Digital Scholarship in the Humanities Advance Access published December 31, 2014

Analysis of Variation Significance in Artificial Traditions Using Stemmaweb

Tara L. Andrews
Digital Humanities, Universität Bern, Switzerland

Abstract

The role of human philological judgement in textual criticism, and particularly in stemmatics, has been at times hotly debated and in computational stemmatology tends to be carefully circumscribed. In this context philological judgement is deployed to distinguish 'significant' from 'insignificant' textual variation—that is, to select those variants that are more or less likely to betray information about the exemplar from which a given text was copied. This article reports on an experiment performed to assess the accuracy of human philological judgement on a set of three artificial traditions, using tools for stemma analysis developed for a prior project and available to the public as the Stemmaweb online service. We show that for most of the artificial traditions, human judgement was not significantly better than random selection for choosing the variant readings that fit the stemma in a text-genealogical pattern, and we discuss some of the implications of these findings.

Correspondence:
Tara L. Andrews,
Digital Humanities,
Muesmattstrasse 45,
CH-3012 Bern.
E-mail:
firstname.lastname@
kps.unibe.ch

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

Since the nineteenth century, the process of stemma construction has been more or less codified and methodical. For all the formalization it has undergone, however, at the core of stemmatics there still lies the question of what role, precisely, philological judgment should play. While modern computational methods allow philologists to delay judgment until most of the analysis is done (in the case of neo-Lachmannian binary tree construction) or even to suspend it altogether (in the case of purely phylogenetic trees presented as stemmata), there has been little assessment of the positive difference that philological intuition makes to the recovery of the transmission history of a text.

Here we report on an experiment designed to assess the weight that can be given to philological judgment in three cases, all artificial traditions in which the true stemma of the text is known. We shall give an overview of each of these traditions, discuss the methods and tools used for

doi:10.1093/llc/fqu072

experimentation, examine the results that were obtained, and draw some general conclusions.

1 Background

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

The history of textual criticism since roughly the time of Lachmann can certainly be understood as a story of attempts to create Bod's 'normal science'—to formalize and generalize the restoration of a text into something approaching a scientific method (e.g. Greg, 1927)—and reactions against these attempts by scholars who believed that no mechanistic approach could ever rival the work produced by the intuition that a genuine master of textual scholarship should possess (e.g. Housman, 1921) or indeed who believed that stemmatic methods tend to produce specious nonsense (e.g. Bédier, 1928). The middle ground after over a century of these debates is perhaps stated most succinctly by West (1973), who explains how a stemma should be created:

> The investigator will not put off the question of the interrelationships of the manuscripts till he has finished collating them: he will be considering it while he collates them, forming and modifying hypotheses all the time. This will not only make the work considerably more interesting to do (which will make him more alert and accurate while doing it), it will also shorten it, as will be explained presently.

As the use of cladistic and other phylogenetic methods accelerated in the last decades of the twentieth century, and as software for automatic collation began to be available, the prevailing attitude changed again: many scholars today (Robinson, 2004; Wattel, 2004; Andrews, 2012a) have advocated best-practice methods in which the collation is produced before any analytical judgment is made concerning the relationships between the texts, on the basis of all available textual information, with as little human interference as possible (although opinion remains divided as to whether the collations should be normalized for orthography, punctuation, and so forth). Only when the collation is finished should the analysis begin. This attitude itself represents a shift in textual criticism back in the direction of 'science' from 'art', insofar as interpretation is separated from that which can be done in a mechanical way with reasonable and undisputed accuracy. Even so, while some scholars have wholeheartedly embraced cladistics to such a degree that they no longer attempt even the orientation of a phylogenetic tree into a more traditional stemma, most others prefer a 'happy marriage of our human philological judgment with the computing power of our algorithm' (Roelli and Bachmann, 2010). Cladistic methods do not make any inherent distinction or judgment concerning the significance of a variant; while arbitrary weightings can certainly be supplied by scholars to be used in the algorithm (Howe *et al.*, 2012), at present these weightings tend to arise from philological judgment rather than any computable property of the text.

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

The most well-known of these is the work of Salemans (2000), who proposed a strict set of formal guidelines for the categorization of textual variation and the selection of those variants that should be deemed 'text-genealogical', that is, significant enough to form the basis for construction of a text-stemmatic tree. Salemans is straightforward about how he constructed these guidelines. Some of them are drawn from his own philological

intuition, informed by the common wisdom of philologists who came before him, for identifying those sorts of variants that are unlikely to occur by chance; others, which appear more strangely restrictive, are meant to ensure that the algorithm he uses can draw up a neat binary tree, as free of contradiction as possible. A few examples of these rules are listed here:

- A place of variation in the text occurs where there are two or more 'competing' readings of the text, while the surrounding readings agree in all text versions; these places should be as small as possible.
- A place of variation suitable for the construction of a stemma is one that contains exactly two competing variants, each attested by at least two witnesses.
- Reordering of words (assuming the reordering is grammatically correct) may be used as a text-genealogical variation, so long as there are at least three words being reordered, none of which are adverbs.
- Nouns and verbs are the most suitable types of readings for creation of a stemma.

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

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

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

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

Between 2010 and 2012, a computational object model was developed, implemented as a Perl library, to represent a given tradition together with the variation in its witnesses as an interlinked graph; a companion model was developed, again based conceptually on a graph, to represent arbitrarily complex manuscript transmission. Use of these models made it possible to perform empirical analysis on a variety of stemmata produced using different methods (Andrews and Macé, 2013). The models also provide the underlying framework for a set of software tools that were used to perform the analysis and subsequently made available to other textual scholars for their own use (Andrews, 2012b). One tool allows the categorization and annotation of the way in which individual variant readings are related, another allows the specification of one or more stemma hypotheses, and a third performs an analysis and cross-correlation of reading variants with their consequences for any of the existing stemma hypotheses. The initial experiments conducted using these tools also corroborated the findings that 'insignificant' variation was surprisingly likely to follow text-genealogical transmission patterns in both artificial text traditions and genuine traditions for which reasonable certainty of the stemma can be had; we concluded that the

application of syntactically based categories of the sort that are relatively straightforward to identify automatically using linguistic analysis parsers (e.g. spelling variation, grammatical variants of the same word, variants that involve different words fulfilling the same grammatical function, which were termed 'lexical' variants in the tools) does not tend to pick out the sorts of variation that are more or less likely to indicate the copying history of the text.

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

2 The Artificial Traditions

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

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

This tradition was created for the comparison of several different methods for computational stemmatology (Baret *et al.*, 2006); this experiment is the only one to date for which the results of 'classical', non-computational methods of stemma creation were included alongside the computational versions. In the published experiment, one of the two non-computational methods came closest to reproducing the true stemma, although the computational methods (none of which are able to infer the sort of contamination that was present in the true stemma) were assessed on the basis of the raw output of the algorithm, without any interpretative intervention. The authors note that 'most philologists' were easily able to observe the shift of exemplar from the collation alone, which suggests that, had the computational methods been subject to interpretation, the outcome may well have been different.

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

The third artificial tradition is a text in Old Finnish, *Piispa Henrikin Surmavirsi*. This text, also known as the 'Heinrichi' tradition, is roughly 1200 words long and was copied by 17 volunteer scribes. Sixty-seven copies were made, of which 47 were made available for analysis (see Fig. 3 for the stemma). The creators of this tradition wished to simulate medieval copying conditions as far as possible in the modern era; in service to that goal they chose a text in an archaic language that was only imperfectly known to most of their scribes (speakers of the modern language), they produced a far larger set of manuscript texts, they had some of the volunteers make two or three copies from different exemplars, and several of the copies were mutilated after the volunteer work of copying had finished to simulate damage to manuscripts that tends to occur over time. This tradition was the primary data set used in a 'computer-assisted stemmatology challenge' run in 2007 (Roos and Heikkilä, 2009); both the *Notre besoin* and the *Parzival* artificial traditions were

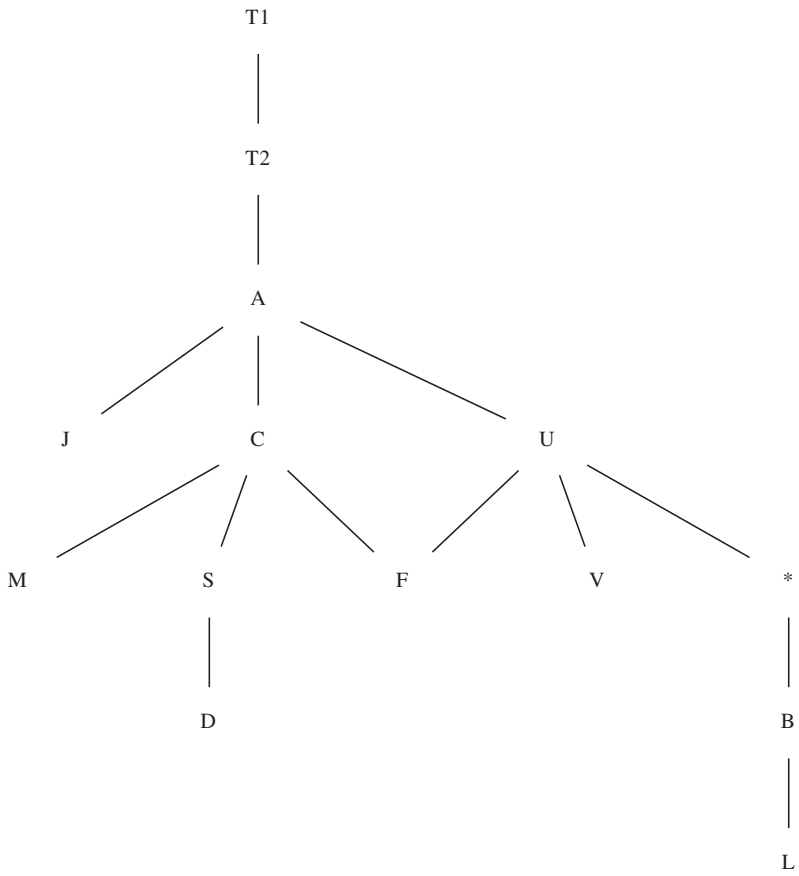

Fig. 1 Stemma for the *Notre besoin* artificial tradition.

also provided to challenge entrants. No attempt at a stemma reconstruction by hand of the *Heinrichi* text was reported during the challenge.

3 The Experiment

For each of the artificial traditions, a volunteer philologist agreed to use the Stemmaweb software (Andrews, 2012b) to categorize the textual variants according to whether, in his or her opinion, the variation was stemmatically significant; in the case of the Parzival text, two volunteers were found. The volunteers were chosen both for their experience in the practice of philological reconstruction of medieval texts and for their native or near-native familiarity with the language of the text. If there were more than two readings in a variant location, then the determination had to be made for each pair of readings with respect to each other at that location. Since the philologist did not consult the stemma, it was impossible to have any external verification of which reading in a set of variant readings came from the archetype, and which were derivative readings.

The premise to be tested is this: a trained philologist should be able to choose variants as 'significant' that do, in fact, genealogically follow the true stemma. The converse is not true; the philologist should not be expected to choose with any certainty those variants that positively contradict the stemma; to call a variant 'insignificant' merely means that it cannot be relied upon to provide

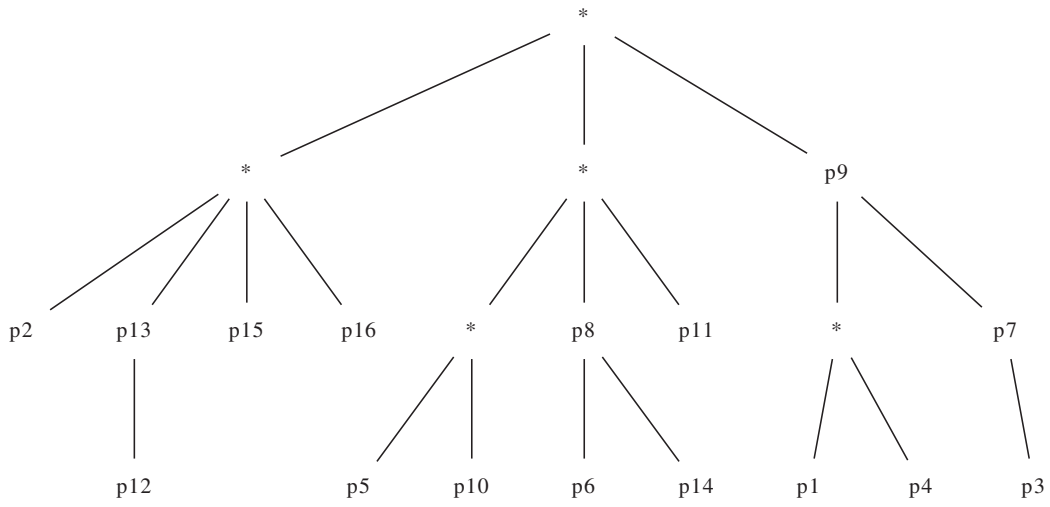

Fig. 2 Stemma for the *Parzival* artificial tradition.

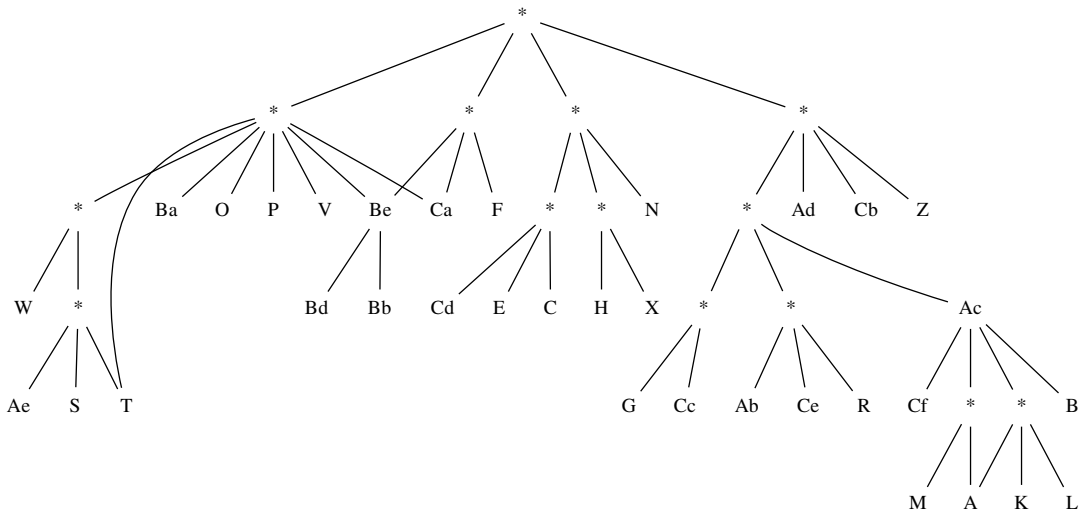

Fig. 3 Stemma for the available texts of the *Heinrichi* artificial tradition.

text-genealogical information. A great many so-called 'insignificant' variations happen to follow the stemma in all three of the texts.

The *Notre besoin* and *Parzival* texts were not normalized in any way; the *Heinrichi* text, due to its sheer size and complexity, was normalized for spelling. Since spelling variation is almost universally considered not to be stemmatically significant, it was felt that this normalization would not harm the philologist's chances of choosing 'significant' variation.

The Stemmaweb text annotation interface presents the variant texts as a unified 'variant graph', in which textual alternatives are represented relative to each other in a continuous presentation of the entire text (c.f. Schmidt and Colomb, 2009; Andrews and Macé, 2013; Dekker *et al.*, 2014). The user may create a

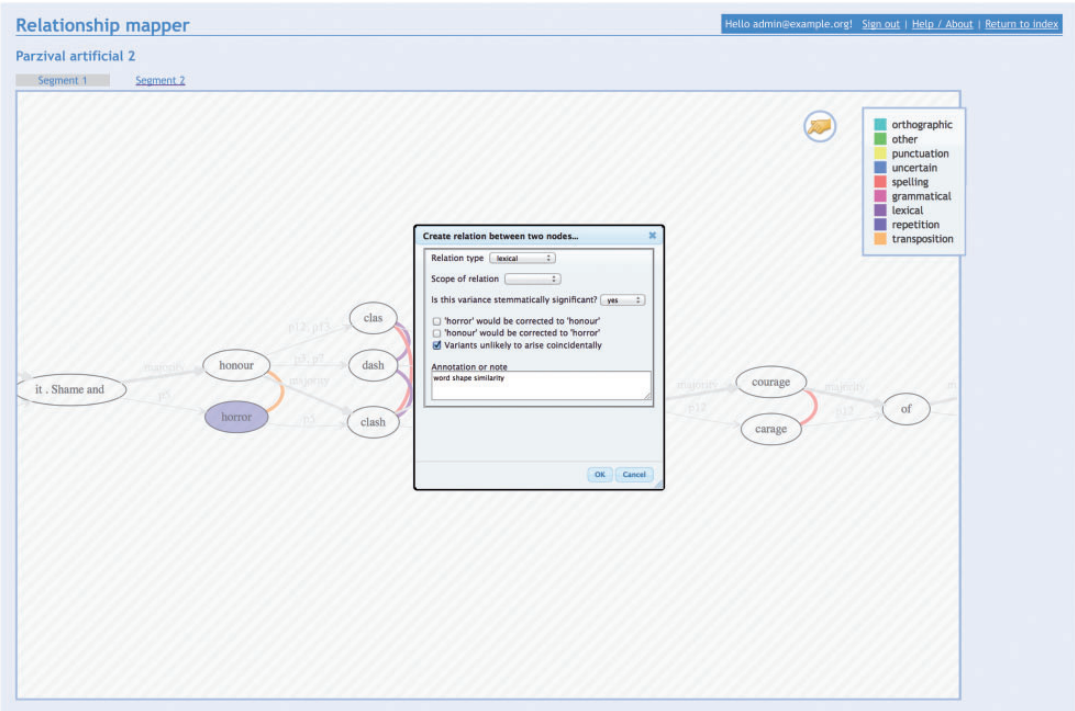

Fig. 4 Variant classification interface for Stemmaweb: creating a relationship between the parallel readings 'honour' and 'horror'.

relationship between two analogous reading nodes, and define several properties of the relationship (see Fig. 4). In this case the philologist had the option of providing any or all of the following information:

- How the readings were related syntactically (e.g. whether it was a spelling, grammatical, or some other sort of variation; whether the readings were variant grammatical forms; whether they were different words filling the same grammatical role in the sentence).
- Whether the variation was significant (possible answers were 'yes', 'maybe', and 'no'.
- Whether the variation was unlikely to have occurred coincidentally.
- Whether a scribe, upon seeing reading A, might 'correct' it to match reading B without reference to another exemplar (or vice versa).

There is currently a deficiency in the Stemmaweb software, so that there is no way to indicate whether a gap (or addition) in the text is stemmatically significant. The volunteer philologists were made aware of this deficiency at the outset of the experiment, and each of them was asked to keep a list of which addition/omission variants might be significant. Two such lists were received, both for the Parzival text; for the other texts, the philologists working on the texts simply stated guidelines to be applied for these variants. In both cases they advised that they were likely to be significant, unless it was purely a question of easily-replaceable readings such as punctuation.

Once annotated, the text variation was compared against the true stemmas for each tradition. For this, the text is subdivided into 'variant locations'—these are places in the text where variation occurs, and in terms of the graph a variant location occurs wherever more than one readings occurs at the same rank (that is, the same number of readings distant from the nearest shared prior reading) in the graph. In order to avoid artificially inflating the number of variants, each graph was compressed before analysis,

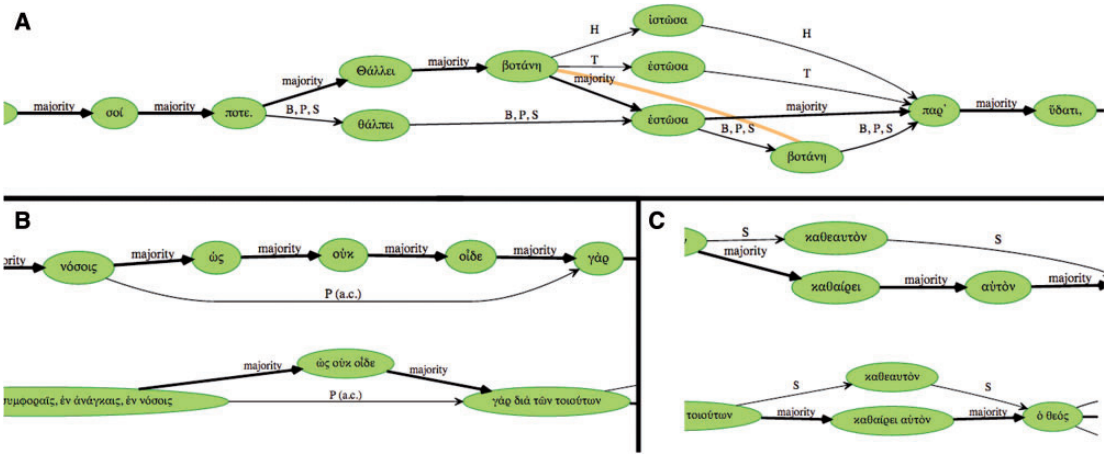

Fig. 5 Examples of reading compression before analysis.

so that individual sequences of readings that did not vary between witnesses, and for which no individual relationships had been made to parallel readings, were treated as a single reading. Three examples of a graph with compression rules applied are given in Fig. 5. In the example marked A, the relationship that marks the transposition of βοτάνη prevents compression, so that the transposition is treated as one variant, and the substitution of *θάλπει* for Θάλλει is treated as a separate reading, even though both of these are characteristic of the witnesses BPS. In example B, on the other hand, the entire phrase *ὡς οὐκ οἶδε* is treated as a single omission in witness P(a.c.), and in example C the two words *καθαίρει αὐτὸν* are treated as a single reading with the alternative *καθεαυτὸν* in witness S.

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

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

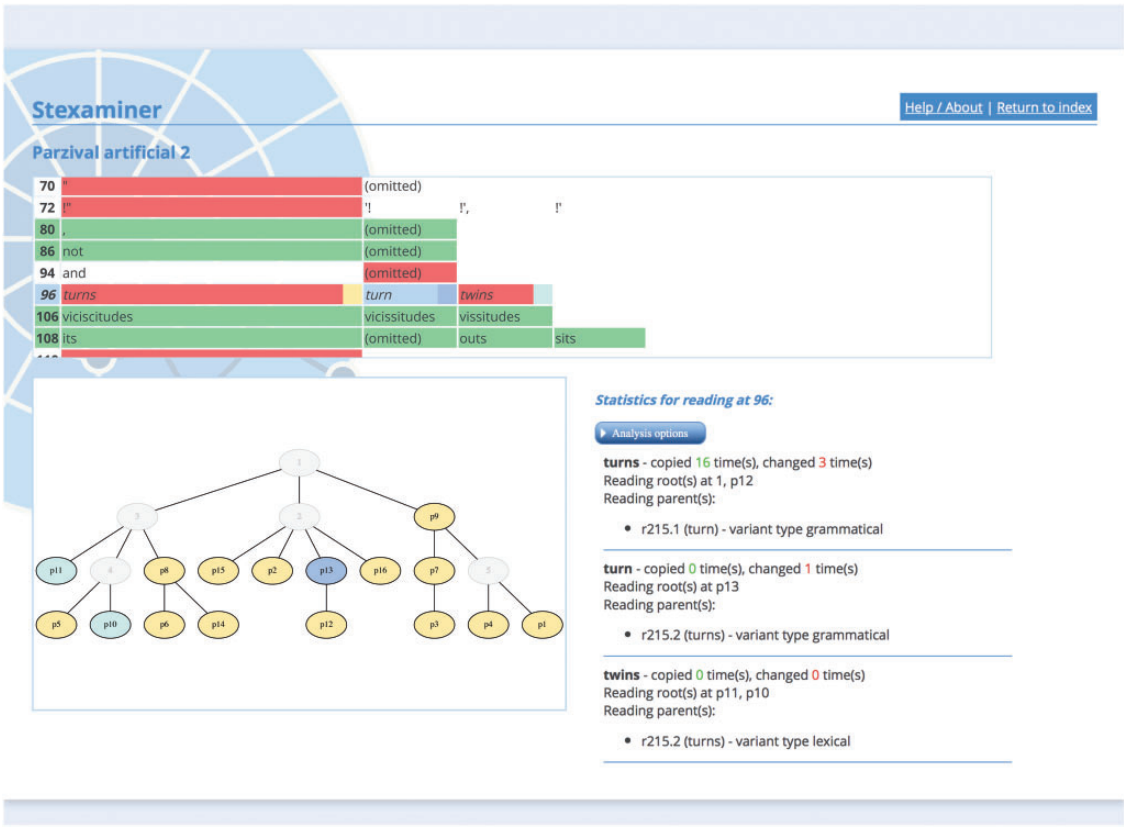

Fig. 6 Analysis of a variant location in *Parzival*. Three instances of variation are recorded: turns -> turn by witness p13, turn -> turns by witness p12, and turns -> twins by witnesses p11 and p10.

4 Results

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

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

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

Table 1 shows the aggregate results. For each text (normalized or not) in each scenario, the number of total variants assigned to each of the significance values 'yes', 'maybe', and 'no' is given, as well as the

Table 1 Aggregate results of variant analysis for the three texts

Classifications	Parzival 1	Parzival 1 normalized	Parzival 2	Parzival 2 normalized	Notre besoin	Notre besoin normalized	Heinrichi normalized
Including addition/deletion assumptions							
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
Excluding addition/deletion assumptions							
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
Excluding only deletion assumptions							
Total yes	15	14	9	9	22	23	194
Total maybe	35	30	11	9	16	14	0
Total no	112	46	141	68	55	25	629
Genealogical yes	11	10	5	5	16	16	103
Genealogical maybe	19	15	7	6	11	9	0
Genealogical no	65	27	82	38	40	18	382

number of variants in each category that were found to follow the stemma in a genealogical fashion. Reading the table, for instance, we can see that within the non-normalized *Parzival* tradition there were 211 variants in total, of which 20 were deemed significant and 51 were deemed potentially significant. Thirteen of 20 (65%) of the readings deemed significant were in fact genealogical according to the stemma; 31/51 (60.8%) of the readings deemed potentially significant were genealogical.

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

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

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

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	

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
9	Yes	Rue	See	reading
45	Yes	Hare	Horse	
53	No	Reveal	Several	
176	Yes	Is	In	
205	Yes	Rate	Note	
205	No	Note	Rate	Reverted
407	No	Cum	And	reading

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

Table 6 shows the results of the chi-square analysis across texts and scenarios. The only text to show a strong correlation between 'significant' and 'genealogical' is *Heinrichi*. In the case where additions and/or deletions are included, however, this extremely strong correlation is highly negative!

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-cahlehisa	routa-cahlehisa	
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	heittelepi	heittelemi	

(continued)

Table 5 Continued

Text position	Genealogical?	Exemplar reading	Copy reading	Note
655	Yes	kijruhti	lähti	
658	Yes	Lalloi	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	dorkuhujjnen	
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	

These are the scenarios where text additions are usually assumed to be significant, and deletions are usually assumed to be in the 'maybe' category. If we refer back to the numbers in table 1, however, we find that 96/186 (51.6%) of significant variants are genealogical, as compared to 382/749 (51%) of insignificant variants but only 115/420 (27.3%) of possibly significant! In this case, the decision to treat additions and deletions in this categorical manner has had a disastrous impact on the result.

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

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

We might thus apply a simpler test: to compare the success rates of the 'significant' and 'possibly significant' categories to the mean success rate of the text as a whole. We can treat this situation as a binomial distribution (with the same caveat concerning the independence of genealogical variants),

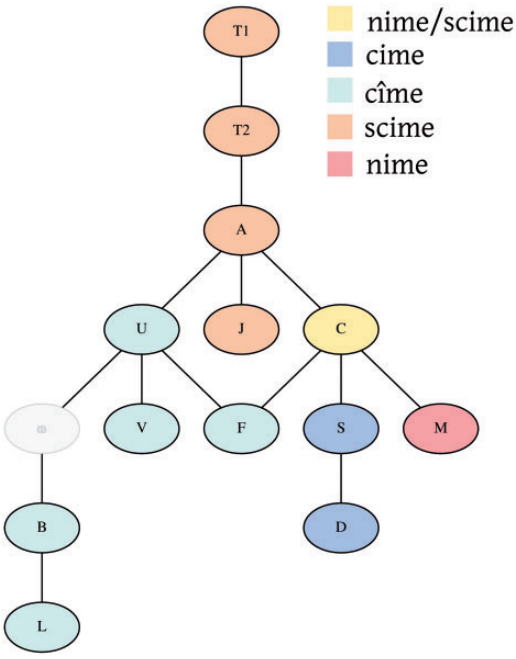

Fig. 7 A variant location that is genealogical only before normalization.

and analyse the 'significant' group as a sample drawn from the whole. In this case, the successful philologist should have constructed a sample of 'significant' variants that should have a markedly higher mean success rate than the wider population of variants. (The same analysis can be performed on the population of 'possibly significant' variants, but we would not expect such a marked difference in the success rate, so we omit that analysis here.) The specific question we ask is: what is the probability that a random sample of variants would have at least the same number of genealogical variants as our significant sample?

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

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

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

Data sets	χ^2 value	*P*-value
All variants		
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
Excluding addition/deletion		
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
Excluding deletions		
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

5 Conclusions

What are we to make of these rather surprising results? Above all it is important to bear in mind that the experiment was done using artificial traditions. Particularly for the *Notre besoin* text, many of whose copyists were themselves philologists, there is a real risk that the volunteers consciously or

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

Data sets	% mean genealogical	% genealogical significant	Likelihood of randomness	Standard deviation
All variants				
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
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
Excluding addition/deletion				
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
Excluding deletion				
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

semi-consciously introduced innovations into their copies in order to make the resulting tradition 'interesting'. On the other hand, also in the case of *Notre besoin*, at the time of the original experiment a philologist using classical methods was able to reconstruct a stemma that was not very different from the true stemma. Is this a case of one philologist simply being better than the other? While that is possible, it is not tremendously likely; over 70% of 'all' variation within *Notre besoin* followed the stemma, which made its reconstruction a comparatively straightforward task no matter what method was used. The results of that experiment bore this out: they showed that every one of the attempted methods, including the computational methods whose results were not manipulated into a 'normal' rooted stemma, could correctly identify the main manuscript groupings. That does in itself raise another question: how accurate must we be in choosing significant variation in order to reconstruct an accurate stemma? Although none of the volunteers in this study attempted to draw a stemma, one of the two philologists for *Parzival* provided a set of observations concerning which manuscripts should be grouped together; these were broadly accurate, even though the selection of individual significant variants was often wide of the mark; it is also worth noting that the philologist quite often cited variants as examples of group affinity that were not judged significant!

Compared to the rest, the *Heinrichi* artificial tradition fared comparatively well. The overall mean rate of genealogical variation in that text was rather lower than in the other two texts, at just under 44%. The *Heinrichi* corpus includes two to three copies per scribe, which increases the possibility of horizontal transmission (particularly for spelling and grammatical idiosyncrasies) in a different way; on the other hand, that tradition appears to have contained many more genuine errors, and the philologist who did the work was accordingly more accurate—leaving aside the question of additions and deletions—in detecting whether variation was significant. The creators of *Heinrichi* seem

to have had more success than the others in creating a tradition that is reasonably close to the 'real-world' situation of a medieval text widely copied.

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

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

Finally, this experiment makes clear that stemmatology has some way to go before it can claim the title of a 'normal science' that Rens Bod has offered. Our systems of categorization are suspect; our very philological sense of what is or is not significant has not fared as well as we ought to expect in the test against artificial traditions. We have more work to do than Bod's simple 'problem-solving'; we have yet to capture in any formal, demonstrable, or falsifiable way the essence of what scribes were likely to copy and what they were likely to change. If stemmatology is indeed to become a science, this is the next task that needs to be done.

Acknowledgements

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 article.

References

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

Andrews, T. L. (2012b). *Stemmaweb - a Collection of Tools for Analysis of Collated Texts.* http://byzantini.st/stemmaweb/(accessed 18 April 2014).

Andrews, T. L., Blockeel, H., Bogaerts, B. *et al.* (2012). Analyzing manuscript traditions using constraint-based data mining. *CoCoMile 2012 - COmbining COnstraint Solving with MIning and LEarning* Montpellier. http://cocomile.disi.unitn.it/2012/papers/cocomile2012_manuscript.pdf.

Andrews, T. L. and Macé, C. (2013). Beyond the tree of texts: building an empirical model of scribal variation through graph analysis of texts and stemmata. *Literary and Linguistic Computing*, **28**(4): 504–2110.1093/llc/fqt032.

Baret, P., Macé, C., and Robinson, P. (2006). Testing methods on an artificially created textual tradition. *The Evolution of Texts: Confronting Stemmatological and Genetical Methods*. Pisa; Rome: Istituti Editoriali e Poligrafici Internazionali, pp. 255–83.

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

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

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

Dekker, R. H., Hulle, D., van Middell, G., Neyt, V., and van Zundert, J. (2014). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project. *Literary and Linguistic Computing*, **29**, doi:10.1093/llc/fqu007.

Greg, W. W. (1927). *The Calculus of Variants: An Essay on Textual Criticism*. Oxford: Clarendon Press.

Housman, A. E. (1921). The application of thought to textual criticism. *Proceedings of the Classical Association*, **18**: 67–84.

Howe, C. J., Connolly, R., and Windram, H. F. (2012). Responding to criticisms of phylogenetic methods in

stemmatology. *Studies in English Literature 1500-1900*, **52**(1): 51–67. doi: 10.1353/sel.2012.0008.

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

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

Roelli, P. and Bachmann, D. (2010). Towards generating a stemma of complicated manuscript traditions: Petrus Alfonsi's Dialogus. *Revue d'histoire des Textes, n.s.*, **5**: 307–21.

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

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

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

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

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

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

Smelik, W. F. (2004). Trouble in the trees! Variant selection and tree construction illustrated by the texts of Targum Judges. In van Reenen, P. T., den Hollander, A., and van Mulken, M. (eds), *Studies in Stemmatology II*. Amsterdam: Benjamins, pp. 167–206.

Spencer, M., Davidson, E. A., Barbrook, A. C., and Howe, C. J. (2004a). Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology*, **227**: 503–11.

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

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

Wattel, E. and van Mulken, M. (1996). Weighted formal support of a pedigree. In van Reenen, P. T., van Mulken, M., and Dyk, J. W. (eds), *Studies in Stemmatology*. Amsterdam; Philadelphia: Benjamins, pp. 135–68.

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