Analysis of Textual Variation by Latent Tree Structures

Teemu Roos and Yuan Zou
Helsinki Institute for Information Technology, HIIT
P.O. Box 68, FI-00014 University of Helsinki, Finland
Email: firstname.lastname@cs.helsinki.fi

***Abstract*—We introduce Semstem, a new method for the reconstruction of so called stemmatic trees, i.e., trees encoding the copying relationships among a set of textual variants. Our method is based on a structural expectation-maximization (structural EM) algorithm. It is the first computer-based method able to estimate general latent tree structures, unlike earlier methods that are usually restricted to bifurcating trees where all the extant texts are placed in the leaf nodes. We present experiments on two well known benchmark data sets, showing that the new method outperforms current state-of-the-art both in terms of a numerical score as well as interpretability.**

***Keywords*-graphical models, latent trees, EM algorithm, stemmatology, textual criticism.**

I. Introduction

In the popular game known as the broken telephone (or Chinese whispers, *Le téléphone arabe*, etc.), a message is successively whispered by one player to another until it reaches all the players. The message typically gets distorted along the way, which makes the game amusing. The accumulation of small changes characteristic to the game is also one of the defining features of evolution. A phenomenon that is perhaps lesser known, but even more fitting as an analogue of the broken telephone process, is encountered in *textual criticism* where texts distorted by transcriptional errors and other changes are reconstructed by identifying such changes and removing them, see [5].

The traditional goal of textual criticism is the reconstruction of the original, or at least the most recent common ancestor of the extant manuscripts.Often the reconstruction is preceded by *stemmatological analysis*, i.e, building a diagram known as a *stemma*, encoding the usually tree-like copying relationships of manuscripts. In biological terms, this corresponds to *phylogenetic analysis* wherein different species are organized in a so called Tree of Life.[1]

The adoption of computational methods in textual criticism, and in humanities at large, is still in its infancy. The current applications mainly involve digitized sources, databases, multimedia, and other relatively "mundane" tools.[2] In contrast, the methodology of the biological sciences has been utterly transformed by mathematical and computational methods; indeed so much that it is now regarded as a new discipline, bioinformatics. It remains to be seen to which extent a similar transformation will take place in the emerging field of *digital humanities.*

Since the processes of textual variation resemble those of biological evolution, it is natural to attempt stemmatological analysis by phylogenetic methods. Indeed this has turned out to be a very successful approach, see e.g. [16], [18], [21]. A plethora of phylogenetic methods are available; for general overviews, see e.g. [3], [8], [20]. The methods can be roughly categorized as distance-matrix based methods (e.g. [19]), parsimony methods (e.g. [7]), and model based methods. The latter group, which is of our primary interest, includes methods based on maximum likelihood [25] and Bayesian inference [17], [27].

However, certain assumptions that are often valid in the biological domain, are problematic when phylogenetic methods are applied to manuscripts. Two such assumptions are: (i) all extant (observed) nodes are always placed in the leaf nodes of the tree, and (ii), all trees are *bifurcating*, i.e., all interior nodes have degree three (one parent, two children). Neither assumption is valid in stemmatology: it is *not* the case that none of the extant manuscripts are ancestors of some other extant manuscripts, and furthermore, it is *not* true that the number of copies made of each manuscript is either two or nil.

Among phylogenetic methods, a method proposed by Friedman et al. [11], called SEMPHY, is particularly relevant. It is based on the structural EM algorithm, proposed earlier by Friedman [10]. The method constructs phylogenetic trees that are essentially tree-structured Bayesian networks with N observed nodes and $N-2$ latent (unobserved) nodes—precisely the kind of bifurcating tree structures that are ubiquitous in phylogenetics. The algorithm is based on alternating between a message passing phase where the distribution of the latent nodes is inferred based on the current structure (the E-step), and building a new tree structure based on the observed nodes and the inferred distribution of the latent nodes [4] (the M-step).

In order to guarantee that the resulting tree is bifurcating and that all the observed nodes are leafs, SEMPHY includes an additional step where the tree obtained in the M-step, which may violate these restrictions, is converted into an equivalent tree with the desired properties. However, from

[1] See, for instance, the Tree of Life web project at tolweb.org.

[2] We do not intend to play down the importance of such tools but to emphasize that their role is somewhat peripheral in the actual scholarly work, compared to their role in the natural sciences.

the stemmatological point of view, there is nothing wrong with multifurcating trees with observed interior nodes. On the contrary, stemmatology proves to be an ideal application for a structural EM approach that, unless specific manipulations are carried out, produces general latent tree structures that are free from the two restrictions mentioned above.

We adapt the structural EM algorithm for stemmatology by applying a model for textual variation, and omitting the aforementioned transformation step. The resulting method is, to our knowledge, the first automatic method for discovering unrestricted tree-shaped structures from textual variants. We demonstrate that our adapted algorithm is able to reconstruct the copying relationships of several manuscripts created by copying texts by hand. The resulting stemmata are more accurate and easier to interpret than traditional trees based on phylogenetic methods. Further applications may include the analysis of computer viruses [26], plagiarism detection [12], and content-based social network analysis [23].

The rest of the paper is organized as follows: In Sec. II we review the model-based approach to phylogenetic and stemmatic analysis. In Sec. III, we describe the structural EM algorithm in detail. In Secs. IV–V, we describe the data and the experimental set-up, and the results. Conclusions and pointers for future work are outlined in Sec. VI.

II. Modeling Textual Evolution

Model-based phylogenetic analysis is preceded by a model specification wherein we construct a probabilistic model describing the evolution of the biological units (individuals or species) under study. The data on which the analysis is based typically consists of genomic sequences. It is usually assumed that the *sites* (positions) in the sequences evolve independently, although this assumption is not strictly speaking biologically valid.[3] Many of the popular evolutionary models can be represented as continuous-time Markov chains (CTMCs) [13]. Such models are characterized by a parametric transition probability matrix, which in the case of DNA sequences can be expressed in the form

$$P(t) = \begin{pmatrix} p_{A\to A}(t) & p_{G\to A}(t) & p_{C\to A}(t) & p_{T\to A}(t) \\ p_{A\to G}(t) & p_{G\to G}(t) & p_{C\to G}(t) & p_{T\to G}(t) \\ p_{A\to C}(t) & p_{G\to C}(t) & p_{C\to C}(t) & p_{T\to C}(t) \\ p_{A\to T}(t) & p_{G\to T}(t) & p_{C\to T}(t) & p_{T\to T}(t) \end{pmatrix},$$

with the interpretation that $p_{x\to y}(t)$ is the probability that a site in state x evolves into state y in time t. There are various ways to define the transition probabilities,[4] and similar models exist for protein sequences.

[3]Models based on more realistic evolutionary assumptions have also been proposed, see e.g. [2], but despite recent advances, their application is still prohibitively inefficient.

[4]Here we use the term *transition probability* to refer to all transitions. The established convention in bioinformatics is to call the probabilities in the top-left and bottom-right 2×2 submatrices transition probabilities, and the remaining ones *translation* probabilities.

The process is in an equilibrium when the state composition of each site is given by the stationary distribution (p_A, p_G, p_C, p_T) for which we have

$$p_x = \sum_{y\in\Sigma} p_y p_{y\to x}(t),$$

for all $x \in \Sigma = \{A, G, C, T\}$ and all $t \geq 0$. Furthermore, the process is said to be time-reversible if

$$p_x p_{x\to y}(t) = p_y p_{y\to x}(t), \tag{1}$$

for all $(x, y) \in \Sigma^2, t \geq 0$.

The models for textual evolution are much less established as those for genomic evolution. The evolution of words can be modeled similarly, although there the above assumptions are even less realistic. Nonetheless, the approach has been shown to be fruitful (proving once again the fact about some models being "wrong" but "useful"). In stemmatology, the time variable, t, does not have a similar role as in biological evolution. Namely, the existing manuscripts may remain in the "stemmatic pool" (akin to the so called *genetic pool*) and can be used as sources for copying after an arbitrarily long time, and there is in principle no reason to assume that a copy made of an older source manuscript should contain more errors than a copy made of a more recent manuscript. Another major difference in modeling text compared to genomic sequences is that the alphabet is not fixed, although in practice, it seems safe to restrict the readings in each site r to the set of readings observed in at least one of the extant manuscripts, $\Sigma^{(r)}$.

For simplicity, we let the diagonal elements of transition matrix for site r to be the same, which is $1 - \alpha$. Thus any other element is $\alpha/(k_r - 1)$ with $0 < \alpha < 1$, and k_r denotes the number of observed unique readings in site r. Hence, each word has the same probability, $1 - \alpha$, of staying unchanged when it is copied, and the probability of the word being changed to another is uniform. We have also experimented with models where the transition probabilities reflect word similarities but the uniform model appears to be more robust in all its simplicity. The probability of change can also be estimated together with the tree structure. However, for simplicity, we assume in this work that $1 - \alpha = 0.95$. In our experiments, the results obtained by estimating $1-\alpha$ or using other constants within the range $[0.8, 1.0)$ results in qualitatively similar results.

The corresponding stationary distribution is easily seen to be uniform, i.e., $p_x^{(r)} = 1/k_r$ for all $x \in \Sigma^{(r)}$. This also implies that the model is time-reversible, i.e.,

$$p_x^{(r)} p_{x\to y}^{(r)} = 1/k_r \cdot \alpha/(k_r - 1) = p_y^{(r)} p_{y\to x}^{(r)}, \tag{2}$$

for all $(x, y) \in {\Sigma^{(r)}}^2, x \neq y$; the case $x = y$ is trivially symmetric.

III. Structural EM

The EM algorithm [6] is an extremely popular technique for dealing with missing data. Its main use is parameter estimation. However, it can also be used for learning the structure of a Bayesian network, as demonstrated by the structural EM algorithm [10]. Unlike most structure learning methods, it is applicable when some of the data are missing or when some of the variables are completely unobserved. The expectation (E) step in the algorithm performs inference on the missing data to obtain suitable statistics, that can be used in the maximization (M) step to construct a model structure. The new structure is then used for obtaining another (better) set of statistics in the next iteration. A phylogenetic method based on the structural EM algorithm, called SEMPHY, has also been presented [11], where the unobserved ancestral sequences are represented as latent variables, and the learned structure is constrained to be a tree.

In this section, we adapt the phylogenetic structural EM method for stemmatology. We start by discussing the relatively straightforward complete data case. We then resort to the structural EM approach for dealing with latent variables and missing data. For the most part, we follow [11].

A. Probability of Stemmatic Trees

Let a stemma, T, be defined as a set of edges, $(i,j) \in \{1,\ldots,N+M\}^2$, where $N+M$ is the number of nodes. We denote the nodes by $X_1,\ldots,X_{N+M}$. Nodes $X_1,\ldots,X_N$ are assumed to be observed. The remaining ones are latent nodes that correspond to undiscovered manuscripts. For the sake of clarity, in the following we do not consider partially observed manuscripts, although they can be handled in a straightforward way using exactly the same structural EM approach. The algorithm we have implemented handles them, and the experimental results in Sec. V address both kinds of missing data.

In the ideal case, when $M = 0$, i.e., we have the complete set of the manuscripts, the probability of the data given a stemma T is easily computed as

$$P_T(X_1,\ldots,X_N) = \prod_{r=1}^{n}\left[P(X_1^{(r)})\prod_{i=2}^{N}P(X_i^{(r)}|X_{\Pi_i}^{(r)})\right], \tag{3}$$

where the number of sites (words) is n,[5] the parent of node $i > 1$ is denoted by Π_i, and we assume without loss of generality that the root node is X_1.

[5]In order to get the data into the format where each manuscript has the same number of sites, and the same site in different manuscripts corresponds to the readings of the same word (if it exists in the given manuscript) requires that the texts be *aligned*. There are various methods that are commonly used in bioinformatics. We apply similar methods but do not discuss the details due to space restrictions.

We re-write Eq. (3) as

$$P_T(X_1,\ldots,X_N) = \prod_{r=1}^{n}\left[\prod_i P(X_i^{(r)})\prod_{(i,j)\in T}\frac{P(X_i^{(r)}|X_j^{(r)})}{P(X_i^{(r)})}\right]. \tag{4}$$

Due to the fact that the model is time-reversible, we have

$$\frac{P(X_i^{(r)}|X_j^{(r)})}{P(X_i^{(r)})} = \frac{P(X_j^{(r)}|X_i^{(r)})}{P(X_j^{(r)})}, \tag{5}$$

which implies that the formula in Eq. (4) is invariant under changing the root variable and reordering all edges to point away from it. Consequently, unless we have prior information about the ordering of the nodes (in the form of, e.g., timings of the manuscripts), different stemmata with the same undirected structure, or *skeleton*, have the same posterior probability.[6]

We consider the logarithm of the likelihood, L_T, and decompose it into a more liable form as follows

$$\begin{aligned} L_T(X_1,\ldots,X_N) &= \log\prod_{r=1}^{n}P_T(X_1,\ldots,X_{N+M}) \\ &= \sum_{r=1}^{n}\left[\sum_{i=1}^{N}\log P(X_i^{(r)}) + \sum_{(i,j)\in T}\log\frac{P(X_i^{(r)}|X_j^{(r)})}{P(X_i^{(r)})}\right]. \end{aligned} \tag{6}$$

The first sum inside the brackets is a constant independent of the stemma, and can therefore be ignored. The latter sum can be written as

$$\sum_{(i,j)\in T}\sum_{(x,y)\in\Sigma^{(r)2}}1\{X_i^{(r)}=x, X_j^{(r)}=y\}\log\frac{p_{x\to y}}{p_y}, \tag{7}$$

where $1\{X_i^{(r)}=x, X_j^{(r)}=y\}$ is the indicator function that takes value one if the argument is true, and zero otherwise.

Since the log-likelihood decomposes as a sum of terms for different edges in the stemma, we can actually maximize the likelihood by casting the problem as a maximum spanning tree problem. The weights, $w_{i,j}$, of each pair of nodes are

$$w_{i,j} = \sum_{r=1}^{n}\sum_{(x,y)\in\Sigma^{(r)2}}1\{X_i^{(r)}=x, X_j^{(r)}=y\}\log\frac{p_{x\to y}}{p_y}, \tag{8}$$

which are symmetric, $w_{i,j} = w_{j,i}$, by the time-reversibility property Eq. (2). For instance, Kruskal's algorithm finds the maximum spanning tree in time $O(N\log N)$. The above procedure amounts to the popular Chow-Liu algorithm [4].

B. Expected Log-Likelihood

The above complete-data case needs to be extended to handle missing data when some of manuscripts are expected

[6]In phylogenetics, this problem is often solved by adding a so called *outgroup* species in the data that is known to be outside the group of species under study. In the case of texts, no such outgroup really exists.

to be lost, namely $M > 0$. Obviously, the actual number of missing manuscripts is very hard, or impossible, to know in advance. Hence, the used number will have to be an educated guess, at best. We follow the convention, originating from phylogenetics, of using $M = N-2$ latent nodes. As it turns out, superfluous latent nodes tend to end up as extra leaf nodes or as sequences of degree-two nodes, both of which can be pruned out without changing the tree topology in any meaningful way.

Consider the conditional distribution of the latent nodes, $X_{N+1}, \ldots, X_{N+M}$, given the observed nodes, $X_1, \ldots, X_N$, and a fixed tree structure, T_t. We let $Q(T : T_t)$ denote the expected log-likelihood of an arbitrary tree structure:

$$Q(T : T_t) = \mathbb{E}[L_T(X_1, \ldots, X_{N+M}) \mid X_1, \ldots, X_N, T_t]. \tag{9}$$

As noted above, the first term inside the brackets in the log-likelihood, Eq. (6), can be omitted as a constant independent of the tree topology. Obviously, the expectation of a constant is a constant as well, and we are left with

$$\sum_{r=1}^{n} \sum_{\substack{(i,j)\in T \\ (x,y)\in\Sigma^{(r)2}}} \mathbb{E}\left[1\{X_i^{(r)} = x, X_j^{(r)} = y\} \log \frac{p_{x\to y}}{p_y} \,\middle|\, Z_t\right], \tag{10}$$

where $Z_t = (X_1, \ldots, X_n, T_t)$, which is easily seen to be equal to

$$\sum_{r=1}^{n} \sum_{\substack{(i,j)\in T \\ (x,y)\in\Sigma^{(r)2}}} \eta(x,y) \log \frac{p_{x\to y}}{p_y}, \tag{11}$$

where

$$\eta(x,y) = P(X_i^{(r)} = x, X_j^{(r)} = y \mid X_1, \ldots, X_N, T_t) \tag{12}$$

denotes the conditional expectation of the indicator function in Eq. (10).

Analogous to the complete-data case of the previous subsection, we now define the weight of a potential edge between nodes i and j as

$$w_{i,j} = \sum_{r=1}^{n} \sum_{(x,y)\in\Sigma^{(r)2}} \eta(x,y) \log \frac{p_{x\to y}}{p_y}, \tag{13}$$

where it is important to note that $\eta(x,y)$ in Eq. (12) depends on the structure T_t as well as the observed nodes $X_1, \ldots, X_N$. Since it is actually a pairwise conditional probability of two nodes taking the values x and y, respectively, it can be evaluated using standard inference algorithms. Furthermore, since the network topology is assumed to be a tree, the classical message passing (belief propagation) algorithm is exact [15]. The weights of all pairs of nodes can be computed in time $O(nN^2|\Sigma_{\max}|^2)$, where $|\Sigma_{\max}|$ denotes the greatest number of variant readings in any given site. We omit further details; see [11].

To warm-start the structural EM, we initialize the tree by the neighbor joining method [19]. The algorithm is run until the expected log-likelihood converges or a maximum number of iterations is reached. In the end, the tree with the highest expected log-likelihood is returned. Pseudo-code for the procedure, which we call Semstem, is given in Algorithm 1.

Algorithm 1: Semstem

```
begin
    initialize T_0 using NJ method;
    let T_max = T_0;
    let Q_max = Q(T_0 : T_0);
    let t = 0;
    repeat
        E-step: compute the weights w_{i,j} for all pairs
        of nodes, i, j under tree T_t;
        M-step: find a new tree T_{t+1} by the MST
        algorithm;
        if Q(T_{t+1} : T_t) > Q_max then
            let T_max = T_{t+1};
            let Q_max = Q(T_{t+1} : T_t);
        let t = t + 1;
    until T_{t+1} = T_t or t > t_max ;
    return T_max;
end
```

C. Local Optima

As the usual (parametric) EM algorithm, structural EM is a greedy method where the expected log-likelihood is never decreased. However, EM tends to get stuck to local optima. To alleviate this problem, Friedman et al. [11] propose to apply a technique similar to simulated annealing. The idea is to add stochastic perturbations to the weights. The magnitude of the perturbations is gradually decreased by adjusting a 'temperature' parameter $\sigma_t \to 0$ as t grows.

We add Gaussian noise with variance σ_t^2 to the elements of the weight matrix:

$$\widetilde{w}_{i,j}(t_{i,j}) = w_{i,j}(t_{i,j}) + \epsilon_{i,j}, \tag{14}$$

To maintain the symmetry of the weight matrix, we let $\epsilon_{i,j} = \epsilon_{j,i}$. The temperature is decreased according to a geometric cooling schedule where $\sigma_{t+1} = \rho\sigma_t$ with $0 < \rho < 1$. In the final stage, σ is set to zero to allow the algorithm to converge to a local optimum. In practice, this happens within a couple of dozen iterations at most.

IV. Data and Experimental Set-Up

To illustrate the method, and to compare its performance against a set of state-of-the-art algorithms applied in stemmatology, we use two artificially generated textual traditions. One could also generate data from a model that produces random copying errors but it is generally believed that

manually created data sets are much closer to real-world textual traditions. The first data set, *Parzival* (see [21]), contains 21 manuscripts of length 1055 words including gaps created by multiple alignment. The second data set, *Notre Besoin de Consolation est Impossible à Rassarier* (*'Notre Besoin'* for short; see [1]) contains 14 manuscripts of length 1035 words. The *Notre Besoin* tradition includes an instance of *contamination*, i.e., a node that has more than one parent. Such cases arise when two or (rarely) more manuscripts are consulted when creating a new copy.

We evaluate the methods based on their success of finding a stemma that is close to the truth. Note that we are comparing two arbitrary latent tree structures, and hence, the usual accuracy measures such as counting the number of shared edges, etc., do not apply. The main problem is that we cannot establish a one-to-one correspondence between the latent nodes in the true stemma and the estimated one—there is no guarantee that even their number will be the same. Instead, we employ the so called *average sign similarity* score that was introduced and used in an earlier benchmarking experiment [18].

To formally define the average sign similarity, let $d_{i,j}$ denote the length of the shortest path (number of edges) connecting nodes i and j in the true stemma, and let $d'_{i,j}$ denote the same for the estimated stemma. For any three distinct nodes i, j, k, we define the local score $u(i, j, k)$ as

$$\begin{cases} 1, & \text{if } \mathrm{sgn}(d_{i,j} - d_{i,k}) = \mathrm{sgn}(d'_{i,j} - d'_{i,k}), \\ 0, & \text{if } \mathrm{sgn}(d_{i,j} - d_{i,k}) = -\mathrm{sgn}(d'_{i,j} - d'_{i,k}), \\ 1/2, & \text{otherwise;} \end{cases} \quad (15)$$

where sgn takes values $-1, 0, +1$, respectively, when the argument is negative, zero, and positive. The average sign similarity score is the average of $u(i, j, k)$ over all distinct *observed* (i.e., not latent) nodes.

Briefly, the greater the score, the more similar the true and the estimated stemmata are, and vice versa. The fact that only triplets involving *observed* nodes are considered makes it possible to apply the average sign similarity to stemmata with different numbers of latent nodes. Furthermore, since the distances are defined in terms of the *shortest* path connecting two nodes, the stemmata need not be tree-shaped—hence, the case of contamination in the *Notre Besoin* data set poses no problems.

In order to investigate how the amount of available data affects the performance of the considered methods, we create subsets by randomly removing complete manuscripts as well as parts thereof from the remaining ones. We first remove $10\%, 20\%, 30\%$ or 40% of the nodes, and then, for each of the remaining manuscript independently, delete $0\%, 10\%, \ldots, 90\%$ of the text in one or more contiguous randomly selected segments. Each combination of the above percentages is repeated 100 times with a new random seed, and a statistical test (Wilcoxon signed rank test) is performed to assess significance.

Other method included in the comparison are neighbor-joining (NJ) [19], least-squares (LS), maximum parsimony [7], all three from the PAUP* package [24], maximum likelihood (ML) [25] from the Phylip package [9], and the RHM method that has been specifically designed for stemmatology [18]. The default settings are used for each algorithm. RHM requires that the number of iterations be specified: we use 25000 in each run which is computationally feasible but usually guarantees convergence to the same solution in multiple repeated runs in the used data sets.

In an earlier comparison on a set of benchmarks, including the two data-sets we are using, maximum parsimony and RHM were found to perform consistently well [18]. The earlier comparison was based on particular subsets of the data with a certain number of missing manuscripts and certain deletions in some of the manuscripts, without randomization and repetitions. Consequently, the conclusions in the earlier comparison were not validated by statistical tests.

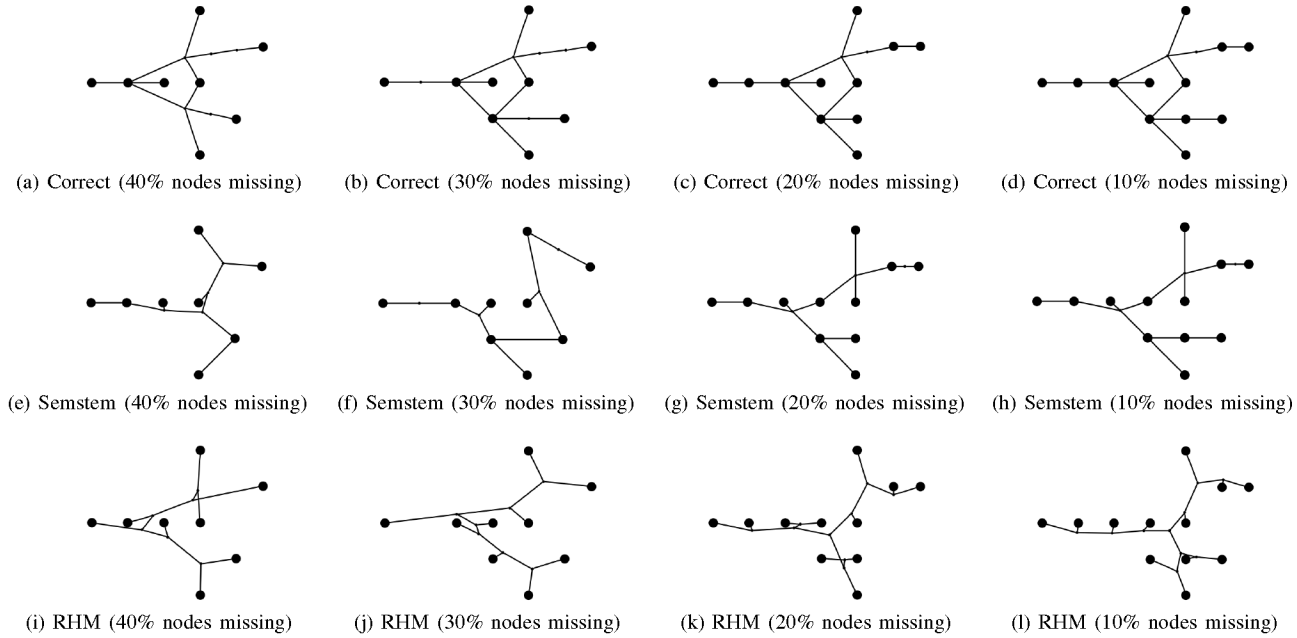
V. Results

To get an idea of the learning task, consider Figs. 1 and 2. They illustrate the original structure and the learned trees by Semstem and RHM for different amount of remaining data. For the plots, we chose RHM since it was found to be consistenly good in earlier experiments [18] as well as ours (see below). The results obtained by other methods such as maximum parsimony were visually similar to those of RHM. In the figures, the positions of observed nodes are fixed to be the same in each graph in order to facilitate comparison. The hidden nodes are placed so as to appropriately show the structure. As mentioned above, RHM as well as all the other methods are only capable of creating trees where the observed nodes are positioned as leafs of the tree. This causes problems for interpreting the resulting trees: especially in the case of *Parzival* (Fig. 2), the stemmata obtained by Semstem are more easily interpreted than the bifurcating trees obtained by RHM and the other methods.

To assess the scores of the methods, Fig. 3 gives the the average sign similarity scores of different methods when the number of missing nodes and the amount of missing text is varied. Tables I and II show numerical results. The highest scoring in each case is highlighted. Statistically significant differences are indicated (see the table caption for details). Semstem outperforms other methods in most cases, achieving in some cases scores as high as 80 %, while other methods typically yield significantly lower scores.

VI. Conclusions and Future Work

We presented a new method for discovering latent tree structures for the analysis of textual variation. Unlike earlier methods, which typically produce bifurcating trees, our method is able to produce unrestricted tree structures where the observed texts can be located either as internal nodes or

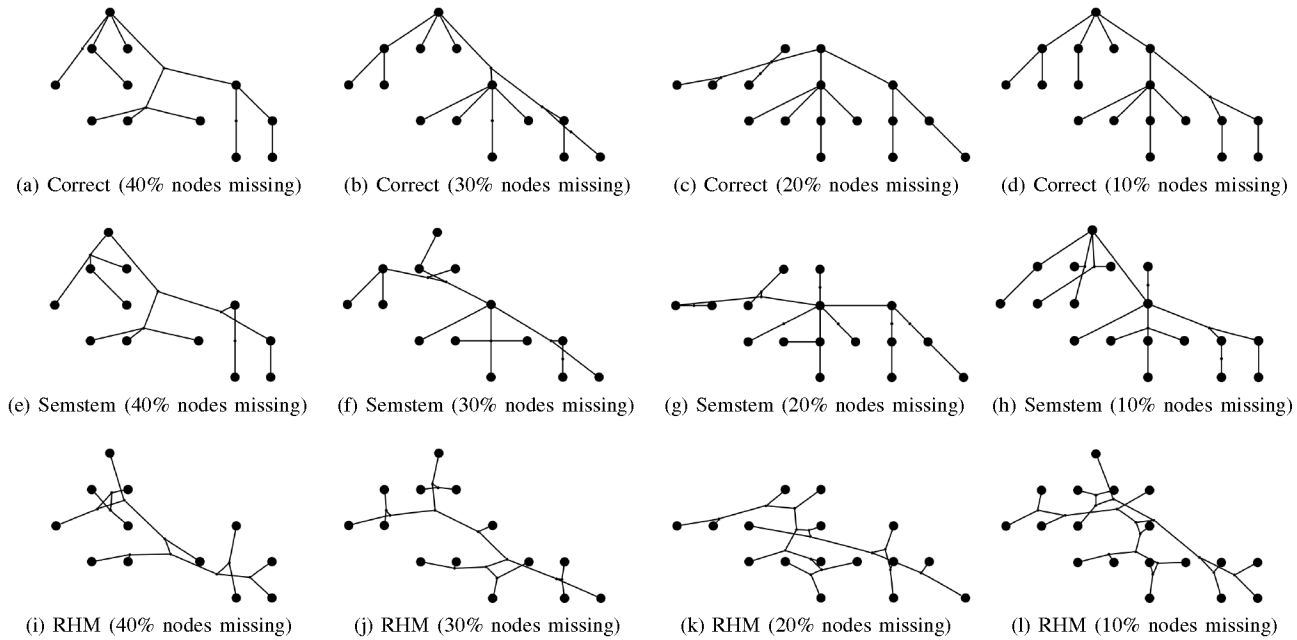


(a) Correct (40% nodes missing) (b) Correct (30% nodes missing) (c) Correct (20% nodes missing) (d) Correct (10% nodes missing)

(e) Semstem (40% nodes missing) (f) Semstem (30% nodes missing) (g) Semstem (20% nodes missing) (h) Semstem (10% nodes missing)

(i) RHM (40% nodes missing) (j) RHM (30% nodes missing) (k) RHM (20% nodes missing) (l) RHM (10% nodes missing)

Figure 1: True stemmata (*a–d*), and trees learned by Semstem (*e–h*) and RHM (*i–l*) for *Notre Besoin* tradition with 40–10 % missing nodes. The amount of missing text in each retained node was 30%.

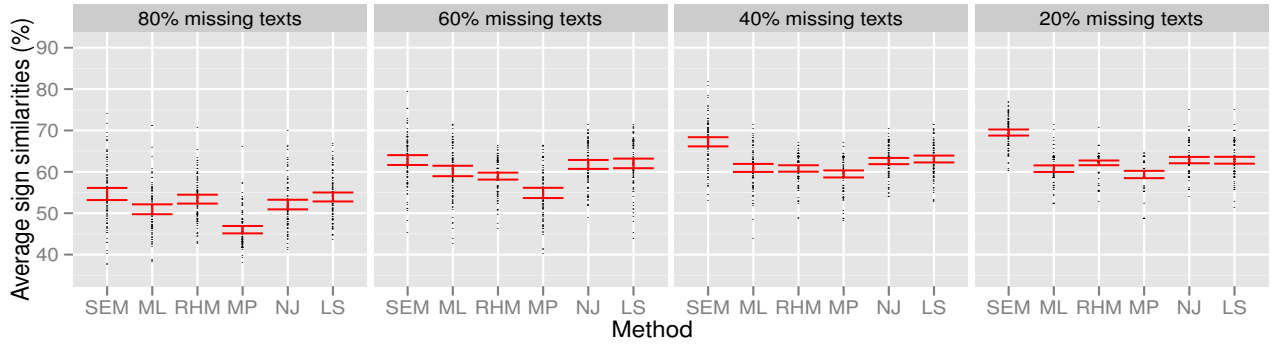


(a) Correct (40% nodes missing) (b) Correct (30% nodes missing) (c) Correct (20% nodes missing) (d) Correct (10% nodes missing)

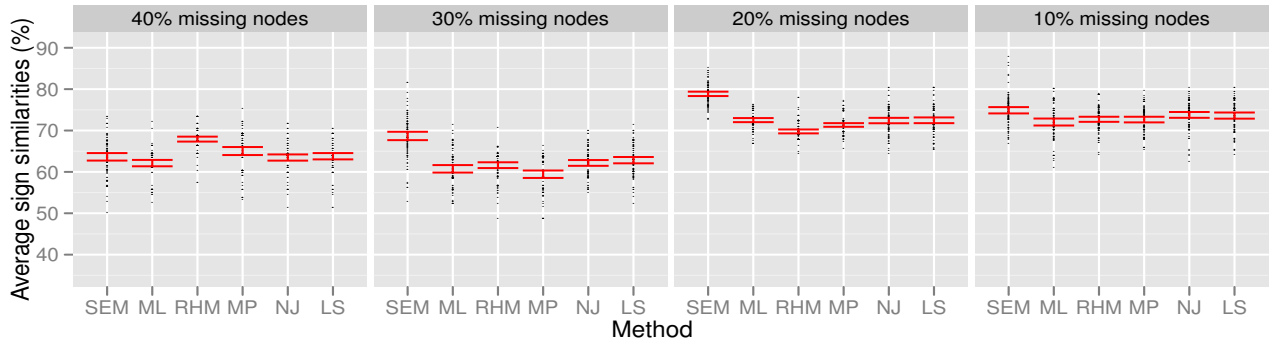
(e) Semstem (40% nodes missing) (f) Semstem (30% nodes missing) (g) Semstem (20% nodes missing) (h) Semstem (10% nodes missing)

(i) RHM (40% nodes missing) (j) RHM (30% nodes missing) (k) RHM (20% nodes missing) (l) RHM (10% nodes missing)

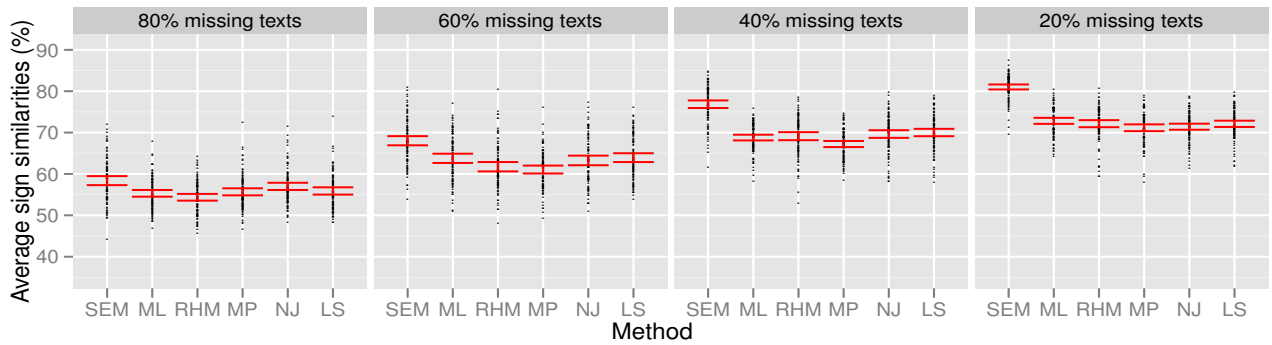
Figure 2: True stemmata (*a–d*), and trees learned by Semstem (*e–h*) and RHM (*i–l*) for *Parzival* tradition with 40–10 % missing nodes. The amount of missing text in each retained node was 30%.



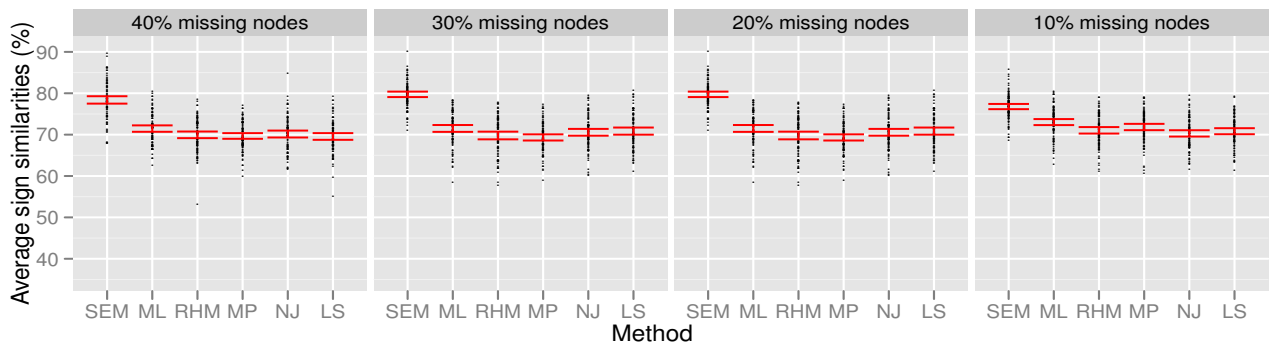
(a) *Notre Besoin*: 30 % missing nodes, 20–80 % missing text



(b) *Notre Besoin*: 10–40 % missing nodes, 30 % missing text



(c) *Parzival*: 30% missing nodes, 20–80 % missing text



(d) *Parzival*: 10–40 % missing nodes, 30% missing text

Figure 3: Scores of different methods with different amounts of missing nodes and text. The boxplot shows the interquartile range in 100 repetitions with different randomly removed subsets of the data. (SEM = Semstem; MP = maximum parsimony; for the other acronyms, see Sec. IV.)

leafs. Another advantage is that the degree of the internal nodes is not limited. These two aspects make the results much easier to interpret. Empirical experiments involving two artificially created manuscript collections demonstrate that the new method achieves higher scores than the compared methods representing the current state-of-the-art.

Future work includes studying the scaling properties of the new method when the size of the data sets is increased. The new artificial data sets created recently in the stemmatology community should provide an ideal basis for this. Studies using simulated data with varying parameters for the transition model describing the copying errors will complement such an investigation. Furthermore, it will be interesting to develop more refined models to be used as a basis of the method. Of particular interest are asymmetric models that could be used for identifying the orientation of the edges, and hence, the root of the stemma.

ACKNOWLEDGMENT

The authors thank the anonymous referees for thoughtful comments that have improved the paper. This work was supported in part by the University of Helsinki Research Funds (project STAM), the Finnish Cultural Foundation (Studia Stemmatologica Workshop), The Academy of Finland (project MODEST), and the European Union Network of Excellence PASCAL.

REFERENCES

[1] P. V. Baret, C. Macé. and P Robinson, "Testing Methods on an Artificially Created Textual Tradition," Linguistica computazionale, pp. 255–281, 2006.

[2] A. Bouchard-Côté, M. Jordan, and D. Klein, "Efficient inference in phylogenetic InDel trees," Advances in Neural Inf Proc Syst (NIPS 2008), pp. 177–184, 2008

[3] L. L. Cavalli-Sforza and A. W. F. Edwards, "Phylogenetic analysis—Models and estimation procedures," Am. J. Hum. Genet., vol. 19, pp. 233–257, 1967

[4] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," IEEE Trans. Inf. Theory, vol. 14, pp. 462–467, 1968

[5] J. Delz, Textkritik und Editionstechnik, in F. Graf (Ed.), Einleitung in die lateinische Philologie, B. G. Teubner, 1997

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J Royal Stat. Soc. Ser. B, vol. 39, pp. 1–38, 1977

[7] A. W. F. Edwards and L.L. Cavalli-Sforza, "The reconstruction of evolution," Ann. Hum. Genet., vol. 27, pp. 105–106, 1963

[8] J. Felsenstein, "Numerical methods for inferring evolutionary trees," Quart. Rev. Biol., vol. 57, pp. 379–404, 1982

[9] J. Felsenstein, "PHYLIP-Phylogeny Inference Package (version 3.2)," Cladistics, vol. 5, pp. 164–166, 1989

[10] N. Friedman, "Learning belief networks in the presence of missing values and hidden variables," Proc. ICML 1997, pp. 125–133, 1997

[11] N. Friedman, M. Ninio, I. Pe'er, and T. Pupko, "A structural EM algorithm for phylogenetic inference," J. Comput. Biol., vol. 9, pp. 331–354, 2002

[12] C. Liu, C. Chen, J. Han, and P. S. Yu, "GPLAG: Detection of software plagiarism by program dependence graph analysis," Proc. KDD 2006, pp. 872–881, 2006

[13] B. Mau, M. A. Newton, and B. Larget, "Bayesian Phylogenetic Inference via Markov chain Monte Carlo methods," Biometrics, vol. 55, pp. 1–12, 1999

[14] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 2nd edition, McGraw-Hill, 1984

[15] J. Pearl, "Reverend Bayes on inference engines: A distributed hierarchical approach," Proc. AAAI 1982, pp. 133–136, 1982

[16] P. Robinson and R. J. O'Hara, "Report on the textual criticism challenge 1991," Bryn Mawr Class. Rev., vol. 3, pp. 331–337, 1992

[17] F. Ronquist and J. P. Huelsenbeck, "MRBAYES 3: Bayesian phylogenetic inference under mixed models," Bioinformatics, vol. 19, pp. 1572–1574, 2003

[18] T. Roos and T. Heikkilä, "Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets," Lit Linguist Comput., vol. 24, pp. 417–433, 2009

[19] N. Saitou and M. Nei, "The neighbor-joining method: A new method for reconstructing phylogenetic trees," Mol Biol Evol, vol. 4, pp. 406–425, 1987

[20] C. Semple and M. A. Steel, Phylogenetics, Oxford University Press, 2003

[21] M. Spencer, E. A. Davidson, A. C. Barbrook, and C. J. Howe, "Phylogenetics of artificial manuscripts," J. Theor. Biol., vol. 227, pp. 503–511, 2004

[22] M. Spencer and C. J. Howe, "How accurate were scribes? A mathematical model," Lit Linguist Comput., vol. 17, pp. 311–322, 2002

[23] J. Sun, S. Papadimitriou, C.-Y. Lin, N. Cao, S. Liu, W. Qian, "MultiVis: Content-based social network exploration through multi-way visual analysis," Proc. SDM 2009, pp. 1063–1074, 2009

[24] D. L. Swofford and D. P. Begle, PAUP: Phylogenetic analysis using parsimony. Version 3.1. User's manual, Smithsonian Institution, Laboratory of Molecular Systematics, 1993

[25] E. A. Thompson, Human Evolutionary Trees, Cambridge University Press, 1975

[26] S. Wehner, "Analyzing worms and network traffic using compression," J. Comp. Secur., vol. 15, pp. 303–320, 2007

[27] Z. Yang and B. Rannala, "Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method," Mol. Biol. Evol. vol. 14, pp. 717–724, 1997

Table I: Means of average sign similarities (%) of different methods for the *Notre Besoin* data set. Results that are better than the others are shown in bold-face, with statistical significance indicated by: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Data set		Method					
Nodes missing (%)	Text missing (%)	Semstem	ML	RHM	MP	NJ	LS
40	90	51.2	47.2	**59.0*****	51.9	47.2	53.9
	80	54.3	52.7	**63.1*****	53.7	52.3	55.6
	70	56.6	54.6	**64.6*****	56.2	55.2	57.6
	60	59.2	58.1	**65.7*****	60.5	58.7	59.8
	50	62.8	61.3	**67.4*****	63.5	61.9	62.0
	40	62.9	62.5	**67.9*****	65.2	62.8	63.1
	30	63.6	62.1	**67.9*****	65.0	63.5	63.8
	20	65.0	62.2	**68.0****	66.4	63.9	63.7
	10	65.1	62.3	67.7	**68.1*****	64.5	64.5
30	90	51.4	49.2	51.3	43.9	49.7	**52.4**
	80	**54.6**	51.0	53.4	46.0	52.1	53.9
	70	**59.0****	55.3	56.2	50.0	56.5	57.2
	60	**62.8**	60.2	59.0	54.9	61.8	62.0
	50	**65.2*****	63.0	61.0	58.8	62.7	62.7
	40	**67.3*****	61.0	60.8	59.5	62.6	63.1
	30	**68.7*****	60.7	61.6	59.4	62.2	62.8
	20	**69.5*****	60.8	62.2	59.4	62.8	62.8
	10	**70.5*****	61.2	61.7	61.3	62.2	63.6
20	90	60.7	58.0	**62.1***	55.5	58.3	55.3
	80	59.7	58.3	**61.3***	55.4	58.8	55.8
	70	63.8	63.7	**65.3***	60.5	63.6	60.6
	60	**70.0***	68.9	68.6	66.5	67.2	66.5
	50	**75.6*****	71.0	69.0	68.8	71.4	69.8
	40	**77.8*****	71.9	70.0	70.7	72.4	71.9
	30	**78.9*****	72.5	69.8	71.3	72.4	72.5
	20	**79.6*****	72.5	70.0	71.6	72.2	72.1
	10	**81.9*****	73.2	70.4	71.6	74.4	73.6
10	90	59.6	58.8	**60.4**	54.6	58.1	54.5
	80	59.0	58.2	**61.6****	55.7	59.3	56.0
	70	63.7	65.6	**67.3****	62.7	64.8	62.4
	60	**69.1**	68.8	68.7	66.6	67.5	66.6
	50	**71.8***	70.0	70.9	70.1	70.9	69.6
	40	**73.7***	71.7	71.7	70.5	72.5	72.6
	30	**74.9***	72.1	72.7	72.6	73.8	73.6
	20	**76.5****	72.9	72.7	73.1	74.1	74.7
	10	**77.0**	73.3	72.4	73.2	76.4	76.6

Table II: Means of average sign similarities (%) of different methods for the *Parzival* data set. Results that are better than the others are shown in bold-face, with statistical significance indicated by: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Data set		Method					
Nodes missing (%)	Text missing (%)	Semstem	ML	RHM	MP	NJ	LS
40	90	58.1	55.4	53.6	**58.4**	57.0	53.8
	80	**59.0**	56.6	54.0	57.1	57.8	54.5
	70	**62.7**	61.5	57.4	60.4	60.6	58.4
	60	**68.7***	66.9	61.4	64.5	65.5	65.3
	50	**73.3*****	69.5	65.6	68.2	66.8	67.2
	40	**75.3*****	71.0	67.4	69.1	68.8	68.5
	30	**78.4*****	71.5	70.0	69.7	70.1	69.6
	20	**78.0*****	71.6	72.3	70.0	69.9	69.9
	10	**78.8*****	73.3	73.8	70.7	70.9	70.6
30	90	**57.0****	53.8	52.7	55.4	55.8	53.8
	80	**58.4***	55.3	54.4	55.7	57.0	55.9
	70	**61.9****	58.9	57.9	57.8	60.0	59.5
	60	**68.0*****	63.8	61.7	61.1	63.3	63.9
	50	**73.7*****	67.8	66.6	65.1	67.6	68.0
	40	**76.9*****	68.8	69.1	67.2	69.7	70.0
	30	**79.7*****	71.5	69.8	69.3	70.6	70.9
	20	**81.0*****	72.8	72.1	71.2	71.4	72.1
	10	**81.7*****	75.9	73.5	73.3	71.8	73.1
20	90.0	**56.9****	54.8	53.1	52.5	55.0	54.4
	80	**58.8**	58.6	55.9	55.6	57.1	56.5
	70	**62.2**	**62.2**	58.7	61.0	60.1	60.0
	60	**67.3**	**67.3**	63.4	66.5	61.9	63.5
	50	70.2	**70.9**	67.3	69.8	65.9	67.7
	40	**74.6**	73.9	68.9	72.9	68.6	69.5
	30	75.5	**75.6**	71.1	74.4	71.2	71.5
	20	76.8	**76.9**	72.3	76.2	72.5	72.9
	10	77.8	**79.6**	74.5	79.3	74.7	74.9
10	90	**56.2***	54.6	52.7	54.8	55.2	53.4
	80	**58.0***	57.1	55.9	55.4	57.2	56.0
	70	**62.9***	61.5	59.2	60.3	59.7	61.3
	60	**67.8****	66.0	63.9	66.1	63.7	64.6
	50	**72.6*****	70.1	67.0	69.0	66.4	67.8
	40	**75.6*****	71.2	68.4	69.8	68.0	69.5
	30	**76.8*****	73.0	71.0	71.8	70.3	70.8
	20	**78.5*****	73.9	72.7	73.8	70.9	71.3
	10	**79.3*****	76.0	73.6	75.3	72.4	72.7