Reconstructing Ancient Literary Texts from Noisy Manuscripts

Moshe Koppel
Dept. of Computer Science,
Bar Ilan University
Ramat-Gan, Israel
moishk@gmail.com

Moty Michaely
Dept. of Computer Science,
Bar Ilan University
Ramat-Gan, Israel
moty.mi@gmail.com

Alex Tal
Dept. of Jewish Thought,
University of Haifa
Haifa, Israel
msaltal@gmail.com

Abstract

Given multiple corrupted versions of the same text, as is common with ancient manuscripts, we wish to reconstruct the original text from which the extant corrupted versions were copied (typically via latent intermediary versions). This is a challenge of cardinal importance in the humanities. We use a variant of expectation-maximization (EM), to solve this problem. We prove the efficacy of our method on both synthetic and real-world data.

1 Introduction

In ancient times, original documents were written by hand and then copied by scribes. Some societies transmitted traditions orally, which were written down and copied at some later date. These copies were inevitably inexact, each scribe introducing some errors into the text. These flawed copies spread around the world where they were then themselves imperfectly copied. Some small subset of these repeatedly corrupted documents survived until modern times.

One of the main tasks of the study of such ancient manuscripts is to reconstruct the original document (the "ur-text") from the corrupted manuscripts that are available. This has traditionally been done using painstaking manual methods. In this paper, we show this reconstruction can be automated using a variant of the expectation-maximization (EM) algorithm (Dempster et al 1977).

The structure of the paper is as follows. In the next section, we consider previous work on the ur-text problem (mostly dealing with a different version of the problem). In Section 3, we define the synoptic form in which we assume the texts are presented. In Sections 4 and 5, we formalize the problem and present our solution. In the three subsequent sections, we consider synthetic, artificial and real-world testbeds, respectively.

2 Previous Work

The reconstruction of ur-texts from corrupted manuscripts using manual methods has a long history (Maas 1958, West 1973). Such methods can be divided roughly into methods designed to select a single best manuscript (a "diplomatic" text) from among the extant ones (Bedier 1928) and methods designed to create an optimal hybrid (an "eclectic" text) out of the extant manuscripts (Lachmann 1853, Timpanaro 2005).

From a computational point of view, it is clear that the Bedierian approach is preferable when the collection of extant manuscripts for a given text is relatively complete (in the sense that the earlier manuscripts from which later manuscripts were copied are also included in the collection), especially if the ur-text itself might be found in the collection. In these cases, the main challenge is to re-

construct the *stemma*, the tree that records which manuscript was copied from which. The root of the reconstructed stemma is hypothesized to be the ur-text.

This challenge is common with bio-informatics (Pupko et al. 2000, Yang 2007) and researchers have applied methods of bio-informatics to the reconstruction of document stemmata (Robinson and O'Hara 1996, Robinson et al. 1998, Roos and Heikkila 2009, Roelli and Bachmann 2010, Andrews and Mace 2013).

Hoenen (2015) considers the problem of automated ur-text reconstruction for cases in which the manuscript collection is relatively complete and compares several methods of post-processing reconstructed stemmata to obtain (possibly eclectic) hypothesized ur-texts.

In the case of ancient documents, which we consider in this paper, the situation in which a collection is relatively complete – and might even include the ur-text – is exceedingly rare. Typically, the available manuscripts might be identifiable as (near or distant) cousins, but will be too sparse to permit even partial stemma reconstruction. Thus, we will develop an entirely new approach that does not focus on stemma reconstruction, as previous work did.

Our approach involves three stages. First, all the manuscripts for a given text must be arranged so that parallel words or phrases are aligned in columns ("synoptic form"). Second, when possible, related manuscripts should be clustered together. Finally, the ur-text can be inferred from the aligned, clustered texts by using statistical methods to make the optimal choice in each column of the synoptic text.

3 Creating a Synopsis

Consider the simple synopsis shown in Figure 1.

United	States	on	the	4th	of	July
USA		on		Fourth	of	July
United	States	in	the	end	of	June

Figure 1: A fragment of a synoptic text.

As is evident even in this simple example, there are a number of subtleties involved in creating such synopses. First, phrases (or any sequence of words that are inter-dependent) should ideally be in a single column, so that columns are as independent of each other as possible. For example, the phrase "United States" (and the acronym "USA") should be in a single column. (As in this example, typical available synopses are not ideal in this sense.) Second, words that differ only in trivial orthographic ways that are not important to us ought to be conflated. For example, we might choose not to distinguish between "4th" and "Fourth". Finally, distinct words that play the same role in the text (fourth/end; June/July) should be aligned, though often this is a matter of judgment.

One important limitation of such synopses is their monotonicity: the words in each row are laid out in the order they are found in the corresponding manuscript. Thus, if some manuscript inverts the order of two strings of text, one of those strings will not correspond to its parallels in other rows.

There have been efforts to automate the process of creating synopses from raw text. One approach adapts alignment methods developed in bio-informatics for aligning strings of DNA (Notredame et al. 2002). However, since the "words" aligned in bio-informatics are chosen from a small alphabet, whereas the words in texts are chosen from a large lexicon, such adaptation is not straightforward. There also are alignment methods designed specifically for text alignment (Robinson 1989, Spencer and Howe 2004, Dekker et al. 2014), as well as methods for aligning parallel texts in multiple languages (Och and Ney 2003).

Existing methods for text alignment are adequate for our purposes. As it happens, for the testbeds considered in this paper, manual synopses were available, allowing us to focus on the more basic issue of ur-text reconstruction.

4 Formalizing the Problem

Suppose now that we have a synopsis of n manuscripts each of which makes some choice with regard to each of m words (tokens) each appearing in a different column. We can think of our synopsis as an m*n matrix $a = \{a_{ij}\}$, where a_{ij} is the word (form) in the j^{th} column according to the i^{th} manuscript. (Some of these words might be blanks, which we treat exactly like any other token.) Given such a synopsis, we wish to choose the most prob-

able choice in each column. The resulting sequence of words is the proposed ur-text.

How can we determine the most probable choice in each column? A straightforward baseline solution is to use simple majority rule (SMR): for each column, choose the token found most frequently in that column. Under certain trivial conditions, Condorcet's Jury Theorem guarantees that this method's accuracy approaches 1 as the number of manuscripts grows.

In real-life, however, the number of manuscripts available is usually quite limited. We will introduce a method that yields considerably stronger results than SMR for the kinds of situations encountered in the real world.

We will assume that each manuscript i has some reliability level, p_i. This means that for any given token, manuscript i has probability p_i of choosing the right token (that is, using the same form that is being transcribed). Of course, the value p_i is not known to us. Our objective will be to show how to simultaneously find the most likely reliability levels of the respective manuscripts and the most likely ur-text.

Our initial generative model is as follows: a single ur-text of length m is copied by each of n scribes. For any token $j \in \{1,\ldots,m\}$, there is probability p_i that the scribe of manuscript $i \in \{1,\ldots,n\}$ will transcribe the token correctly. If he fails to transcribe a token correctly, there are k_j equiprobable potential distinct forms other than the original. (Note that the number of potential forms might be different for different tokens.) One limitation of this generative model is that we assume, perhaps unrealistically, that for any given scribe the probability of an error is the same for every word.

We do not assume that for all i, $p_i > .5$. Rather, for the binary case ($k_j=1$), we assume only the almost trivial condition that $\prod p_i > \prod(1-p_i)$; for non-binary cases, the necessary condition is even weaker. Note that for the binary case, the necessary condition is weaker in the limit than the necessary condition for Condorcet's Jury Theorem (Berend and Paroush 1998): $\lim_{n\to\infty} \text{average}(p_i-1/2)\sqrt{n} = \infty$.

Thus our synopsis $a=\{a_{ij}\}$ is such that each column has at most k_j+1 distinct choices: one correct choice and k_j equiprobable potential alternative forms. We arbitrarily map each choice to a number in the set $\{1,\ldots,k_j+1\}$.

An ur-text reconstruction is a mapping from the synopsis $a=\{a_{ij}\}$ to a proposed text in $\{1,\ldots,k_j+1\}^m$. Our objective is to find an optimal reconstruction, given no information other than the synopsis $a = \{a_{ij}\}$.

5 Our Proposed Method

We treat our problem as an instance of judgment aggregation in which each of a set of judges (manuscripts) makes judgments regarding multiple issues (words). This problem has been handled (Baharad et al 2011, Bachrach et al 2012, Hovy et al. 2013) using variants of EM; we adapt this approach here for our purposes.

In principle, given the set of scribal reliabilities $\{p_i\}_i$ and the probabilities in each column of each distinct form being the correct (original) form, $\{p(t_j=w|w\in\{1,\ldots,k_j+1\})\}_j$ (or for short, $\{p(t_j=w)\}_j$) we could compute the conditional probability of obtaining the synopsis a. Thus, given some synopsis a, optimality is obtained by the values of $\{p_i\}_i$ and $\{p(t_j=w)\}_j$ that maximize the likelihood of a. As shown below, the values $\{p(t_j=w)\}_j$ can be determined from a and $\{p_i\}_i$. Thus, denoting by $p(a;\{p_i\})$ the likelihood of a given the parameters $\{p_i\}_i$, our objective is to maximize $p(a;\{p_i\})$.

Our algorithm, which we'll call UR, finds a local maximum for $p(a;\{p_i\})$ as follows. First, for each token j, we estimate the value of k_j by simply assuming that every distinct form with non-zero probability actually occurs in one of the manuscripts (i.e., k_j is one less than the number of distinct forms that appear in the column). Clearly, this estimate is only plausible when the number of manuscripts is large, but we find that it is good enough for our purposes.

We assign some initial constant value to $\{p_i\}_i$. Then we repeat the following two steps until convergence:

1. Use the manuscript reliabilities $\{p_i\}_i$ to recompute the values $\{p(t_j=w)\}_j$.

2. Use the values $\{p(t_j=w)\}_j$ to estimate the maximum likelihood values of the manuscript reliabilities $\{p_i\}_i$.

For the first step, we assume that for every j the prior $\{p(t_j=w\}$ is equal for every $w\in\{1,\ldots,k_j+1\}$.

Then we have by Bayes' rule that for each j and each $w \in \{1,\ldots,k_j+1\}$,

$p(t_j=w|\ a) = p(t_j=w|\ a_j) = \frac{p(a_j|t_j=w)}{Z}$ where a_j is the j^{th} column of a and Z is a normalization factor. This can easily be computed by substituting

$p(a_j|t_j=w) = \prod_{a_{ij}=w} p_i * \prod_{a_{ij}\neq w}(1-p_i)/k_j$.

For the second step, we compare the values $\{p(t_j=w)\}_j$ to the judgments of individual i, in order to compute the maximum-likelihood values of $\{p_i\}_i$. Specifically, the maximum likelihood value of p_i is equal to the average (over j) probability that $a_{ij} = t_j$. Thus, our updated value of $p_i = \frac{1}{m}\left(\sum_j p(t_j = a_{ij}|a)\right)$.

It can be shown that the method converges to a local maximum of $p(a;\{p_i\})$.

5.1 Handling Dependencies

The above method would guarantee a (locally) optimal solution if it were the case that manuscripts are independent of each other. In fact, however, manuscripts are copied from one another, so that various extant manuscripts might have some common ancestor (subsequent to the ur-text). Thus, for example, third-generation manuscripts fall naturally into clusters, reflecting the second-order manuscript from which they are copied. The errors in manuscripts in the same cluster tend to be similar.

Even when, as is usually the case, we don't have a sufficiently complete collection of manuscripts to reconstruct a stemma, we might have enough (possibly external) information to at least divide the collection into several flat clusters of related manuscripts. Given such a clustering, we can use the UR method to identify the ur-text for each cluster and then use UR once again to reconstruct the original ur-text from the multiple second-generation ur-texts.

In most real life cases, domain experts are able to identify flat clusters (but not a full stemma) using external evidence. In cases where the clusters are not known, automatic clustering methods must be used to identify them.

6 Experiments – Synthetic Synopses

6.1 Direct Transcription

We first test our method on synthetic manuscripts. For our initial experiment, we assume that there is a single ur-text T consisting of m words. T is copied directly by each of n scribes. (In these experiments, we use n=20.) Each manuscript is assigned some random reliability p_i (the probability of copying a given word correctly) chosen from a uniform distribution between 0.20 and 0.99. In addition, for each word w_j, we let k_j have equal chances of being either 1 or 2. If a word is copied incorrectly, it is randomly replaced by one of k_j possible other words.

For each trial, we use the method described above to reconstruct the ur-text. As a baseline method, we use simple majority rule (SMR) to decide which word to choose in each column of the synopsis.

We run 1000 trials as described above, showing results for different manuscript lengths. For each algorithm in each trial, we check the proportion of words that are reconstructed correctly and we average the results over all trials. In Figure 2, we show the results.

UR improves initially as manuscript length increases since its estimates of manuscript reliability improve, while SMR is indifferent to manuscript length.

In Figure 3, we show results for the same setup where document length is fixed at 100 but the number of manuscripts varies. UR clearly outperforms the baseline simple majority rule.

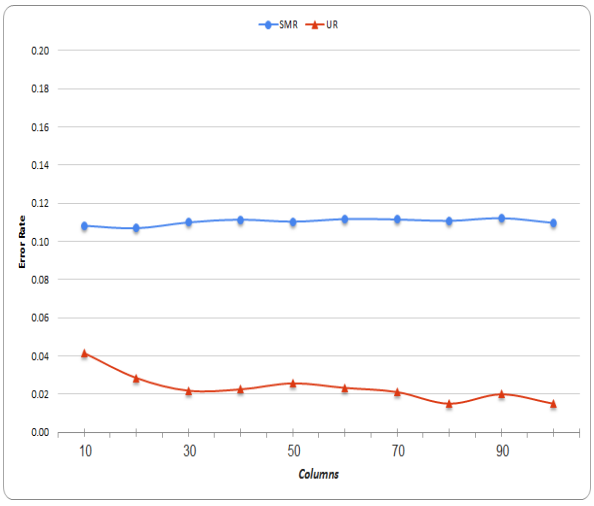

Figure 2: Word error rate in reconstruction using UR and SMR, respectively, for varying manuscript lengths.

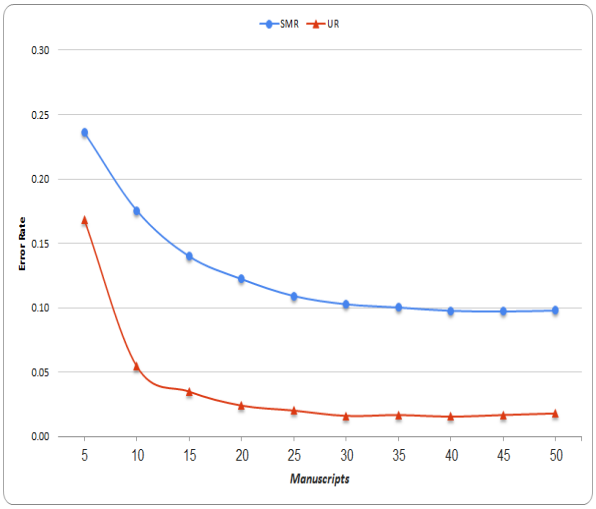

Figure 3: Word error rate in reconstruction using UR and SMR, respectively, for varying numbers of manuscripts.

6.2 Latent Manuscripts

For our next set of experiments, we drop the assumption that all extant manuscripts are copied directly from the ur-text. Instead, we assume that our manuscripts are copies of copies. We generate 20 second-generation manuscripts by noisily copying the ur-text *T* 20 times, exactly as above. Now we generate 200 third-generation manuscripts, each time randomly choosing one of the second-generation manuscripts and copying it noisily according to some randomly-chosen reliability (from the same distribution as above). These 200 third-generation manuscripts serve as input.

We call the second-generation manuscripts to which we do not have access "latent" manuscripts and we call the set of third-generation manuscripts that are generated from a given second-generation manuscript a "cluster". In these experiments, we assume that the clusters are known.

For each trial, we use each of the following algorithms for regenerating the ur-text:

1. SMR
2. UR
3. Recursive SMR
4. Recursive UR

The recursive methods run the algorithm on each cluster separately and then again on the results of the respective clusters.

In Figure 4, we show accuracy results averaged over 1000 trials as described above, showing results for different manuscript lengths. We find that UR that ignores clustering performs very poorly but Recursive UR is much stronger, outperforming both versions of SMR. (For all data-points, standard error is <.005, too small to be seen.)

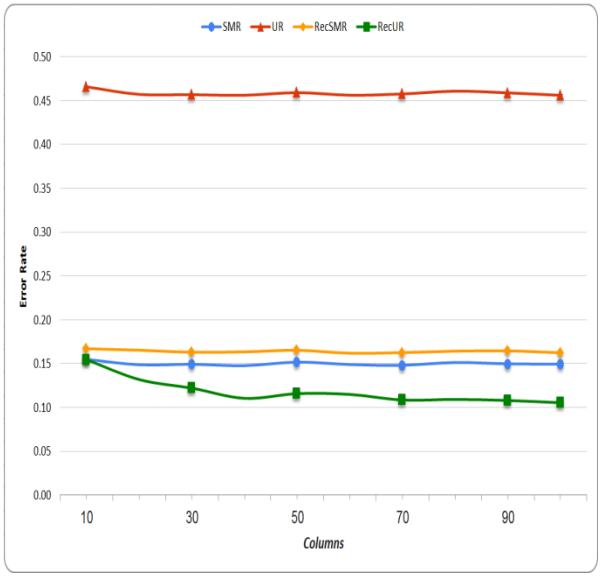

Figure 4: Word error rate in reconstruction using UR and SMR, respectively, for varying manuscript lengths, with and without clustering of manuscripts.

7 An Artificial Manuscript Testbed

As noted above, our method is appropriate for cases in which only a fraction of the manuscripts in the stemma are extant. In cases where the bulk of the stemma – possibly including the ur-text itself –

is extant, it would be better to attempt to reconstruct the stemma and identify the actual ur-text.

Notre Besoin (Baret et al 2006) is an artificial collection of manuscripts generated by having "scribes" successively copy an Old French manuscript. Thirteen manuscripts of length 1020 were generated in this fashion. The full set of manuscripts (including the ur-text itself) was used as a basis for comparing several methods for stemma reconstruction (Roos and Heikkila 2009) and ur-text reconstruction (Hoenen 2014).

Although this is a situation in which we regard our method as less appropriate than stemma reconstruction methods, we run it for comparison purposes.

Hoenen found that an automated method (PAML) for stemma reconstruction (Yang 2007), yields an ur-text with word error rate of 4.7% and that post-processing the obtained stemma using a method akin to Recursive SMR lowers the word error rate to 4.1%. We find that applying Recursive UR to three non-hierarchical clusters – the descendants of the three highest-level non-root nodes in the stemma reconstructed by PAML (provided by A. Hoenen) – while ignoring all other information in the stemma, yields an ur-text word error rate of 4.6%. Thus the complete stemma reconstruction offers no clear benefit beyond the shallow clustering method for our purposes.

8 A Real-World Manuscript Testbed

Finally, we consider a real-world example. The Babylonian Talmud is a 6th century Aramaic compendium transmitted orally and written down several centuries later in Hebrew letters in Iraq. We use a synoptic version of a single chapter of the Talmud (the second chapter of Tractate Beitzah), consisting of 8564 columns and 20 manuscripts, seven of which are relatively complete and the rest of which are very fragmentary. A domain expert established that, based on external evidence, the manuscripts split naturally into six identifiable clusters (containing 8, 4, 3, 3, 1, and 1 manuscripts, respectively).

Several pre-processing steps are applied to the raw synopsis. First, we automatically identify minor orthographic variants within a given column and standardize them so that they are treated as identical. Furthermore, since the raw synopsis includes single words, rather than phrases, in a given column, there are many dependencies among consecutive columns. To eliminate the most egregious such dependencies, we iteratively conflate to a single column all perfectly correlated consecutive columns. After conflating dependent columns in this way, we remain with 5912 columns. In Figure 5, we show the proportion of these columns contain a single form, two variant forms, and so on. As can be seen, only for 17% of the columns do all manuscripts agree.

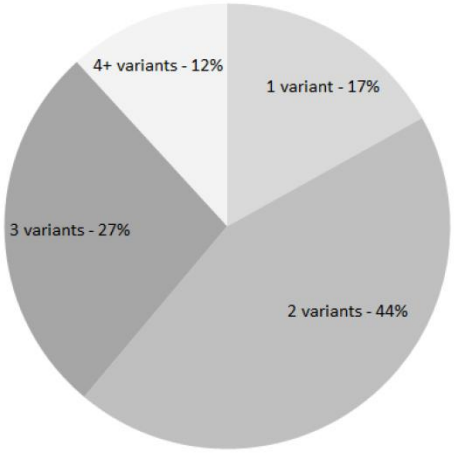

Figure 5: The proportion of columns in the Beitzah corpus containing a given number of word forms

We apply Recursive UR, as well as Recursive SMR as a baseline method, to the processed synopsis. Recursive UR assigns the six clusters reliabilities ranging from 0.46 to 0.78, with the highest reliability assigned to a cluster consisting of a single manuscript indeed considered to be particularly ancient and trustworthy.

The two methods disagree for 448 of the columns and agree for the rest. Our domain expert (who did not know which word choice came from which method) provided the most likely correct word according to his own judgment for those columns for which the two methods disagree. Of the 448 disagreements, he determined that 80 were significant and resolvable. In 66 of these 80 cases (82.5%), the expert's judgment coincided with the form chosen by UR and in only 14 cases (17.5%), his judgment coincided with SMR.

9 Conclusions

We have found that ur-texts can be reconstructed using automated methods far more effectively than using a simple majority rule. Furthermore, this can be done to some extent even using only manuscripts from the third-generation and later.

We have assumed that the correct clustering of manuscripts is known. Left for future work is the case in which the clusters are identified using automated clustering methods.

More importantly, perhaps, we have assumed throughout that a given manuscript has some fixed reliability over all words. In fact, it might be the case that reliability varies over different tokens (or types) and that, moreover, not each distinct form of a word is an equally probable alternative to the original. Graphical models could be used to generalize our approach to handle such cases.

References

Andrews, TL and C. Mace. 2013. Beyond the tree of texts: Building an empirical model of scribal variation through graph analysis of texts and stemmas. *Literary and Linguistic Computing*, 28(4):504–521.

Bachrach, Y., T Graepel, T Minka, J Guiver (2012). How To Grade a Test Without Knowing the Answers--A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing*, Proceedings of ICML.*

Baharad, E., Goldberger, J., Koppel, M. and Nitzan, S. (2011), Distilling the Wisdom of Crowds: Weighted Aggregation of Decisions on Multiple Issues, *JAAMAS* 22(1), 31-42.

Baret, P., Macé, C. and Robinson, P. (2006), Testing methods on an artificially created textual tradition, in C. Mace, P. Baret, A. Bozzi, L. Cignoni (eds.), *Linguistica Computazionale. The evolution of texts: confronting stemmato-logical and genetical methods*, XXIV-XXV, Pisa-Roma, Istituti Editoriali e Poligrafici Internazionali, pp. 255-283.

Bedier, J (1928). *La tradition manuscrite du 'Lai de l'Ombre': Reflexions sur l'Art d'Editer les Anciens Textes*. Romania, 394:161–196, 321–356.

Berend, Daniel and Paroush, Jacob (1998). When is Condorcet's Jury Theorem valid?. *Social Choice and Welfare* 15 (4)

Dekker, R., D. van Hulle, G. Middell, V. Neyt, J. van Zundert (2014). Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project, *LLC: Digital Scholarship in the Humanities* 25. 452-470

Dempster, A. P., N. M. Laird and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*. Series B 39(1): 1-38.

Hoenen, A (2015). Lachmannian Archetype Reconstruction for Ancient Manuscript Corpora. *HLT-NAACL* 2015: 1209-1214

Hovy, D., Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy (2013*).* Learning Whom to Trust with MACE. *Proceedings of NAACL-HLT* 2013

Lachmann, K (1853). *In T. Lucretii Cari De rerum natura libros commentarius*. Georg Reimer.

Maas, P. (1958). *Textual Criticism (tr. B. Flower)*, Oxford

Notredame, C. (2002) Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, 3, 131–144.

Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics* 29 (1), 19-51.

Pupko, T., Itsik Pe'er, Ron Shamir, and Dan Graur. 2000. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol*., 17(6):890–896.

Robinson, P. M. W. (1989). The Collation and Textual Criticism of Icelandic Manuscripts (1): Collation. *Literary and Linguistic Computing* 4(2), 99-105.

Robinson, P. and R. J. O'Hara. 1996. Cladistic Analysis of an Old Norse Manuscript Tradition. *Research in Humanities Computing* 4.

Robinson, P., A. Barbrook, N. Blake, and C. Howe. 1998. The Phylogeny of The Canterbury Tales. *Nature,* 394:839.

Roelli, P. and Dieter Bachmann. 2010. Towards Generating a Stemma of Complicated Manuscript Traditions: Petrus Alfonsis Dialogus, *Revue d'histoire des textes*, 5(4):307–321.

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

Spencer, M. & Howe, C. J (2004). Collating Texts Using Progressive Multiple Alignment. *Computers and the Humanities* 38, 253-270.

Timpanaro, S (2005). *The Genesis of Lachmann's Method*, (ed. and trans. Glenn W. Most), U. of Chicago Press

West, ML (1973). *Textual Criticism and Editorial Technique*, Stuttgart

Yang, Ziheng. 2007. *PAML 4: phylogenetic analysis by maximum likelihood.* Mol. Biol. Evol., 24(8):1586–1591.