	rantainen	
561	yes	rautainen	tauroinen	
562	yes	kukersi	kaukan	
567	yes	walcoinen	waleoinen	
570	no	fildin	tildin	
573	yes	njn	siju	
577	yes	wandi	wanki	
577	yes	wandi	waneli	
600	yes	takoa	kackoa	
600	yes	takoa	tokra	
625	yes	pannahinen	lallinlainen	
631	yes	kiukahalda	luikahalda	
631	yes	kiukahalda	kirikahalda	
632	yes	parku	lauleli	
639	yes	wielä	sulle	
640	yes	se	olutta	
641	yes	sun	tarjoapi	
641	no	sun	suu	
645	yes	wielä	wiila	
646	yes	päänsi	leiwän	
646	yes	päänsi	päänni	
647	yes	päristelepi	päällystelepi	
649	yes	sirgotelepi	virgotelepi	
650	yes	heitlelepi	heitlelemi	
655	yes	kijruhti	lähti	
658	yes	Laloi	Lakoi	
659	yes	cuin	ain	
662	yes	walehteli	certoili	
691	yes	heitti	kejtti	
692	yes	tuhkia	luhkia	
696	yes	siwui	silleni	
698	no	lahtarinsa	lahtaunsa	
700	yes	pitkän	pilkän	
707	yes	wuoldu	wuceldu	
722	no	suxen	suten	
726	no	siasta	piasta	
728	yes	sitten	sinen	
728	no	sinen	sitten	Reverted reading
758	yes	wandi	wouti	
760	yes	corkuhujnen	dorkuhujnen	
765	yes	tacoa	tuloa	
765	yes	tacoa	taloa	
774	no	kuhunga	kuhunsa	
775	yes	luuni	luuhi	
775	yes	luuni	kuuni	
776	yes	lendelepi	laudelepi	
778	no	suaneni	suoleni	
790	no	oroin	aroin	
813	yes	Nousiaisten	Pargahisten	
815	no	hieta-cungahan	hieta cangahan	

820	yes	haudattihin	handotti	
847	yes	kewät	kewät [850]	Transposition
854	no	sijne	sijtte	Reverted reading
854	yes	sijtte	sijne	
1111	no	ja	jo	

369

370 In all texts but *Heinrichi*, the philological determination of stemmatic significance
371 fared surprisingly poorly. If human intuition is to be a reasonably reliable and accurate
372 tool for assessing variation, one would expect to see a relatively much higher proportion
373 of text-genealogical variation marked as significant than as potentially-significant; the
374 “maybes” should probably, in turn, be higher again than that not marked as significant at
375 all.

376 How, in this instance, do we define ‘poorly’? One way to examine the data is
377 through use of a chi-square analysis on each of the text scenarios: if our philologists are
378 successful at identifying genealogical variation, we should expect to find that there is a
379 positive correlation between ‘genealogical’ and ‘significant’. If, on the other hand, the
380 philologists are not successful, we will not be able to demonstrate the correlation with
381 any degree of certainty. The chi-square test is not foolproof, both because the amount of
382 variation classed significant is fairly low for most of our texts, and because it may not be
383 safe to assume that each variant is entirely independent of the others in whether or not
384 it is genealogical. It can nevertheless work as a first approximation.

385 Table 6 shows the results of the chi-square analysis across texts and scenarios.
386 The only text to show a strong correlation between ‘significant’ and ‘genealogical’ is
387 *Heinrichi*. In the case where additions and/or deletions are included, however, this
388 extremely strong correlation is highly negative! These are the scenarios where text
389 additions are usually assumed to be significant, and deletions are usually assumed to be
390 in the ‘maybe’ category. If we refer back to the numbers in Table 1, however, we find
391 that 96/186 (51.6%) of significant variants are genealogical, as compared to 382/749

392 (51%) of insignificant variants but only 115/420 (27.3%) of possibly-significant! In this
 393 case, the decision to treat additions and deletions in this categorical manner has had a
 394 disastrous impact on the result.

395 **Table 6: Results of chi-square analysis across all text scenarios**

All variants	χ^2 value	<i>P</i> -value
Parzival 1	1.95	0.38
Parzival 1 normalized	2.06	0.36
Parzival 2	2.60	0.27
Parzival 2 normalized	1.85	0.40
Notre besoin	0.04	0.98
Notre besoin normalized	0.23	0.89
Heinrichi normalized	75.28	0.00

Excl. addition/deletion	χ^2 value	<i>P</i> -value
Parzival 1	0.65	0.72
Parzival 1 normalized	1.39	0.50
Parzival 2	0.07	0.96
Parzival 2 normalized	0.09	0.95
Notre besoin	0.17	0.92
Notre besoin normalized	0.24	0.89
Heinrichi normalized	5.10	0.02

Excl. deletions	χ^2 value	<i>P</i> -value
Parzival 1	1.63	0.44
Parzival 1 normalized	1.83	0.40
Parzival 2	0.16	0.92
Parzival 2 normalized	0.39	0.82
Notre besoin	0.10	0.95
Notre besoin normalized	0.25	0.88
Heinrichi normalized	3.38	0.07

396
 397 Once addition and deletion is excluded, the news for *Heinrichi* is much improved:
 398 we can say with roughly 98% certainty that there is indeed a positive correlation
 399 between ‘genealogical’ and ‘significant’. For the other two texts, the chi-square test
 400 rather spectacularly fails to demonstrate any correlation at all!

401

402 An objection to the chi-square test could be raised here, however: the text that
 403 demonstrated a convincing correlation also happens to be the text for which an order of
 404 magnitude more variation existed to be analyzed. The test is not usually recommended
 405 unless all combinations of category contain at least 10 instances, and that criterion is not
 406 quite met by any of the texts besides *Heinrichi*. In the case of *Parzival 2* in particular, the
 407 philologist has marked relatively few variants as significant at all.

408 We might thus apply a simpler test: to compare the success rates of the
 409 'significant' and 'possibly-significant' categories to the mean success rate of the text as a
 410 whole. We can treat this situation as a binomial distribution (with the same caveat
 411 concerning the independence of genealogical variants), and analyze the 'significant'
 412 group as a sample drawn from the whole. In this case, the successful philologist should
 413 have constructed a sample of 'significant' variants that should have a markedly higher
 414 mean success rate than the wider population of variants. (The same analysis can be
 415 performed on the population of 'possibly-significant' variants, but we would not expect
 416 such a marked difference in the success rate, so we omit that analysis here.) The specific
 417 question we ask is: what is the probability that a random sample of variants would have
 418 at least the same number of genealogical variants as our significant sample?

419 Table 7 shows the results of our binomial distribution. In every case except for
 420 that of *Heinrichi* there were fewer than 15 genealogical variants classed as significant,
 421 the 'plus-four' rule has been applied to the data in order to compensate for the small
 422 sample size (Moore, Craig, and McCabe, 2012).

423 **Table 7: 'Significant' variants treated as samples from a binomial distribution**

All variants	<i>% mean genealogical</i>	<i>% genealogical significant</i>	<i>Likelihood of randomness</i>	<i>Std. deviation</i>
Parzival 1	55.5%	62.5%	18.5%	0.63
Parzival 1 normalized	51.9%	60.9%	14.1%	0.79
Parzival 2	54.2%	57.1%	31.6%	0.19

424

Parzival 2 normalized	49.3%	57.1%	19.6%	0.50
Notre besoin	73.7%	69.2%	62.9%	-0.48
Notre besoin normalized	72.1%	66.7%	67.0%	-0.58
Heinrichi normalized	43.8%	53.1%	0.5%	2.55

Excl. addition/deletion	<i>% mean genealogical</i>	<i>% genealogical significant</i>	<i>Likelihood of randomness</i>	<i>Std. deviation</i>
Parzival 1	60.1%	64.7%	26.8%	0.34
Parzival 1 normalized	60.6%	62.5%	34.6%	0.14
Parzival 2	59.9%	53.8%	57.0%	-0.37
Parzival 2 normalized	59.7%	53.8%	56.6%	-0.36
Notre besoin	72.2%	70.8%	48.0%	-0.13
Notre besoin normalized	69.4%	68.0%	48.5%	-0.14
Heinrichi normalized	68.6%	81.9%	0.4%	2.38

Excl. deletion	<i>% mean genealogical</i>	<i>% genealogical significant</i>	<i>Likelihood of randomness</i>	<i>Std. deviation</i>
Parzival 1	58.6%	68.4%	13.5%	0.77
Parzival 1 normalized	57.8%	66.7%	15.8%	0.67
Parzival 2	58.4%	53.8%	52.7%	-0.28
Parzival 2 normalized	57.0%	53.8%	48.5%	-0.19
Notre besoin	72.0%	69.2%	55.3%	-0.29
Notre besoin normalized	69.4%	66.7%	54.8%	-0.28
Heinrichi normalized	58.9%	53.1%	97.4%	-2.03

425

426 With this analysis, we can see a differentiation of results between the three texts.

427 The results for *Notre besoin* were by far the worst: there was no scenario where the
428 variants treated as significant were more likely than average to be genealogical. Both
429 *Parzival* texts fared slightly better when additions and deletions were taken into
430 account; since these were the two texts for which a positive list of additions and
431 deletions were received, and in light of the overall small sample size, this is not
432 particularly surprising. *Heinrichi* again appears to be the most convincing case of
433 success, when additions and deletions are disregarded; the philologist was correct about
434 82% of the time, as opposed to the 69% that random chance might yield.

435

436 Conclusions

437 What are we to make of these rather surprising results? Above all it is important
438 to bear in mind that the experiment was done using artificial traditions. Particularly for
439 the *Notre besoin* text, many of whose copyists were themselves philologists, there is a
440 real risk that the volunteers consciously or semi-consciously introduced innovations
441 into their copies in order to make the resulting tradition “interesting”. On the other
442 hand, also in the case of *Notre besoin*, at the time of the original experiment a philologist
443 using classical methods was able to reconstruct a stemma that was not very different
444 from the true stemma. Is this a case of one philologist simply being better than the
445 other? While that is possible, it is not tremendously likely; over 70% of *all* variation
446 within *Notre besoin* followed the stemma, which made its reconstruction a
447 comparatively straightforward task no matter what method was used. The results of
448 that experiment bore this out: they showed that every one of the attempted methods,
449 including the computational methods whose results were not manipulated into a
450 ‘normal’ rooted stemma, could correctly identify the main manuscript groupings. That
451 does in itself raise another question: how accurate must we be in choosing significant
452 variation in order to reconstruct an accurate stemma? Although none of the volunteers
453 in this study attempted to draw a stemma, one of the two philologists for *Parzival*
454 provided a set of observations concerning which manuscripts should be grouped
455 together; these were broadly accurate, even though the selection of individual
456 significant variants was often wide of the mark; it is also worth noting that the
457 philologist quite often cited variants as examples of group affinity that were not judged
458 significant!

459 Compared to the rest, the *Heinrichi* artificial tradition fared comparatively well.
460 The overall mean rate of genealogical variation in that text was rather lower than in the

461 other two texts, at just under 44%. The *Heinrichi* corpus includes 2-3 copies per scribe,
462 which increases the possibility of horizontal transmission (particularly for spelling and
463 grammatical idiosyncrasies) in a different way; on the other hand, that tradition appears
464 to have contained many more genuine errors, and the philologist who did the work was
465 accordingly more accurate—leaving aside the question of additions and deletions—in
466 detecting whether variation was significant. The creators of *Heinrichi* seem to have had
467 more success than the others in creating a tradition that is reasonably close to the ‘real-
468 world’ situation of a medieval text widely copied.

469 One substantial conclusion to be found in the data, and one that reinforces
470 findings made previously, is that ‘insignificant’ variation is really not that insignificant at
471 all. We have seen that some philologists prefer to exclude it entirely; others (e.g. Wattel
472 and van Mulken, 1996) include the information but give it as low a weighting as
473 possible. This experiment, together with several others, strongly suggests that our
474 practices for handling this sort of ‘insignificant’ variation are in dire need of revision.

475 A second conclusion concerns the effect of the adoption of blanket
476 generalizations: in this case, the guidelines from two of the philologists for how to
477 handle certain variants. They advised that, “in general”, additions and deletions should
478 be treated in a certain way; when these rules were duly applied in a general fashion, the
479 resulting proportion of “significant” genealogical variation was badly impacted. This
480 aspect of the experiment suggests that we must be extremely careful before adopting
481 any sort of rule-based guideline for the classification of variants, especially if the
482 guidelines are meant to be applied in a regular computational way. It is far too easy to be
483 led blindly into poor results.

484 Finally, this experiment makes clear that stemmatology has some way to go
485 before it can claim the title of a ‘normal science’ that Rens Bod has offered. Our systems

486 of categorization are suspect; our very philological sense of what is or is not significant
487 has not fared as well as we ought to expect in the test against artificial traditions. We
488 have more work to do than Bod's simple 'problem-solving'; we have yet to capture in
489 any formal, demonstrable, or falsifiable way the essence of what scribes were likely to
490 copy and what they were likely to change. If stemmatology is indeed to become a
491 science, this is the next task that needs to be done.

492

493 **References**

494 **Andrews, T. L.** (2012a). The Third Way: Philology and Critical Edition in the
495 Digital Age. *Variants*, 10: 1–16.

496 **Andrews, T. L.** (2012b). *Stemmaweb - a collection of tools for analysis of collated*
497 *texts*. <http://byzantini.st/stemmaweb/> (accessed 18 April 2014).

498 **Andrews, T. L., Blockeel, H., Bogaerts, B., Bruynooghe, M., Denecker, M., De**
499 **Pooter, S., ... Ramon, J.** (2012). Analyzing manuscript traditions using constraint-based
500 data mining. In *CoCoMile 2012 - COmbining COnstraint solving with MIning and LEarning*.
501 Montpellier. http://cocomile.disi.unitn.it/2012/papers/cocomile2012_manuscript.pdf.

502 **Andrews, T. L., and Macé, C.** (2013). Beyond the Tree of Texts: Building an
503 Empirical Model of Scribal Variation through Graph Analysis of Texts and Stemmata.
504 *Literary and Linguistic Computing*, 28(4): 504–21. 10.1093/lc/fqt032.

505 **Baret, P., Macé, C., and Robinson, P.** (2006). Testing Methods on an Artificially
506 Created Textual Tradition. In *The Evolution of Texts: Confronting Stemmatalogical and*
507 *Genetical Methods*. Pisa; Rome: Istituti Editoriali e Poligrafici Internazionali, pp. 255–83.

508 **Bédier, J.** (1928). La tradition manuscrite du Lai de l'Ombre. Réflexions sur l'art
509 d'éditer les anciens textes. *Romania*, 54: 161–96, 321–56.

510 **Blake, N., and Thaisen, J.** (2004). Spelling's Significance for Textual Studies.
511 *Nordic Journal of English Studies*, 3(1): 93–108. (accessed 28 March 2013).

512 **Bod, R.** (2013). *A New History of the Humanities: The Search for Principles and*
513 *Patterns from Antiquity to the Present*. Oxford University Press.

514 **Dekker, R. H., Hulle, D. van, Middell, G., Neyt, V., and Zundert, J. van.** (2014).
515 Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital

516 Manuscript Project. *Literary and Linguistic Computing*, fqu007. 10.1093/llc/fqu007.
517 **Greg, W. W.** (1927). *The calculus of variants: an essay on textual criticism*. Oxford:
518 Clarendon press.

519 **Housman, A. E.** (1921). The Application of Thought to Textual Criticism.
520 *Proceedings of the Classical Association*, 18: 67–84.

521 **Howe, C. J., Connolly, R., and Windram, H. F.** (2012). Responding to Criticisms
522 of Phylogenetic Methods in Stemmatology. *Studies in English Literature 1500-1900*,
523 52(1): 51–67. 10.1353/sel.2012.0008.

524 **Moore, D. S., Craig, B. A., and McCabe, G. P.** (2012). *Introduction to the practice*
525 *of statistics* (7th ed., international ed.). New York: W. H. Freeman.

526 **Robinson, P.** (2004). Making electronic editions and the fascination of what is
527 difficult. *Linguistica Computazionale*, 20–21: 415–38.

528 **Roelli, P., and Bachmann, D.** (2010). Towards Generating a Stemma of
529 Complicated Manuscript Traditions: Petrus Alfonsi’s Dialogus. *Revue d’histoire des textes*,
530 n.s. 5: 307–21.

531 **Roos, T., and Heikkilä, T.** (2009). Evaluating methods for computer-assisted
532 stemmatology using artificial benchmark data sets. *Literary and Linguistic Computing*,
533 24(4): 417–33. 10.1093/llc/fqp002.

534 **Salemans, B. J. P.** (1996). Cladistics or the Resurrection of the Method of
535 Lachmann. In van Reenen, P. T., van Mulken, M., and Dyk, J. W. (eds.), *Studies in*
536 *Stemmatology*. Amsterdam; Philadelphia: Benjamins, pp. 3–70.

537 **Salemans, B. J. P.** (2000). *Building Stemmas with the Computer in a Cladistic, Neo-*
538 *Lachmannian, Way: The Case of Fourteen Text Versions of Lanseloet van Denemerken*.
539 Ph.D. thesis, Katholieke Universiteit Nijmegen.

540 **Schmid, U.** (2004). Genealogy by chance! On the significance of accidental
541 variation (parallelisms). In van Reenen, P. T., den Hollander, A., and van Mulken, M.
542 (eds.), *Studies in Stemmatology II*. Amsterdam: Benjamins, pp. 127–43.

543 **Schmidt, D., and Colomb, R.** (2009). A data structure for representing multi-
544 version texts online. *International Journal of Human-Computer Studies*, 67: 497–514.

545 **Schøsler, L.** (2004). Scribal variations: When are they genealogically relevant—
546 and when are they to be considered as instances of “mouvance”? In van Reenen, P. T.,
547 den Hollander, A., and van Mulken, M. (eds.), *Studies in Stemmatology II*. Amsterdam:
548 Benjamins, pp. 207–26.

549 **Smelik, W. F.** (2004). Trouble in the trees! Variant selection and tree

550 construction illustrated by the texts of Targum Judges. In van Reenen, P. T., den
551 Hollander, A., and van Mulken, M. (eds.), *Studies in Stemmatology II*. Amsterdam:
552 Benjamins, pp. 167–206.

553 **Spencer, M., Davidson, E. A., Barbrook, A. C., and Howe, C. J.** (2004).
554 Phylogenetics of Artificial Manuscripts. *Journal of Theoretical Biology*, 227: 503–11.

555 **Spencer, M., Mooney, L., Barbrook, A., Bordalejo, B., Howe, C. J., and**
556 **Robinson, P.** (2004). The effects of weighting kinds of variants. In van Reenen, P. T., den
557 Hollander, A., and van Mulken, M. (eds.), *Studies in Stemmatology II*. Amsterdam:
558 Benjamins, pp. 227–39.

559 **Wattel, E.** (2004). Constructing initial binary trees in stemmatology. In van
560 Reenen, P. T., den Hollander, A., and van Mulken, M. (eds.), *Studies in Stemmatology II*.
561 Amsterdam: Benjamins, pp. 145–65.

562 **Wattel, E., and van Mulken, M.** (1996). Weighted Formal Support of a Pedigree.
563 In van Reenen, P. T., van Mulken, M., and Dyk, J. W. (eds.), *Studies in Stemmatology*.
564 Amsterdam; Philadelphia: Benjamins, pp. 135–68.

565 **West, M. L.** (1973). *Textual Criticism and Editorial Technique: Applicable to Greek*
566 *and Latin Texts*. Stuttgart: B. G. Teubner.

567