Trees of Texts - Models and methods for an updated theory of medieval text stemmatology

Tara L. Andrews and Caroline Macé
Katholieke Universiteit Leuven

The construction of a stemma codicum from the manuscript witnesses to a given text remains a somewhat divisive issue in philology. To a classical philologist, it is a necessary first step toward creating a critical edition of the text; to a scholar who adheres to the principles of new philology, the assumptions inherent in the construction of a stemma are so fundamentally unsound that the exercise should simply not be attempted.

The advent of digital philology demands a fresh look at the practice of stemmatology; indeed there has been a great deal of work in the past two decades on the subject, and in particular on the applicability of similar work from the field of evolutionary biology on the reconstruction of phylogenetic relationships between species. More recent work has begun to overcome the limitations of these techniques when applied to manuscript texts, such as the biological assumption that no extant species is descended from another, or the assumption that any ancestor species has precisely two descendants, or the converse assumption that any descendant species has precisely one ancestor.

Despite this considerable methodological progress, the fundamental assumptions at the core of stemmatology have yet to be formalised, or even seriously questioned. Before an analysis even begins, variants are regularized, classified, and judged according to the sense of the editor as to whether they are likely to have genealogical significance. This process is a relic of the times before computer-assisted analysis was possible, and is designed to exclude large volumes of possibly-insignificant data that might obstruct a sensible result. Given the vast quantities of data that can now be produced about a set of texts, and given the speed with which computational analysis can be applied to large datasets, there should no longer be any rationale for such a priori exclusion of evidence.

The purpose of this paper is to revisit the assumptions and simplifications that have lain at the heart of stemmatology since the early nineteenth century; to propose a model for the representation of texts in all their variations that is tractable to computational analysis; to set forth

---

1 Maas, P., Textkritik, (Leipzig, 1960); Pasquali, G., Storia della tradizione e critica del testo (Firenze, 1971); West, M. L., Textual criticism and editorial technique: applicable to Greek and Latin texts (Stuttgart, 1973).


a method for the evaluation of stemma hypotheses with reference to text variants, and vice versa; and to discuss the means by which we can provide the field of stemmatology with a formal underpinning appropriate for the analytical abilities of the era in which we now live.

Modelling a text tradition
The first step to computational analysis of any text tradition is to model it programmatically. This act of object modelling is distinct from, and orthogonal to, text encoding - the component texts may reach us in a variety of forms, including TEI XML and its critical apparatus module, CSV-format alignment tables, a base text and a list of variants exported from Microsoft Word, or even individual text files that remain to be collated. All of these must be parsed into a uniform computational object model for further programmatic analysis.

The model we have devised is relatively simple, taking into account only those features of the text that are useful for computational purposes. We define a tradition as the union of all the versions of a text that exist, represented by its witnesses. A collation is the heart of the tradition, representing in a unified way the entire text with all its known variations. We have adapted the variant graph model proposed by Schmidt and Colomb\(^6\) for this purpose. In our model, a collation is a graph, or more precisely a superimposed set of graphs based on the same collection of vertices. Each vertex represents a reading, usually a word, within the text. The witness texts themselves are represented in the graph as a set of paths. Each witness path is directed and acyclic, taking a route through the reading vertices from beginning to end, according to the order of its text. The witness path may branch, if for example the text contains corrections by the original scribe(s) or by later hands; the paths and their possible branches are marked as edge attributes.

Taken together, these witness paths form the base graph layer of the collation. This graph may be trivially transformed into a directed acyclic graph for analysis purposes (through the separation of readings transposed between texts into multiple vertices), and may be equally trivially transformed into an alignment table, with transpositions handled either as separate rows in the table or as variant characters within the same row.

Another view of the graph concerns relationships between readings, or types of variation that occur within the graph. It is often useful to specify that two different readings are spelling or grammatical variants of the same word, or that they have some lexical relationship to each other e.g. synonyms or antonyms. Each relationship is an undirected edge between two (or more) readings, with the constraint that the related readings must all appear in different witnesses, as it is generally expected that related readings collate together. The relationship information allows association and regularization of reading data, producing thereby more exact versions of alignment tables for distance analysis or other statistical procedures.

Analysis of variants with arbitrarily complex stemmas
The goal of our research is to arrive at an empirical model of medieval text transmission, and set forth a formalized means of deducing copying relationships between texts. As such, we have two tasks before us. We must be able to evaluate a stemma hypothesis according to the variation present in a text tradition, and we must be able to evaluate variants in a tradition according to a given stemma hypothesis.

The core assumption of a stemma is that, for any variant that appears in more than one witness, it is either 'genealogical' - that is, the manuscript that carries this variant betrays information about its exemplar - or else it is 'coincidental', carrying no such information. Stemmata are constructed on the basis of identifying the 'genealogical' variation and using these to determine which manuscripts must have descended from which others.

The problem becomes computationally tractable once we realize that a stemma hypothesis is also a directed acyclic graph. No matter the complexity of the stemma, no matter the level of “horizontal transmission”, no matter the possibility of more than a single archetype, a stemma can be represented as a directed acyclic graph, where each extant or hypothesized manuscript is a vertex, and each edge describes a source relationship from one manuscript to another. In this paper we will describe the graph-theoretical methods we have developed to evaluate text variants as “genealogical” or not, according to a given stemma graph. Similar methods are applicable to unrooted bifurcating trees, as produced by cladistic methods of computational stemmatology; our methods make it straightforward to pinpoint the variant locations that have been implicitly rejected as “genealogical” for a given distance tree.

An empirical approach to text variation
The idea of distinguishing 'text-genealogical' variants from the coincidental sort is not new; with the exception of Schmid (2004), none of the approaches taken have been empirical, in accordance with text relationships known from outside the variant tables. Given the means to quickly evaluate a stemma in detail according to the variants present in the text, distinguishing those variants that follow the stemma from those that do not, we should now easily be able to arrive at an empirical classification of genealogically ‘significant’ and ‘insignificant’ variants. This threatens to create a problem of circular logic: how to retrieve the genealogical variants in the real world? We build hypotheses on the basis of variation; we accept or reject variants as ‘genealogical’ based on the hypothesis. How do we break the cycle?

We may adopt two approaches to anchor our methods in external results. First we test our methods on artificial traditions that have been created for the testing of stemmatological methods. These traditions are convenient in that the true stemmata are known, but they cannot reflect the true conditions under which medieval texts were copied. Thus we must also use ‘real’ text traditions, whose true stemmata are not fully known; in most traditions, however, some external information (e.g. colophon notations, physical damage or other physical characteristics of the texts) survives and can be used to establish or verify copying relationships.

Conclusion
When examining a medieval tradition, we can rarely be certain what the scribe has copied in a text, and what has changed coincidentally or independently. To date, stemmatic reasoning has relied on the question of whether the scholar finds it likely that a given variant was preserved in copies. Given the models and methods for analysis proposed herein, we have an opportunity to remove this limitation of ‘assumed likelihood’ and take all the evidence of a text into account, to build a statistical model independent of the constraints of evolutionary biology, using statistical probability rather than scholarly instinct alone.

---

7 Salemans (2000); also see papers by Wachtel, van Mulken, Schmid, Smelik, Schøsler, and Spencer et al. in P. van Reenen et al. (ed.), Studies in Stemmatology (Amsterdam, 2004).