Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmata

Tara L. Andrews and Caroline Macé
KU Leuven, Belgium

Abstract

Stemmatology, or the reconstruction of the transmission history of texts, is a field that stands particularly to gain from digital methods. Many scholars already take stemmatic approaches that rely heavily on computational analysis of the collated text (e.g. Robinson and O'Hara 1996; Salemans 2000; Heikkilä 2005; Windram *et al.* 2008 among many others). Although there is great value in computationally assisted stemmatology, providing as it does a reproducible result and allowing access to the relevant methodological process in related fields such as evolutionary biology, computational stemmatics is not without its critics. The current state-of-the-art effectively forces scholars to choose between a preconceived judgment of the significance of textual differences (the Lachmannian or neo-Lachmannian approach, and the weighted phylogenetic approach) or to make no judgment at all (the unweighted phylogenetic approach). Some basis for judgment of the significance of variation is sorely needed for medieval text criticism in particular. By this, we mean that there is a need for a statistical empirical profile of the text-genealogical significance of the different sorts of variation in different sorts of medieval texts. The rules that apply to copies of Greek and Latin classics may not apply to copies of medieval Dutch story collections; the practices of copying authoritative texts such as the Bible will most likely have been different from the practices of copying the Lives of local saints and other commonly adapted texts. It is nevertheless imperative that we have a consistent, flexible, and analytically tractable model for capturing these phenomena of transmission. In this article, we present a computational model that captures most of the phenomena of text variation, and a method for analysis of one or more stemma hypotheses against the variation model. We apply this method to three 'artificial traditions' (i.e. texts copied under laboratory conditions by scholars to study the properties of text variation) and four genuine medieval traditions whose transmission history is known or deduced in varying degrees. Although our findings are necessarily limited by the small number of texts at our disposal, we demonstrate here some of the wide variety of calculations that can be made using our model. Certain of our results call sharply into question the utility of excluding 'trivial' variation such as orthographic and spelling changes from stemmatic analysis.

Correspondence:
Tara L. Andrews,
KU Leuven,
OE Griekse Studies,
Blijde-Inkomststraat
5 bus 3000,
3000 Leuven, Belgium.
E-mail:
tara.andrews@arts.
kuleuven.be

Literary and Linguistic Computing, Vol. 28, No. 4, 2013.

doi:10.1093/llc/fqt032 Advance Access published on 25 June 2013

1 Background

Although stemmata are commonly constructed by the editors of classical and medieval texts (Robinson and O'Hara 1996; Windram *et al.* 2008; Heikkilä 2005; Andrews 2009), the principles of their construction have often been questioned (Timpanaro 2005: Appendix C), and their utility is rejected entirely by some practitioners of the new philology (e.g. Driscoll 2010). One of the most glaring deficiencies of the traditional stemmatic model is the apparent inability to proceed once more than a trivial amount of 'contamination' (i.e. the conflation of readings from two or more exemplars into a single witness) has been detected or hypothesized. This limitation has convinced many medieval text editors, faced with complex and fluid traditions, that stemmatics is not a tool they can use.

The promise of computational methods was recognized early by stemmatologists. In recent decades, several computationally assisted stemmatic methods have arisen, both neo-Lachmannian (Roelli and Bachmann 2010; Salemans 2000; Wattel and van Mulken 1996a) and phylogenetic (Howe *et al.* 2004; Roos and Zou 2011), to assess the variant data less prejudicially and produce a best-fit approximation of a stemma. Despite the general success of these methods, there is still some disagreement within the field concerning how—or indeed whether—to select variants or weight their significance for analysis.[1] In addition, few of these methods have approached the question of how to produce or handle more complex stemmata, including cases of conflation of exemplar.

The rise of these formal computational methods, particularly of the neo-Lachmannian ones, has moreover reduced most models to that of the binary tree. Although it has been acknowledged that the stemmata of text need not be bipartite and that far too many of the published stemmata for medieval texts are (Timpanaro 2005: Appendix 3), the relative ease of computation of a binary tree has given rise to a trend in the field to embrace the binary tree as the only practical representation of the genesis of the text. Greg's 'Calculus' (Greg 1927) was an early example of a method for constructing a binary tree from a subset of variants found in the text; this was later adapted and refined (Dearing 1974). By 1996, it was considered 'naïve to pretend that pedigree-building can be used for reconstructing the real, historical transmission of a text' (Wattel and van Mulken 1996a), and Salemans (2000) even called for the scholar to restrict the evidence under consideration to 'type-2' variants—those with precisely two alternative readings, each of which appears in at least two manuscripts. Although the rationale for the binary stemma has its merits as the model most easily calculated from a set of text variants, simplicity of construction cannot be allowed to stand as a theoretical justification for dismissing the reality of non-binary text transmission. Philologists must be able to accurately model whatever complex and disorderly situation the historical evidence suggests.

Many scholars today share the desire for a computational model of stemma construction that can natively support internal nodes (i.e. surviving manuscripts that were themselves copied), multifurcation, conflation of exemplars, and even inclusion of external physical evidence of copying (such as damage to the manuscript unmistakably reproduced in later copies, or colophon notices of who copied what when). The most recent algorithmic contribution to computer-assisted stemmatological method, Semstem (Roos and Zou 2011) specifically addresses the first two of these. Although it is possible to detect a shift in exemplar within a manuscript (Wattel and van Mulken 1996b; Windram *et al.* 2008), other forms of conflation are much more difficult to approach with the current set of computational tools. The NeighborNet algorithm (Bryant and Moulton 2004; Spencer *et al.* 2004a) is at present the only option for detection of other forms of conflation, and that requires a high degree of judicious interpretation from the editor. Likewise, any external evidence of the text transmission history must be incorporated into the stemma by the scholar after the computational analysis is done.

The case for phylogenetic methods of stemma construction and the current state-of-the-art has been summarized recently by Howe *et al.* (2012); the authors argue that most genuine limitations of phylogenetic methods—the limited support for detecting conflation of exemplar, the need to

judge the significance (weighting) of variants before running any analysis, the need to include text-external evidence where applicable—are shared by any form of stemmatic analysis. We are inclined to agree, but what then is our way forward? There is a clear need for a more empirically based approach to the study of text variation. Only a few empirical studies have ever been carried out, each limited to the study of a single text (e.g. Schmid 2004; Schøsler 2004; Spencer *et al*. 2004b), but a broader study has not yet been attempted. A quarter-century ago, Timpanaro observed that it was not practically feasible to collect statistics on the genealogical significance of many variations across a wide range of texts. Without precisely this base of empirical data on textual variation, however, philologists will be limited to their own deductive reasoning when assessing the significance of variants. Although a full database remains some way in the future, the steady technological advance of the past few decades means that the data collection can no longer be said to be unfeasible. An empirical profile of texts and their variations would pave the way for better weighted models, less guesswork and deduction surrounding the significance of individual variants, and more and better methods for inference of stemma hypotheses from unsanitized (i.e. real) textual data. Philological judgment and human reasoning will always be necessary to the field, but how much better for the judgment to be supported by evidence?

2 An Empirical Model for Text Variation

The necessary first step in any empirical analysis is to consider the question of how to represent text variation. The vast majority of text editions represent variation in the form of an apparatus criticus or collation table. Although this format is ultimately flexible, published apparatuses tend not to be sufficiently consistent, complete or rigorously defined for our purposes—of the several critical apparatuses analyzed during the course of our research, each of them contained some error or ambiguity that required the editor to be consulted for clarification.

We may consider the available models for text variation with the fragment expressed in Table 1 in the usual form of a collation table. The left-hand column contains words (readings) as they appear in the base text, and each variant is expressed to the right with the list of witnesses in which the variant appears.

Most automatic collation programs will create an alignment table such as that given in Table 2; unlike a collation table, this requires no base text. The witnesses are arranged by column, with the complete text of each witness appearing sequentially in its column and matching words aligned in the rows. The alignment table has the advantage of being a more complete representation of the collated texts, but it cannot satisfactorily display overlapping variation or transposed readings. This is nevertheless the format most often required by computational methods for stemma construction, particularly those based upon phylogenetic algorithms.

More recently, collation programs have appeared that express the text variation as a graph. The idea of the variant graph was first introduced by Schmidt and Colomb (2009) and is used in the 'nmerge' tool of Schmidt; an alternative form of variant graph was developed for the CollateX tool (Dekker and Middell, 2011). The variant graph in either form is an elegant way to represent concurrence as well as variation within a text tradition; one may imagine a text running from beginning to end with a number of points of divergence, where each witness takes a single path through these divergences. Figure 1 shows a variant graph produced by CollateX for the texts in the table aforementioned; the directed lines indicate witness paths, and the dashed line on the lower right indicates that CollateX has detected a reading transposition.

We have adopted and extended the variant graph used by CollateX for modeling text variation in

Table 1 Sample collation table for a fragment of Sermo Augustini 158 (Boodts and Partoens 2011)

	'Apostolus insignes quae pertinent ad deum'	
insignes	insignis Vb12	in his *add.* Vb18 Vb21
quae	qui Vb18	
pertinent	pertineant Va6 Vb12 Vb20 Vb9	*tr. post* deum Vb11
deum	eos Vb20	christum Vb9

Table 2 The same variants as an alignment table

Va6	Vb11	Vb12	Vb18	Vb20	Vb21	Vb9
Apostolus	Apostolus	Apostolus	Apostolus	Apostolus	Apostolus	Apostolus
insignes	insignes	insignis	insignes	insignes	insignes	insignes
			in		in	
			his		his	
quae	quae	quae	qui	quae	quae	quae
pertineant		pertineant	pertinent	pertineant	pertinent	pertineant
ad	ad	ad	ad	ad	ad	ad
deum	deum	deum	deum	eos	deum	christum
	pertinent					

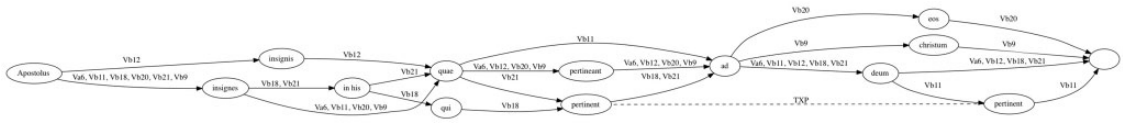

Fig. 1 CollateX-generated variant graph for the text fragment shown in Table 2.

sufficient depth for stemma analysis. The graph may easily be converted into an alignment table and back again, simply by treating each identical word on a table row as a single reading vertex and vice versa. It is also possible to construct a more traditional collation table from the graph by choosing a sequence of readings to act as the base text and noting a variant every time there is a divergence from the base text in the witness path.

The problem posed by transposition remains—the graph in Figure 1 mixes the meaning of its edges (connecting lines) and complicates the computability of the graph. For purposes of analysis, our graph should be directed and acyclic: only when each text has a common beginning, a common end, and a single direction of flow can we sensibly speak about readings that occupy the same place in the witness texts. We can preserve this characteristic by modelling the text tradition as two separate graphs containing the same set of vertices. The sequence graph is what remains of our variant graph with the transposition removed: vertices represent readings, and edges represent reading sequences within individual witnesses. The second graph describes the relationships between readings, including transposition. This graph is unrooted, undirected, and potentially cyclic, where the vertices represent the same readings, but edges represent variant relationships. We begin with the relationship graph in Figure 2 empty of any relationships; to this we may add edges that signify relationships of any sort, be they transpositions (identical readings in different locations) or syntactic or semantic relationships between variants in the same location. In Figure 3, our transposition is marked along with three links that represent a grammatical relationship between collocated variant readings. In the sequence graph, each edge is attributed with the list of witnesses that follow it; likewise, in the relationship graph, each edge is attributed with the category of relationship between the readings in question.

Our text model thus consists of the superimposition of these two graphs: the sequence of readings and the relationships between them. The relationship graph is constrained by the variant graph it shadows; its readings are identical, and as we will see in the following section, we may define relationships with constraints taken from the variant graph (e.g. that transpositions may not be defined between readings in the same place on the sequence graph, or between readings that occur in the same set of witnesses).

We have applied our model to a number of text collations—both artificial traditions and real

Fig. 2 Relationship graph with no relationships yet defined.

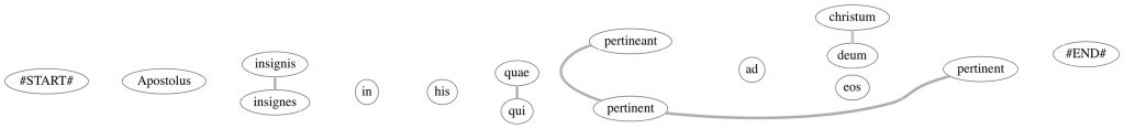

Fig. 3 Relationship graph containing defined relationships.

traditions provided by scholarly editors. The initial graph may be constructed from an alignment table, from a critical edition produced in Classical Text Editor or published in a parallel-segmentation format as laid down by the guidelines of the Text Encoding Initiative (TEI), or from an automatic collation program such as CollateX. In some cases, the original collation included variant relationship information that was taken into the graph; in other cases, the relationship information was added by scholars or students working for the project, sometimes with the aid of morphological tagging tools such as those provided by the Perseus Project (http://www.perseus.tufts.edu/) and Perseus under PhiloLogic (http://perseus.uchicago.edu/). In all cases, the representation of the text in a consistent model was a necessary first step for our subsequent stemmatic analysis.

There are many possible typologies of variant relationship that one might choose to analyze. The ones most commonly used for traditional analysis, such as that given by Reynolds and Wilson (1991), have at their core the presumption that 'error' can be distinguished from 'true' (i.e. original or archetypal) readings and are therefore not suitable for the construction of an empirical model wherein all such presumptions are discarded at the outset. We chose for our initial analysis a schema designed to facilitate automated classification, given a suitable base of linguistic data. Briefly, our categories are as follows:

- orthographic/spelling—the same reading, represented differently
- grammatical—the readings share a lemma or root but vary in their morphology
- lexical—the readings share a morphology or part of speech but come from different lemmata or roots
- transposition—the same (or nearly the same) reading, displaced within the text
- repetition—the same (or nearly the same) reading, repeated in one or more of the witnesses (primarily dittographies)
- uncertain—the readings are clearly related, but one or both are nonsensical; therefore, the relationship cannot be categorized
- other—the readings are related in some way not defined here
- unknown—no explicit or implicit relationship has been determined

The final category of variation that can appear in a text is that of the addition or deletion of a reading. Although, in the case of the other categories, it is a question of the relationship between two positive readings, addition/deletion is instead a question of the relationship between a reading and the absence of a corresponding reading in another manuscript. As such, these cannot be marked explicitly in our variant graph, but may be inferred automatically wherever a witness lacks a reading for a given location. Whether any particular case should be regarded as an 'addition' or a 'deletion' depends entirely on the stemma hypothesis—i.e. on whether the reading in question appears in a manuscript taken to be archetypal—and as such the graph model is purely agnostic on this point.

We will see in our results later in the text that this relationship classification, based as it is purely on orthographic and morpho-syntactic features of the

text, does not provide any new insight into the mechanics of text transmission. A semantic tagging scheme, for instance, one that includes information on the extent to which the variation changes the meaning of the surrounding text, would be interesting to analyze, but it was not feasible for this project. Despite their deficiencies from an empirical point of view, the relationship definitions we used do allow us to demonstrate the methods by which variation and the relationships between variants may be analyzed within a text.

3 Properties of a Stemma Graph

Armed with a sufficiently robust representation of textual variation, we may now turn to the modelling of the genealogical relationships between texts—i.e. to their stemmata. The definition of a stemma as a graph (or a network) is the essence of the idea of text genealogy; in his handbook on the subject, West (1973) gives an example of a graph that includes a case of multiple transmission and is therefore not tree-like. Unfortunately, this example is intended to illustrate West's conclusion that 'if contamination is present in more than a slight degree, it will be found that no stemmatic hypothesis is satisfactory'. The result of this assumption—that a contaminated stemma is problematic and should be avoided—has led to the more common definition of a stemma as a tree—that is, a graph that only branches and never merges.

West's observation may well be true for ancient texts, in that the text tradition can often be so attenuated that the classical scholar cannot hope to sketch the complexity of what may have once existed, and therefore he or she omits to produce a stemma. It becomes problematic when this limitation is accepted without reservation in medieval text philology, in which manuscripts that represent a conflation of exemplars routinely survive along with one or more of the exemplars, and hints are often left within manuscript colophons about the order of transmission of their texts (e.g. Robinson and O'Hara, 1996; Andrews 2009). In these cases, an arbitrarily complex 'historical' stemma that is a closer approximation to the true transmission history may be drawn, but how are we to analyze it?

Let us begin with a precise, if simple, definition of our terms. A stemma is a directed acyclic graph that indicates the network of copying relationships between manuscript versions of a text, beginning from a presumed (or, in rare cases, extant) archetype. It is immaterial for the model whether the archetype ever existed in written form; all that matters is that, for multiple texts considered to be versions of the same work, the origin of the stemma represents that work.

Within the stemma, the archetype takes its place as the single root vertex within the graph and each extant witness is represented as another vertex. Also included is any witness that is not extant, but for which there is evidence (e.g. a notice in a colophon) or whose existence must be postulated to explain a feature of the text transmission. The vertices are connected with directed edges; each represents an assertion that the target witness was copied (in whole or in part) from the source witness. A witness may be connected to zero or more targets (apart from the archetype, which must be the source for at least one); likewise, a witness (apart from the archetype, which itself has no source) may be connected to one or more sources.

At the core of stemmatology lies the idea that certain variant readings follow the stemma or are 'genealogical', in that, if a witness carries the reading then its descendant witnesses will also carry that reading (or a new variant, if another change has been introduced). This phenomenon is readily defined in terms of the stemma graph. A 'genealogical' variant has precisely one originating witness within the stemma (c.f. Figure 7 later in the text). That is, for all the witnesses that contain the reading, only one of these will not have an exemplar that also contained the reading. A 'coincidental' variant reading will, conversely, have multiple originating witnesses—multiple witnesses unconnected to each other in the stemma that did not derive the reading from their own ancestors (c.f. Figure 9 later in the text).

Part of the complexity of text transmission that must be adequately represented in a stemma is the incidence of scribal corrections. Medieval manuscripts were almost never copied without mistakes; thus, they almost never come down to us without

corrections marked throughout the text. Sometimes, these corrections are judged trivial (e.g. the immediate replacement of an omitted letter in a word that makes no sense without it), but corrections can often be a sign that the scribe was reverting a perceived mistake in the exemplar, or even that the text was checked against a second exemplar. How can these different textual layers be represented within the stemma?

Each vertex of our graph can carry at most one reading; wherever there are two possible readings, we must imagine two vertices for the text. The source witness(es) provide the text for both these vertices, but the uncorrected version might also be said to serve as a source for the corrected version. The resulting graph segment might be shaped like it is shown in Figure 4.

Here, the source witness α provides the text for both $T^{(a.c.)}$ and T, but the uncorrected text $T^{(a.c.)}$ is also the basis for the corrected text T and is thus represented as another source.

However, this model may not always be satisfactory. Consider the case of a manuscript W that has been copied from one witness A and corrected (perhaps by a later hand) from another witness B; the second witness serves as exemplar to W, but does not serve as exemplar to $W^{(a.c.)}$. This gives the graph segment shown in Figure 5.

Here, any reading in W that arises from a trivial correction to $W^{(a.c.)}$ is rendered genealogical via the $W^{(a.c.)}$ -> W link; the readings in W that arise from B are rendered genealogical via the B -> W link. If, however, B and $W^{(a.c.)}$ share a reading, that is a coincidence of variant and will be recognized as such by the graph model.

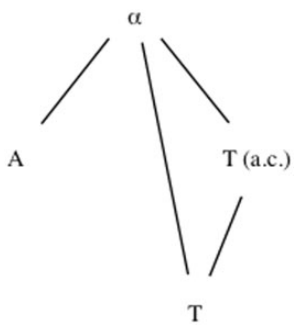

Fig. 4 Stemma fragment representing scribal correction.

This leads us to a solution for a more complicated situation that occasionally arises in medieval philology. It is not necessarily the case, particularly in much-copied and much-corrected traditions such as the Greek New Testament (Mink 2004), that the transmission of the text went in only one direction. Let us imagine, for example, that a witness A might have been copied with reference to two witnesses B and C, D copied from A, and B altered with reference to D. If we were to represent the texts in our graph without correction layers, the result would be like that shown in Figure 6.

This is a cycle—the presence of the path from B to A to D back to B means that the graph is no longer acyclic, which is a requirement if our analysis is not to descend into an endless loop. To return to a directed acyclic graph, we must represent B as two layers of text, if possible. Just as in the simpler case of Figure 5, any reading present in both $B^{(a.c.)}$ and D is a coincidence of variant, but any reading present in B and D may be treated as text-genealogical.

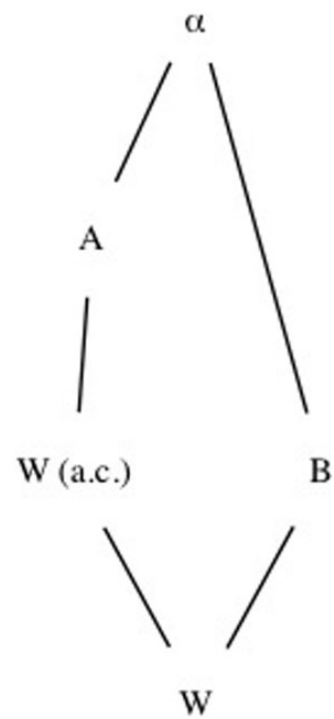

Fig. 5 A witness corrected against a second source.

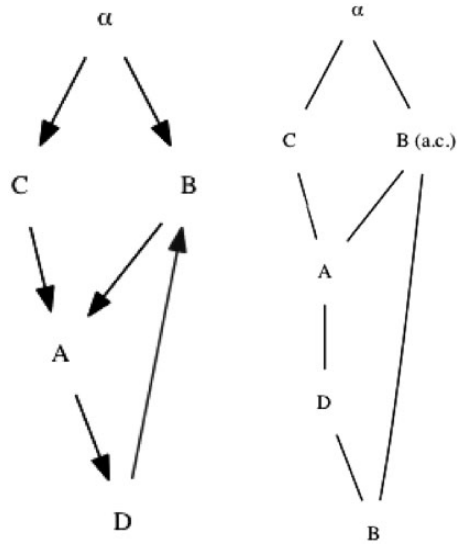

Fig. 6 A cyclic stemma and its acyclic equivalent.

With a combination of these methods, it is possible to represent a stemma hypothesis of nearly unlimited complexity, taking into account both conflations of archetype and correction layers within the text. These stemmata are not different from many complex stemmata that have been published, but their explicit modeling as a directed acyclic graph allows for the analysis techniques to which we will now turn.

4 Analysis of a Stemma Graph

Given a variant graph model of a text and a reasonably reliable stemma model of its witnesses, we may begin to construct an empirical profile of the copying process for that text. We begin on a small-scale with questions such as:

(1) For each variant, is it genealogical (that is, does it have a single origin within the stemma)? If not, how often did it arise coincidentally?
(2) For each variant, from which variant(s) was it copied (that is, what reading existed in the source(s) for this witness?) How are the readings related?
(3) For each variant, how often was it altered, and to what?
(4) For each variant, how often was it copied without alteration?
(5) For each variant, how often was it reverted to a reading that occurred in an ancestor witness of its source?

These questions may be answered on a variant-by-variant basis through cross-correlation of the stemma with the variant graph. Our analysis is based on a graph search algorithm (Andrews *et al.* 2012) implemented in the constraint programming language IDP (Wittocx *et al.* 2008). The analysis problem has been proven to be NP-complete (i.e. programmatically hard); this means that the amount of computation power necessary to solve it grows extremely quickly with any increase in complexity of the stemma, particularly when there is more than a trivial amount of witness conflation; our analysis would therefore not have been feasible until recently without access to large-scale scientific computing clusters. Constraint programming is a means of solving these problems as efficiently as possible. This approach allows us to assess for any stemma—even one with many instances of conflation—whether a given variant is genealogical, how often it was copied or changed, and/or whether it represents a reverted reading. Cross-correlation with the relationship information in our variant graph allows us to answer the remaining questions. An implementation of the solver is available online (Andrews 2012a) for general use with any uploaded text tradition.

A simple analysis example can be seen in Figure 7. This example is taken from the 'Parzival' artificial tradition (Spencer *et al.* 2004b); the majority of witnesses have the reading 'clash', whereas two have 'clas' and two have 'dash'. The stemma displayed is the true stemma for the tradition; our solver has detected that the variation in the witnesses does follow the stemma. Each reading has a single root (origin) within the graph, and we may additionally say that the archetypal 'clash' was changed once to 'clas' and once to 'dash'. Each of the non-archetypal variants was copied once and nowhere reverted.

Figure 8 shows an example of a detected reading reversion. In this case, the archetypal reading is the word 'motley'; the scribe of witness p9

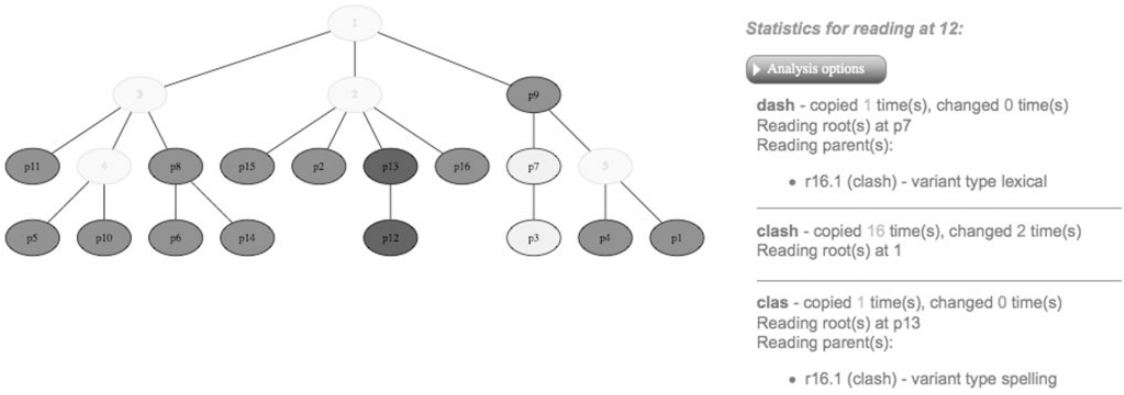

Fig. 7 Analysis of a genealogical variant within the Parzival artifical tradition.

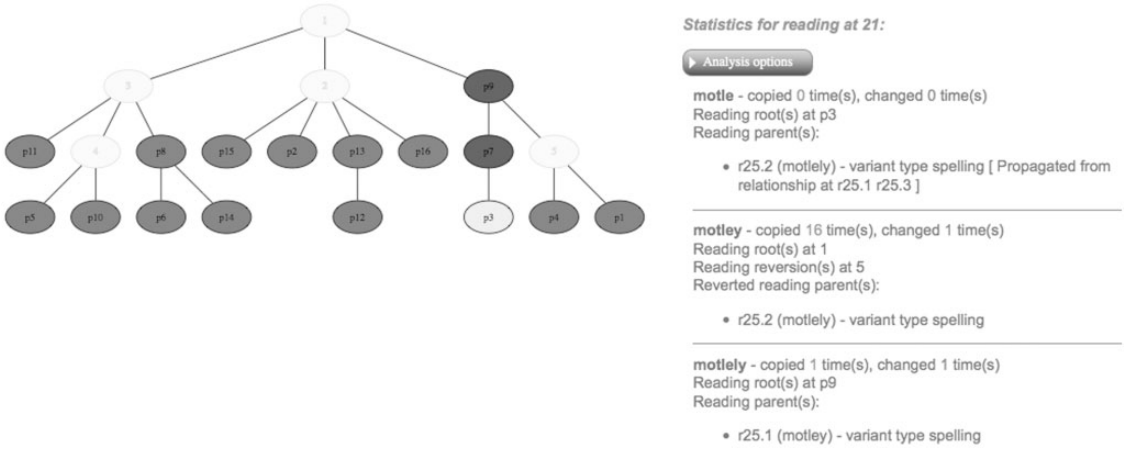

Fig. 8 Analysis of a reading reversion in Parzival.

introduced a spelling error. Although p7 retained the error and p3 mutated it farther, our solver has detected that the error was reverted to the original in the witnesses p4 and p1 (and presumably, from the point of view of the calculation, in their lost exemplar).

In Figure 9, we see an example of the third of our possibilities, an independently arising variant. In this case, the archetypal reading was 'base'; a witness in each of the three main groupings (p2, p3, and p10) has changed this to 'bare'. Both these words are adjectives in context, making this a 'lexical' variation in our classification.

By combining these analyses, we may look at how frequently variation occurs within the text, how many variant readings tend to occur in a single variant location, and the breakdown in types of variation. The 'Parzival' text is 834 words long; we find 262 variant locations in the graph, broken down in Table 3 by number of variant readings per location.

In all, 145 variant locations consist only of so-called 'type-1' variants, in which a single manuscript that was not copied by any other manuscript carries a different reading from the rest. These variants are not genealogically informative—by our graph criteria, they always count as 'genealogical' simply because they occurred only once and were neither copied nor changed. As such, we exclude them from the remainder of our analysis.

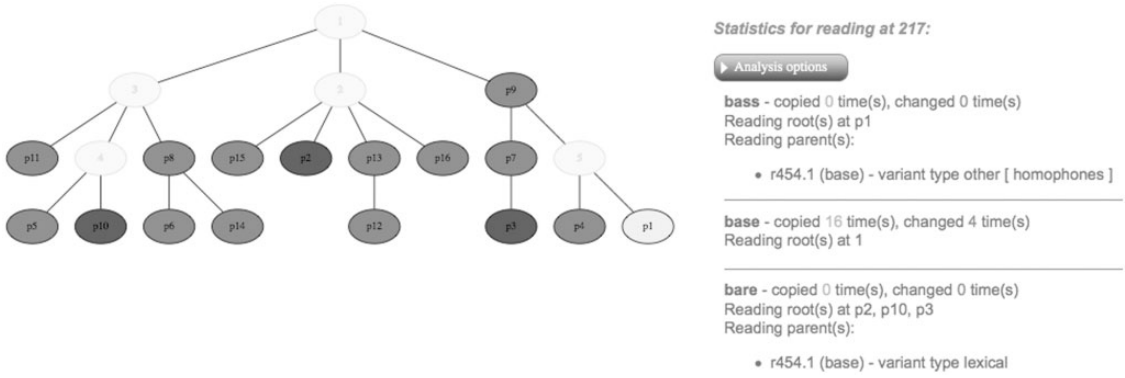

Fig. 9 Analysis of a coincidental variant reading in Parzival.

Table 3 Breakdown of variant readings per location in Parzival

Variant readings per location	Number of locations
2	211
3	36
4	10
5	5

Of the remaining 232 genealogically informative variants (in 117 locations), we can now gather a profile of the categories of variation, i.e. what relationship the variant had with the variant in its source witness (if any) at the point of mutation. The categories displayed in the following graphs are those named aforementioned on page X, where orthographic and spelling variation is combined into the single category 'sameword'. Additions and deletions are also included here, oriented into their respective categories according to our stemma hypotheses. The vertical axis in each graph represents the absolute number of variants analyzed.

For 'Parzival', the variants fall into the categories shown in Figure 10. This is a text in English, translated from medieval German with some consequent unusual language, copied by volunteers in Cambridge who were, by and large, fluent in the language and well-educated (Spencer *et al.* 2004a). The greatest total amount of variation was in word orthography and spelling, as might be expected for a text copied under such experimental conditions. Addition and deletion of readings was also reasonably common, although the relatively high incidence of coincidental reading deletion (i.e. deletion of the same reading in manuscripts not related in the stemma) seems surprising. Closer inspection shows that many of these coincidental deletions are of punctuation, but several incidences remain where words or phrases that would normally be regarded by philologists as substantial are deleted in unrelated manuscripts. It is not known whether any of the texts in this tradition were copied by the same scribe, or whether any of the scribes had a prior familiarity with the text; in any event, the presence of these 'significant' but non-genealogical variants is intriguing. The relatively high incidence of 'lexical' variation is largely explained by misreadings of letter or word shape, e.g. 'dash' for 'clash' or 'companion' for 'comparison'; here, our classification system would have benefitted from a category to indicate similarity of word or letter shape.

The other artificial traditions give different profiles. A case in point is the 'Notre Besoin' tradition, a text of 1015 words in French translated from a Swedish original (Baret *et al.* 2006). The total amount of variation is markedly lower than in the slightly shorter 'Parzival'; this may be partially explained by the smaller number of exemplars, but in fact the variation is often so slight that it is difficult to produce a stemma according to traditional philological principles.

The vast majority of the variation as seen in Figure 11 does in fact follow the stemma; what little coincidence in variation is found primarily

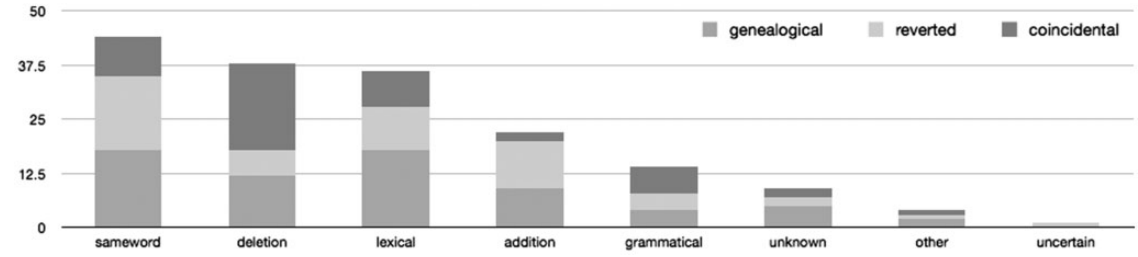

Fig. 10 Breakdown of variation by relationship type in Parzival.

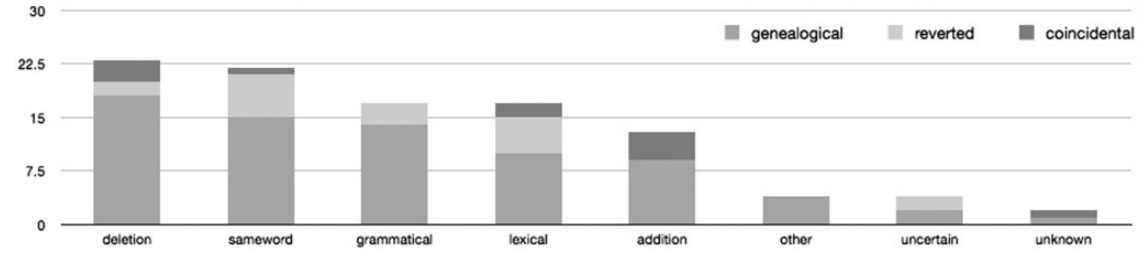

Fig. 11 Breakdown of variation by relationship type in Notre besoin.

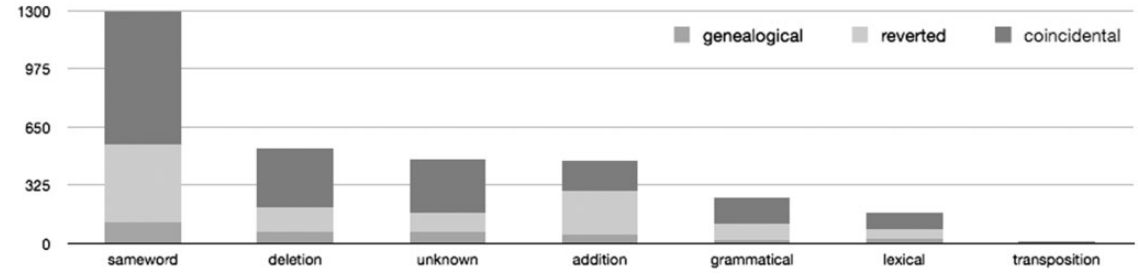

Fig. 12 Breakdown of variation by relationship type in Heinrichi.

concerns spelling or punctuation within the text. As in the 'Parzival' tradition, many of the 'lexical' variants could be traced back to letter or word shape similarity, e.g. 'avides' for 'arides' or 'joie' for 'jour'. In this sense, although the 'Notre besoin' tradition certainly contains features in common with traditions copied in medieval times, it has far too little variation to reflect accurately the features of a medieval tradition.

The 'Heinrichi' artificial tradition (Roos and Heikkilä 2009), a text of roughly 1100 words in Old Finnish, goes perhaps too far the other way. The creators of the tradition set out to simulate as realistically as possible the situation of copying a Latin text in the medieval era; most of their volunteers were native speakers of modern Finnish, for whom Old Finnish would have been a little unfamiliar. The tradition does contain a much larger set of witness texts and therefore a much larger amount of variation (2655 variant readings in 917 locations) than either of the two other traditions. Its variants fall into the categories shown in Figure 12.

The lack of standardized spelling in Old Finnish serves to explain the much higher amount of

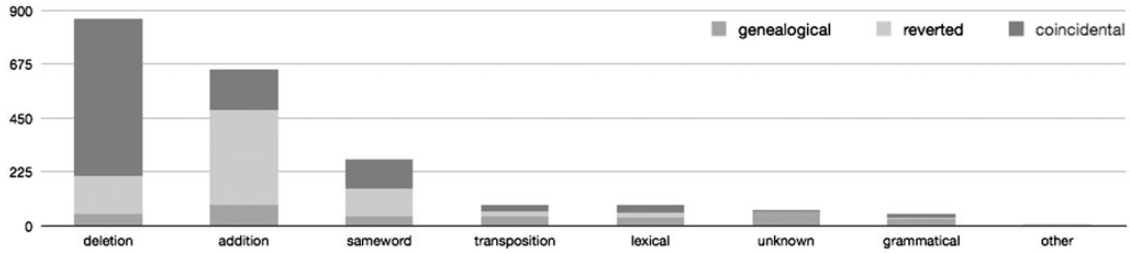

Fig. 13 Breakdown of variation by relationship type in the Legend of Bishop Henry.

variation in general; the 'unknown' category here may reflect the fact that at least one of the texts contained a translation into Latin of the original Finnish, a situation for which our categorization did not account. What is striking about this tradition is that the vast majority of variation conflicts with the stemma somehow; although there is some evidence of reversion (i.e. scribal correction), most of the coincidence in variation cannot be explained that way. Here, the 'Heinrichi' tradition served as a good test case for our tools, but how much can it tell us about the medieval phenomenon of text copying?

To answer that question, we must turn to the other texts at our disposal. Alongside the artificial traditions, we had a number of genuine medieval texts in the form of witness collations contributed by different scholars. The four we examined for this project were written in Latin, Greek, and Armenian; two had been regularized for spelling and three for punctuation.

Let us first consider the 'real-life' analogue to the 'Heinrichi' artificial tradition: the 'Legend of Bishop Henry', a 13th-century Latin version of the saint's life and deeds that survives in more than 50 copies ranging from the 14th–16th centuries and that was edited by one of the creators of the artificial tradition (Heikkilä 2005). The stemma used for our evaluation is that prepared by the editor, who used a combination of computational methods and deductions based on text-external evidence in the manuscripts to arrive at his conclusions. In this case, as in every case involving a genuine tradition without a certainly known stemma, our observations and conclusions must be more or less tentative. Given in Figure 13 is the categorization of variants for the Legend.

Much as in the case of the artificial 'Heinrichi', little of the variation within the 'Legend' appears to follow the stemma. There is a great deal of deletion and only slightly less addition; this may arise from the fact that many of the copies omit entire sections that are present in other copies. Variations in spelling and orthography make up the next most frequent category, as seems often to be the case with medieval texts.

The features of the 'Heinrichi' tradition are therefore more or less accurately reflected in the 'Legend', given the stemma provided, but this is far from the case in other texts. The Armenian-language 'Chronicle' of Matthew of Edessa survives in more than 30 manuscripts, of which 20 were used to create a stemmatic analysis and critical edition of certain segments of the text (Andrews 2009). The excerpts edited include 2908 variants in 1861 locations; their analysis, a summary of which is given in Figure 14, serves as a counter example to 'Heinrichi' and the 'Legend' in the magnitude of coincidental variation and also differs from all the artificial traditions in the distribution of variants across categories.

The first striking feature of this text is the relative lack of non-revertible coincidental variation that sets it apart from the other texts. In the absence of more data on Armenian texts, medieval chronicles in general, and relative linguistic features, it is not currently possible to give an explanation for this variation; the example serves simply to demonstrate the grave dangers of attempting to generalize from too little data.

The second striking feature is that, unlike in our prior two examples, the 'grammatical' category here outweighs any of the others. This reflects the fact

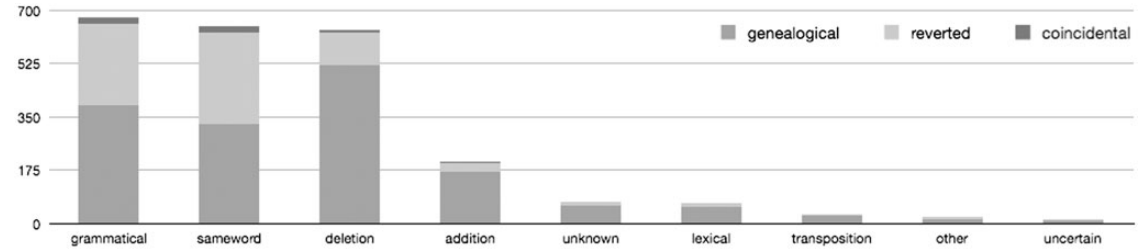

Fig. 14 Breakdown of variation by relationship type in the Chronicle.

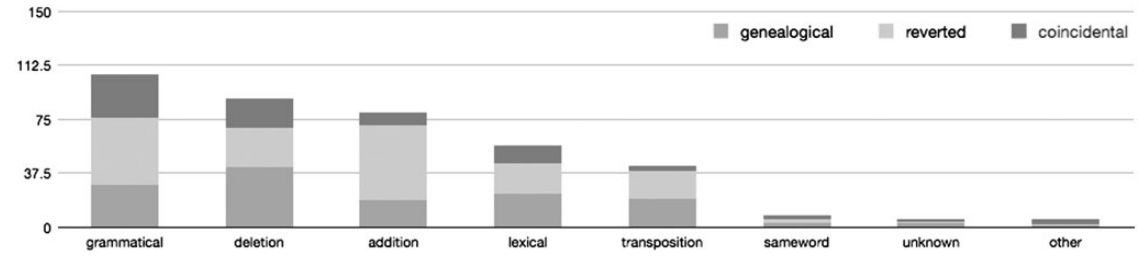

Fig. 15 Breakdown of variation by relationship type in Sermo Augustini 170.

that the most common variation by far is the omission or insertion of a definite article suffix to a noun, which was classed as a grammatical variation by our tagger and would not, moreover, be considered a particularly stable text-genealogical feature of any Armenian text. The example demonstrates the way in which a linguistic feature peculiar to a language may dramatically alter the logic behind a classification of variant relationships and may require a modification to the entire schema.

The two remaining medieval traditions, the 'Sermo Augustini 170' and the 'Florilegium Coislinianum B', were both collated by hand; minor variation in punctuation and orthography was therefore implicitly neglected in both these texts. Each of the texts has a stemma based partially on text-external evidence, derived through a combination of traditional Lachmannian and phylogenetic means. The 'Sermo Augustini 170' text has the added advantage that the oldest manuscript is believed to be the archetype for the entire tradition; this situation is exceedingly rare in medieval text studies and therefore the case has an added interest. Analysis of that text gives the results as shown in Figure 15.

Here, there is relatively little variation that cannot be explained by scribal reversions/corrections, as compared with the 'Heinrichi' and 'Legend' texts, although there is a bit of variation across categories that relies on an assumption that the variants were reversible.

The 'Florilegium Coislinianum B' has a similar, though not identical, profile, seen in Figure 16. The vast majority of the variation here can be explained by the provided stemma. Just as in the case of the 'Legend', there are several manuscripts that contain only parts of the full text, which explains the prominence of the 'deletion' category here; unlike the 'Legend', these relatively large omissions do seem to give a good indication of the copying history of the manuscripts.

5 'Significant' Variation in Transmitted Texts

These examples have served to highlight a few examples of the wide variety of ways in which texts could have been copied, mutated, and restored by medieval scribes. Any eventual empirical analysis of

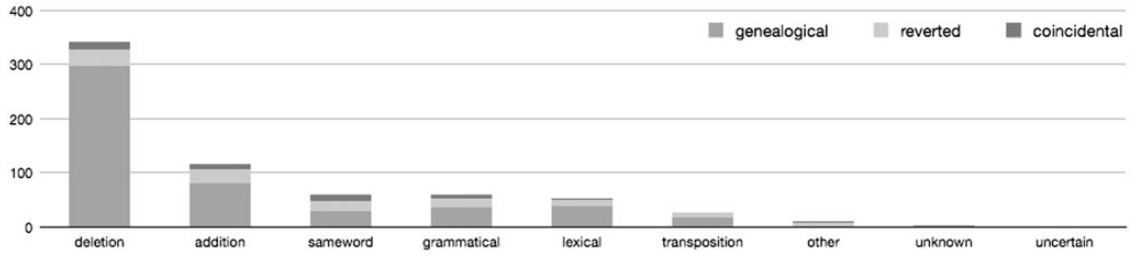

Fig. 16 Breakdown of variation by relationship type in Florilegium Coislinianum B.

Table 4 Text traditions available for analysis

Text name	Language	Approx. length	No. of witnesses
Parzival artificial (Spencer *et al.* 2004a)	English	834 words	16
Notre besoin artificial (Baret *et al.* 2006)	French	1015 words	13
Heinrichi artificial (Roos and Heikkilä, 2009)	Finnish	1100 words	35
Legend of Bishop Henry (Roos and Heikkilä, 2009)	Latin	1200 words	53
Chronicle of Matthew (Andrews 2012b)	Armenian	1800 words	19
[a]*Sermo Augustini 170* (Boodts 2012)	Latin	2800 words	18
[a]*Florilegium Coislinianum B* (De Vos *et al.* 2010)	Greek	3400 words	13

[a]non-diplomatically transcribed.

text transmission as a whole must take into account a large number of texts; it must also use a relationship model that represents enough information to give a definitive answer to the question with which we began: If we refrain from assuming *a priori* what constitutes 'error' in text copies, what features of variation are most likely to give us information about the order of transmission of the copies?

At present, we lack both a sufficient number of texts and a sufficiently complete system for tagging variant relationships. Even the small number of texts we have allows us to demonstrate the shortcomings of our current relationship typology. The seven texts available to us, each described in the preceding section, are listed in Table 4. The artificial traditions, although they have the advantage of a known and certain stemma, display none of the diachronic linguistic features (e.g. dialect shifts and non-standardized spelling) that frequently occur in medieval traditions.[2] Genuine traditions in their turn can only have stemmata that are partially certain at best; it is often also the case that no diplomatic transcription of the witnesses exists, and that the editors have excluded much of the variation (that which they have judged 'insignificant') from their collations. Of our four genuine traditions, only two ('Legend of Bishop Henry' and the 'Chronicle' of Matthew) have been diplomatically transcribed, and one of those transcriptions (the Legend) preserves no scribal correction information. All of our texts are moreover prose texts in a single language; the transmission problems presented by poetry and translated works are not represented here.

All seven texts had different empirical profiles in terms of the amount of variation, the frequency with which their variants followed the stemma, and the proportion of variation that fell into each of our relationship categories.

It is a commonplace of stemmatics that orthographic and spelling variation is more likely to reflect shifts in dialect and scribal practice than the features of the exemplar; these variants are considered likely to be coincidental and should be excluded from stemma construction on that basis. We would therefore expect to find that variation of this type is consistently less likely to be genealogical in all our texts, but this was the case only for the 'Chronicle' of Matthew. We would likewise have

	(none)	Spelling	+Grammatical	+Lexical	+Transposition	+Addition	+Deletion
Mean	0.59172	0.60717	0.63936	0.62397	0.60860	0.64015	0.65368
Std. dev	0.25388	0.25657	0.26988	0.26128	0.25884	0.23786	0.21736

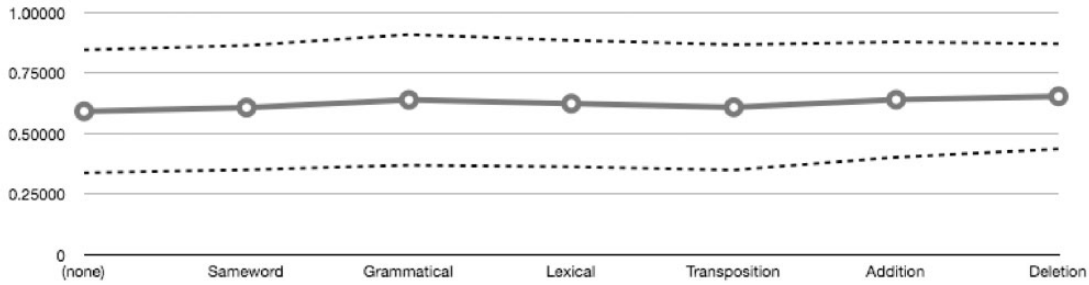

Fig. 17 Percentage of genealogical variation when the given relationship types are excluded.

expected that, when these variants are treated as the same reading, the proportion of genealogical variation in the remaining readings will rise substantially relative to the coincidental (and reverted) variation. This is not the case, as we can see from the values given in the first two columns of Figure 17. These show the average percentages of genealogical variation when each relationship type is excluded in turn. Thus, on average across the texts, 59.2% of the variation is genealogical; the number rises only slightly, to 60.7%, when we exclude orthographic and spelling variation. (The relatively high standard deviation of each text from the mean reflects the wide variety in the profiles of our individual texts.) Our results moreover appear to support the findings of Blake and Thaisen (2004), who demonstrated that a scribe's system of spelling can reflect the exemplar that was used.

Given that spelling regularization has so little effect on the extent to which variation in a text matches its stemma, could any of our categories be said to have a definite effect? We excluded in turn each of the other relationships in our typology (in addition to our spelling regularization, in each case) and found only small differences as shown in the remainder of Figure 17.

The standard deviation values are relatively high in all cases, which reflects the fact that none of the individual texts have profiles that are particularly consistent with each other.

The implications for what might constitute a 'significant' variant are interesting. Philologists have long agreed that trivial variation is more likely than anything else to be coincidental, and some have devised elaborate rules for the selection of stemmatically 'significant' variants; at the same time, it is well known to many scholars of medieval texts that even those variants that seem the most significant (for instance, a text with headwords that, in certain witnesses, have been elaborated and explained) might have a hidden source of transmission that is entirely separate from the original text (e.g. a widely available lexicon that includes entries for the headwords in question) and can therefore not be counted as genealogical. Our findings broadly support the conclusions reached by Spencer *et al.* (2004b) and suggest that even the most trivial changes, taken in aggregate, have some text-genealogical significance that should not be discounted.

6 Conclusion and Future Directions

The aim of this article has been to consider how philologists might take advantage of computational methods and the ability to process data on a large scale to arrive at an evidence-based, rather than an assumption-based, body of knowledge on how to reconstruct the copying history of medieval texts.

The first and most necessary step along that path has been to formalize some of the phenomena of text transmission, creating a computer model that expresses these phenomena as robustly as possible while representing them in a way that renders them tractable to large-scale data analysis. Our model consists of an overlaid pair of graphs that indicate both the horizontal sequence of readings within manuscript witnesses to the text and the vertical relationships between readings across the different witnesses. In the same way, and for the same purpose, we have made formal and explicit the properties of a stemma hypothesis, as it has long been understood by philologists; formally speaking, a stemma is a directed acyclic graph with a single common root that indicates which manuscripts served as exemplars for others.

A great deal of work remains to be done for the creation of a full empirical profile of text variation. At present there are far too few digital diplomatic transcriptions of manuscript texts, and this remains in our view the primary stumbling block to a 'better' evidence-based stemmatology.

A secondary concern is the categorization system that we used for our analysis. These categories were chosen for their relative lack of ambiguity and the relative ease with which they can be assigned automatically, independent of meaning or context within the text. The results of our analysis showed that these categories were not particularly helpful in determining the empirical genealogical value of any given variant. A promising avenue for future research will be to attempt an answer to the question: are there common features of genealogical variation that can be detected so that these variants can be categorized and weighted accordingly?

Despite these obstacles and unanswered questions, our purpose here has been to show a way, methodologically, in which philologists can begin to work toward the construction of a full empirical and statistical model of text variation using the knowledge of true stemmatic relationships wherever that knowledge exists, and built on the analysis methods described here. The resulting empirical information—whether generally applicable for hand-copied texts or specific to a particular language, period, or type of text—could then be used as statistical data to refine existing methods for stemma construction or even give rise to methods as yet unrealized.

Acknowledgements

The authors thank Hendrik Blockeel, Marc Denecker, and Stef De Pooter for their invaluable collaboration on the graph analysis, as well as Jeroen Van Heel and Annika Asp for their work in tagging the Latin and Finnish traditions, and Helma Dik of the University of Chicago for her generous provision of the Perseus and PhiloLogic morphological data for Latin and Greek. We are also grateful to Joris van Zundert of Huygens ING and Troy Griffitts of the INTF in Münster for their collaboration on tools made available within the Stemmaweb site.

Funding

This work was supported by the Special Research Fund (BOF) of the KU Leuven, as CREA project # 3H100334.

References

Andrews, T. L. (2009). *Prolegomena to a Critical Edition of the Chronicle of Matthew of Edessa, with a Discussion of Computer-Aided Methods Used to Edit the Text*. Oxford: University of Oxford. http://ora.ouls.ox.ac.uk/objects/uuid%3A67ea947c-e3fc-4363-a289-c345e61eb2eb (accessed 12 June 2013).

Andrews, T. L. (2012a). *Stemmaweb - a Collection of Tools for Analysis of Collated Texts*. http://byzantini.st/stemmaweb/ (accessed 12 June 2013).

Andrews, T. L. (2012b). *Excerpts from the Chronicle of Matthew of Edessa (Matt'ēos Uṙhayec'i)* http://byzantini.st/ChronicleME/ (accessed 12 June 2013).

Andrews, T. L., Blockeel, H., Bogaerts, B. *et al.* (2012). Analyzing Manuscript Traditions Using Constraint-based Data Mining. In: *CoCoMile 2012 - COmbining COnstraint Solving with MIning and LEarning*. Montpellier, pp. 15–20. http://cocomile.disi.unitn.it/2012/papers/cocomile2012_manuscript.pdf (accessed 12 June 2013).

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

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

Boodts, S. (2012). A New Critical Edition of Augustine's Sermo 170. With a Tentative Analysis of the Stemmatic Position of the De Lapsu Mundi Collection. *Sacris Erudiri: a Journal on the Inheritance of Early and Medieval Christianity*, **50**: 185–225.

Boodts, S. and Partoens, G. (2011). The Manuscript Transmission of the De uerbis Apostoli Collection. State of the Art and New Perspectives. In: *Tractatio Scripturarum. Philological, Exegetical, Rhetorical, and Theological Studies on Augustine's Sermons*. Turnhout: Brepols, pp. 79–96.

Bryant, D. and Moulton, V. (2004). Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**(2): 255–65.

Dearing, V. A. (1974). *Principles and Practice of Textual Analysis*. Berkeley: University of California Press.

Dekker, R. H. and Middell, G. (2011). Computer-supported collation with CollateX: managing textual variance in an environment with varying requirements In: *Supporting Digital Humanities: Conference Proceedings*. Copenhagen. http://crdo.up.univ-aix.fr/SLDRdata/doc/show/copenhagen/SDH-2011/proceedings.html (accessed 12 June 2013).

Driscoll, M. J. (2010). The words on the page: thoughts on philology, old and new. In: *Creating the Medieval Saga: Versions, Variability, and Editorial Interpretations of Old Norse Saga Literature*. Odense: University Press of Southern Denmark, pp. 85–102.

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

Heikkilä, T. (2005). *Pyhän Henrikin Legenda*. Helsinki: Finnish Literature Society.

Howe, C. J., Barbrook, A. C., Mooney, L., and Robinson, P. (2004). Parallels Between Stemmatology and Phylogenetics. In: *Studies in Stemmatology II*. Amsterdam: Benjamins, pp. 3–11.

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**: 51–67.

Mink, G. (2004). Problems of a highly contaminated tradition: The new testament: stemmata of variants as a source of genealogy for witnesses. In: *Studies in Stemmatology II*. Amsterdam: Benjamins, pp. 127–43.

Reynolds, L. D. and Wilson, N. G. (1991). *Scribes and Scholars: A Guide to the Transmission of Greek and Latin Literature*. Oxford: Clarendon Press.

Robinson, P. and O'Hara, R. J. (1996). Cladistic Analysis of an Old Norse Manuscript Tradition. In Hockey, S. and Ide, N. (eds), *Research in Humanities Computing*, vol. 4. Oxford: Oxford University Press, pp. 115–37.

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): 417–433.

Roos, T. and Zou, Y. (2011). Analysis of textual variation by latent tree structures. In: *IEEE International Conference on Data Mining*. Vancouver. http://www.cs.helsinki.fi/u/ttonteri/pub/icdm2011.pdf (accessed 12 June 2013).

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. Nijmegen: Katholieke Universiteit Nijmegen. http://www.neder-l.nl/salemans/diss/salemans-diss-2000.pdf (accessed 12 June 2013).

Schmid, U. (2004). Genealogy by chance! On the significance of accidental variation (parallelisms). In: *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: *Studies in Stemmatology II*. Amsterdam: Benjamins, pp. 207–26.

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: *Studies in Stemmatology II*. Amsterdam: Benjamins, pp. 227–39.

Timpanaro, S. (2005). *The Genesis of Lachmann's Method*. Translated by Most, G. W. Chicago and London: University of Chicago Press. http://www.loc.gov/catdir/toc/ecip0513/2005015897.html (accessed 12 June 2013).

De Vos, I., Gielen, E., Macé, C., and Van Deun, P. (2010). La lettre B Du Florilège coislin: editio princeps. *Byzantion*, **80**: 72–120.

Wattel, E. and van Mulken, M. (1996a). Weighted Formal Support of a Pedigree. In: *Studies in Stemmatology*. Amsterdam, PA: Benjamins, pp. 135–68.

Wattel, E. and van Mulken, M. (1996b). Shock Waves in Text Traditions: Cardiograms of the Medieval Literature. In: *Studies in Stemmatology*. Amsterdam, PA: Benjamins, pp. 105–21.

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

Windram, H. F., Shaw, P., Robinson, P., and Howe, C. J. (2008). Dante's Monarchia as a Test Case for the Use of Phylogenetic Methods in Stemmatic Analysis. *Literary and Linguistic Computing*, **23**(4): 443–63.

Wittocx, J., Mariën, M., and Denecker, M. (2008). The IDP System: a model expansion system for an extension of classical logic. In: *Proceedings of the 2nd Workshop on Logic and Search*. Leuven: ACCO, pp. 153–165.

Notes

1 See (Salemans 2000) and (Spencer *et al.* 2004b) for opposing opinions.

2 The Heinrichi tradition could arguably be an exception; its creators hoped that the non-standard spelling of Old Finnish and the relative unfamiliarity of the language would simulate diachronic shifts. The situations, however, are not really the same, and the extremely high incidence of non-genealogical variation within Heinrichi compared with real medieval traditions may indicate that it is not a sufficient simulation.