Multi Modal Distance - An Approach to Stemma Generation with Weighting

Armin Hoenen
CEDIFOR
Goethe University Frankfurt
hoenen@em.uni-frankfurt.de

Abstract

Stemma generation can be understood as a task where an original manuscript M gets copied and copies – due to the manual mode of copying – vary from each other and from M. Copies $M_1, .., M_k$ which survive historical loss serve as input to a mapping process estimating a directed acyclic graph (tree) which is the most likely representation of their copy history. One can first tokenize and align the texts of $M_1, .., M_k$ and then produce a pairwise distance matrix between them. From this, one can finally derive a tree with various methods, for instance Neighbor-Joining (NJ) (Saitou and Nei, 1987). For computing those matrices, previous research has applied unweighted approaches to token similarity (implicitly interpreting each token pair as a binary observation: identical or different), see Mooney et al. (2003). The effects of weighting have then been investigated and Spencer et al. (2004b) found them to be small in their (not necessarily all) scenario(s). The present approach goes beyond the token level and instead of a binary comparison uses a distance model on the basis of psycholinguistically gained distance matrices of letters in three modalities: vision, audition and motorics. Results indicate that this type of weighting have positive effects on stemma generation.

Keywords: stemmatology, weighted distance, multi modal

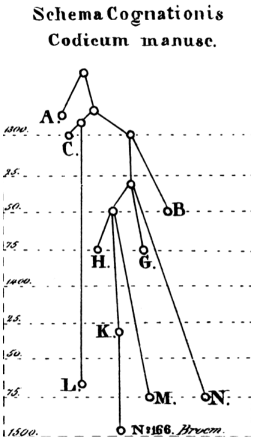


Figure 1: First modern stemma by Schlyter, 1827, from O'Hara (1996) with texts= nodes, copy processes= edges.

1. Introduction

Stemma generation is the process of determining the most likely tree[1] with manuscripts being represented by nodes in the usually directed acyclic graph[2] and edges representing copy processes or chains of such, see Figure 1. Since the late 1950ies computational methods have been applied to stemmatological tasks (Ellison, 1957) and in the last decade evaluation against benchmark datasets (or artificial traditions) has been conducted (Baret et al., 2004; Spencer et al., 2004a; Roos and Heikkilä, 2009; Hoenen, 2015a). These datasets have been generated by first giving one text (root) to volunteers to be handcopied (or dictated). Its copies have then been handcopied again and so forth. The true *vorlage*[3]-copy relations (edges in the true stemma) have been recorded by the authors, so that we know the entire true tree with all edges and the position of root. Those manuscript texts have been digitized and manually aligned for a *stemmatology challenge* (Roos and Heikkilä, 2009) where from a subset of manuscripts several teams attempted to reconstruct – manually or automatically – the true tree.[4] In this paper, these datasets are taken as basis to a new method which uses external data in the form of psycholinguistically generated letter and phoneme distance matrices in order to a) generate and evaluate stemmata and b) assess how large the influence of low level perceptual processes is. From the alignments of the artificial traditions, pairwise distance matrices of the single manuscripts (texts, nodes) are built. Each manuscript pair is compared tokenwise using some metric resulting in an overall distance. This metric can be described as weighted, where the external data serves for determining the weights. Concerning token comparison, philology describes a whole range of types of variation and their implications, see e.g. (Roelli and Macé, 2015; Andrews and Macé, 2013). Philologically motivated classification has been used for weighting token pairs upon distance computation. Categories such as "Word variant, changes meaning" or "Word change affecting rhyme" (Mooney et al., 2003, p.287) have been applied. A stemmatologically relevant distinction and driving force behind the will to weight variants is that between genealogically informative and accidental variation (Andrews and Macé, 2013). The implication is that some innovations in the text induced

[1] Insights and historical considerations on the tree as a preferred technical model for stemmatology and concurrent graph theoretical considerations are discussed for instance in Hoenen et al. (2017), Flight (1994) and sources therein.

[2] Note, that we target stemmata for closed traditions (Pasquali and Pieraccioni, 1952) that have no multiple originals, as is probable for orally transmitted epics (Lord, 1960). Hoenen (2017) has attempted to reconcile tree and network perspectives on stemmata.

[3] Vorlage is a German loan, which is used in philology to describe the model or the original of a copy.

[4] The webpage of the challenge is `https://www.cs.helsinki.fi/u/ttonteri/casc/`. Here, the challenge datasets can be obtained. For the current study the authors provided the full datasets additionally via direct contact.

by copying are idiosyncratic and hardly revertable, for instance when some non syntactically crucial word is accidentally left out: *this is a really big challenge* → *this is a big challenge* or when some content word gets replaced by one equally fitting into the context: *the clay dust shimmered* → *the day dust shimmered*. Such errors imply,[5] that the whole subbranch rooted by the manuscript having the new version at first will have it. In this way the innovation is genealogically informative. That is the information helps us locate the manuscript on the stemmatic tree, whereas other innovations could easily happen independently in different copy processes such as the introduction of punctuation at some point in time or some shift in definiteness *I heard the magpie* → *I heard a magpie*. Often variation can be multicausally explained and is analyzed on a case-by-case basis. One process which could be responsible for both kinds of innovations is the confusion of letters. In philological discussions on the complex processes which can lead to variation, the confusion of letters such as <cl> with <d> has been discussed early on, probably already in antiquity (Vanek, 2007, p.276). Reynolds and Wilson (2013, p.222) identify conscious and inadvertent processes as underlying such and other processes.

The current paper tries to determine how well the true stemmatic trees for artificial benchmark datasets in stemmatology (gold standards) can be approximated from external data on the confusion of letters. As Spencer et al. (2004b), we use manually provided alignments, derive pairwise distance matrices and then use the Neighbor-Joining (NJ) algorithm for stemma generation from the distance matrix. In computing the distance matrix of pairwise variant text distances, we compare each position of the alignments and implement three metrics, the simple binary (same or different variant?) Hamming distance (Hamming, 1950), the Levenshtein distance (Levenshtein, 1965) and the weighted Levenshtein distance.[6] For weighting, we do not consider philological classes of variation but distance matrices from psycholinguistic research on letter distances. These have been gained in experimental set-ups and do thus suffer less from a weighting bias introduced through subjectivity as mentioned in Spencer et al. (2004b). Comparing stemma generation with philologically inspired weighting against unweighted stemma generation (Hamming distance), Spencer et al. (2004b) found no crucial differences in the resulting stemmata for their data set but stated (p. 236) that 'different weightings could lead to completely different stemmata' concluding (p.238) that 'Determining appropriate weightings in these cases is an open problem'. The main aim of this paper is to assess a part of this problem through using external data for weighting.

2. Artificial Data Sets

For evaluation, we use three most used artificial datasets, called Parzival PRZ (English), Notre Besoin NB (French) and Heinrichi HR (Finnish) (Baret et al., 2004; Spencer et al., 2004a; Roos and Heikkilä, 2009) both in their entirety. A fourth[7] and fifth (Hoenen, 2015a)[8] are not focussed.[9] PRZ has 21 manuscripts and the alignment has 855 lines, NB features 13 manuscripts of 1035 lines and HR 64 manuscripts of 1208 lines.[10] From a machine learning perspective, these data sets are quite small and from a historical perspective, they may not represent but a tiny fraction of possible scenarios. Results are thus to be taken with utter caution. Nevertheless, these are the only data in the field for which an indisputable gold standard exists.

3. Method and Model

Of all pairwise manuscript comparisons, in large numbers of cases both manuscripts do not share an edge in the true stemma. Hence, those comparisons will include word pairs which stem from remotely distant manuscripts on the stemmatic tree. On each edge on the shortest path between them some event(s) may have happened with the implication that one is most often looking at variation reflecting more than one copy step and a back and forth of directionality. This is unfortunate but unavoidable if one doesn't know the true relations in advance. To illustrate, Figure 2 gives a small toy example of a tradition, where each manuscript contains only one word and where thus the comparison of the concurrent word pairs correspond to manuscript pair comparisons. Looking at Figure 2, when comparing and aligning words, all pairs are different in terms of a binary classification. Using the Levenshtein distance, token pair c gets the same distance as pair b, but only b corresponds to an edge. Counting differing alignment positions assigns the same distance to b and c as well. Only carefully chosen weights achieve an overall weighting that assigns the three lowest values to the pairs corresponding to edges and the largest to that with the longest path in the true stemma. Weights in the example are set intuitively to mimick confusability of the aligned letter units. Realistic confusability patterns of letters and phonemes have been researched

[5] Terminologically, there are some slightly differing terms which imply similar things: variant, innovation, error, change, alteration. Since 'error' implies a knowledge of the correct form, the term can sometimes lead to controversies. Here, we use all terms quasi-interchangeably.

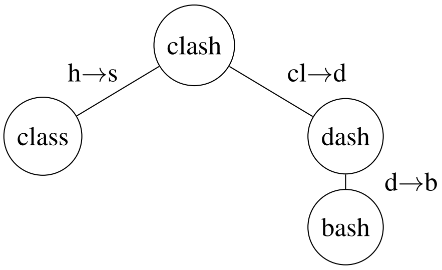
[6] Weights for transpositions are not immediately derivable from psycholinguistic letter confusion matrices. Additionally, there are long distance transpositions or transpositions of vowels of adjacent syllables which would require some additional linguistically carefully modelled distance. The java library debatty `info.debatty.java.stringsimilarity` was used for implementation of the weights.

[7] https://phylomemetic.wordpress.com/2015/02/12/artificial-textual-tradition-julies-caesar/ :last accessed on 04.02.2018

[8] Available under Tascfe at https://www.texttechnologylab.org/applications/corpora/.

[9] Both have been used scarcely in the literature, compare Robinson (2015) and the correct stemma available online for the first of them does not feature all node names. The second requires substantial additional modelling since it has a) multiple roots, b) is written in the Arabic writing system of Persian usage and c) because there are considerably less psycholinguistic resources for this constellation. Additionally the text is rather short. Experiments were conducted and results are briefly summarized below.

[10] We include the PRZ_loss challenge data set (17 ms) for comparison.



Index	comp	pos	Lev	MMD*	l(path)	EDGE
a	c-l-a-s-**h** c-l-a-s-**s**	1	1	0.4	1	T
b	**cl**-a-s-h **d** -a-s-h	1	2	0.2	1	T
c	**cl**-a-s-h **b** -a-s-h	1	2	0.8	2	
d	**cl**-a-s-**s** **d** -a-s-**h**	2	3	0.2+0.4 0.6	2	
e	**cl**-a-s-**s** **b** -a-s-**h**	2	3	0.8+0.4 1.2	3	
f	**d**-a-s-h **b** -a-s-h	1	1	0.3	1	T

Figure 2: An example stemma and all corresponding (text=) word pair comparisons. All word pairs are manually aligned, corresponding comparisons (column comp.) highlighted. The number of such comparisons or positions (column pos) is compared to the Levenshtein distance (lev) and a modally weighted version of it (*Multi Modal Distance, MMD, with one addend for each comparison). Path length (l(path)) between the nodes of a pair and whether this corresponds to an edge serve evaluation. Only MMD achieves an optimal ranking.

and can be inferred from psycholinguistic experiments, see next section, which brings external data into the model and which might help to avoid overfitting and subjectivity. Another question is how much these linguistically speaking low-level phenomena are responsible for the variation observed in copies.

3.1. Model

We operate with a number of observed (survived) manuscript variant texts $\mathbb{M}$, which are arranged in a provided token level alignment A. For each variant text pair $(M_i, M_j), i \neq j$, we sum the weighted Levenshtein distances of all words $M_{i,j_{1..k}}$ (implying different letter level alignments) according to the different weighting schemas of the modalities and then weight again each modality with a linear factor.

$$\begin{aligned} &\Delta(M_{i_k}, M_{j_k}) = \\ &\alpha \cdot wLev_{vis}(M_{i_k}, M_{j_k}) + \\ &\beta \cdot wLev_{ac}(M_{i_k}, M_{j_k}) + \\ &\gamma \cdot wLev_{mot}(M_{i_k}, M_{j_k}), i \neq j, \end{aligned} \tag{1}$$

where $wLev_{modality}$ is the weighted Levenshtein distance according to the values from modally determined (**vis**ual, **ac**oustic, **mot**oric) psycholinguistic letter distance matrices, M_{i_k} is the $k-$th token (alignment position, often word) of the $i-$th manuscript and α, β, γ are the respective weights for the modalities. The final distance of a variant text pair is then simply

$$\sum_{k=1}^{length(A)} \Delta(M_{i_k}, M_{j_k}). \tag{2}$$

In the even simpler conditions for comparison, the distance function Δ simply returns 1 (in case of difference) or 0 (in case of identity) of the elements in (M_{i_k}, M_{j_k}) or in the other condition $Lev(M_{i_k}, M_{j_k})$.

3.2. Modalities

Copying is a very complicated process and builds on many cognitive processes, compare Hoenen (2014), amongst others reading, retaining the read in memory and writing are involved. These make use of vision, probably acoustics (as far as retention in memory is involved) and motor innervation of the muscles responsible for the movements leading to writing. Among human modalities or senses, those three are assumed to be the decisive ones for the copy process. Whilst human languages differ profoundly in a number of parameters, the basic receptory and cognitive apparatus is essentially the same for all humans. Consequently, basic confusability patterns should across time and language be roughly stable. Therefore, we believe one can use psycholinguistically derived confusability information for a weighting regardless of the time period or language from which the textual material may stem.

Vision and Reading In comparison to the other modalities, vision is not only the most important one, but witnesses by far the largest body of research on confusability of letters. In order to model the values of the visual modality, matrices of visual confusability have to be used. Müller and Weidemann (2011) have compared 55 papers from 1886 until 2011 that describe 74 experiments (the majority using psycholinguistic approaches (ca. 82%)) to establish letter discriminability matrices for the Latin alphabet. As many tables as were readily available from the supplied paper links have been extracted and it was ensured that they were labelled for

1. modality (visual, motoric, acoustical)
2. directionality (symmetric matrix?, $\Delta(<a>, <d>) = \Delta(<d>, <a>)$?)
3. letter set (upper case, lower case, numbers, mixed case)
4. polarity (similarity or distance)

However, some matrices or data reported in the papers were not used, since they either analysed irrelevant data (perception in pigeons (Blough, 1985), discrimination of the Braille alphabet, (Gilmore et al., 1979)), reported a poor predictive performance (Coffin, 1978), provided incomplete data (Uttal, 1969), featured very few observations (Banister, 1927) or were hardly extractable due to age or condition of the pdfs. We ended up with 27 matrices.

In order to make all matrices comparable, the values were normalized to a number between 0 and 1 by using the largest value as 1, if and only if the reported values were not already in that range explicitly representing percentages.

The values were transformed if necessary turning similarities into distances. Furthermore, distances were averaged if directional differences existed: $\Delta(a, b) = \Delta(b, a)$. This was primarily done since the direction of copy when comparing two manuscripts is not apriori known. All non observed letter combinations would receive the maximal distance. For numbers Keren and Baggen (1981) provided a table and for mixed case only Boles and Clifford (1989) reported confusability values. This gave 1 matrix for numbers and mixed case visual confusion and 6 matrices for lower case to lower case letters and 17 for upper case to upper case letters. We combined those and obtained and tested 102 combinations of visual uppercase, visual lowercase, visual mixed case, visual number confusabilities with acoustic and motoric confusion matrices. Matrices have been made available on GitHub.[11]

Acoustic Modality For acoustic confusion, the process of modal transition from and into the visual medium must be modelled as an additional step. Naturally, one could choose grapheme-to-phoneme (g2p) and p2g based approaches. However, since the aim of the present study is to analyse explicitly modally motivated errors, we alternatively do the following and leave g2p/p2g as an alternative for future research.

Cutler et al. (2004) provide phoneme-based confusability matrices. We use the ones for initial vowels and consonants discriminated by natives. In word initial position, the phonemes do usually not become subject of heavy coarticulation.[12] Additionally, there was a high canonical correlation between initial and final confusability values (vowels initial and final:0.99, consonants in onset and coda: 0.81). For the mapping of phoneme pair distances to graphemic units (GU), Van Berkel (2005)'s *basic*, *contextual* and *word specific* spellings were used for English.[13] For instance, the presumably confusable GU pairs potentially representing /aʊ/ and /aɪ/ constructed from this were: <ou>:<i>, <ou>:<y>, <ou>:<ie>, and <ow>:<i>, <ow>:<y>, and <ow>:<ie>. The same corresponding normalized distance value from the matrix of phoneme distances was assigned to each of them and used with the acoustically weighted Levenshtein distance. If one GU pair could represent multiple phoneme pairs, all of its values were averaged. For Finnish and French similar resources were used to obtain GUs (Lyytinen et al., 2013; Lehtonen, 2013; Wiik, 1965; International-Phonetic-Association, 1999; Guex and Pithon, 1975; Dryer and Haspelmath, 2013; O'Grady et al., 1997).

Those acoustic distances between phonemes for which Cutler et al. (2004) have provided no values have been estimated using the average of the values of all observed pairs, which had a similar distance. This distance was measured in terms of numbers and qualities (backness, height, roundedness) of edges in the vowel diagram or number and quality (place, manner and voice) of steps in the consonant table of the International Phonetic Alphabet (IPA). See Figure 3.2., here, to come from /a/ to /ɜ/ requires a shift in height (to reach ɛ) and then one in backness summing 2 steps. To estimate the value of a pair where at least one phoneme was not observed by Cutler et al. (2004), and where their distance was one shift in height and one in backness, we sum and average all distances for observed vowel pairs from the chart which have that very distance. Analogously, for consonants, for instance /n/ to /m/ has 4 place steps, /n/ to /d/ 4 place and 1 manner steps, /n/ to /t/ an additional voice step etc. For diphthongues with missing values, the distinction was made between such diphthongue pairs which shared at least one sound and such which did not and the concurrent observed averaged values were assigned.

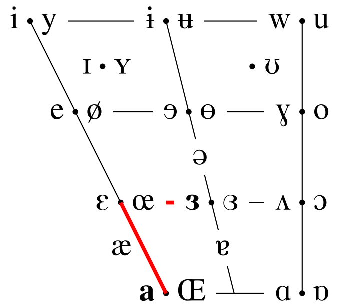


Figure 3: The vowel diagram from the IPA. Highlighted in red and bold is the path from /a/ to /ɜ/.

Motor Modality While biologically muscular neurology is well-understood, research focussing on letter production from a motor-perspective is comparatively rare. In Müller and Weidemann (2011), the only mentioned study focussing on letter production is Miozzo and Bastiani (2002), where production errors of one patient are reported, who suffered a brain damaging intoxication. They found letter substitutions to occur predominantly between letters with common strokes such as <b> and <p> and remarked that letter frequency, consonant-vowel status and letter gemination were affecting such errors. Their data was used to define all motoric weights and truly corresponds to handwriting. Non observed pairs received the maximum distance.[14]

4. Stemmatological Application of MMD

We used the artificial datasets to compute a pairwise distance matrix with the MMD (ignoring gaps in the alignment) using each combination of matrices (102 combinations). From the pairwise distances, we computed a stemma using NJ from the R package *ape* and then scored it using the so called Average Sign Distance (ASD),[15] an accuracy

[11] https://github.com/HoenenA/MultiModalDistance/. References to all in Müller and Weidemann (2011).

[12] Coarticulation is a linguistic phenomenon whereby some phonemes are influenced by previous or subsequent ones.

[13] Van Berkel (2005) analyses the English spelling system postulating for each phoneme a *basic spelling* which reflects the most frequent spelling for this phoneme, a *contextual spelling* representing a frequent but not the most frequent spelling and *word specific spellings*. Corresponding phonemes have been mapped from American to British English in the process.

[14] Values on $< a/e >$ were not used.

[15] While on the level of path comparison operating on distance, in terms of the overall manuscript comparison, the ASD is rather a similarity and referred to as Average Sign Similarity by other authors.

Trad.	Bin	Lev	MMD	RH09
NB	69.35	58.74	69.35	**77**
PRZ	72.11	67.58	72.11	
PRZ_loss	76.04	71.28	76.04	**87**
HR	72.71	72.9	74.32	

Table 1: Comparison of stemmatological evaluation results with ASD on the percentage of shared words (Bin), the Levenshtein distance (Lev) and the MMD (psycholinguistically weighted Levenshtein distance). RH09 gives the best achieved results of the 2009 study of Roos and Heikkilä.

Trad.	vis	ac	mot
NB	$-0.18(58)$	$\mathbf{0.56}(66)$	$-0.39(45)$
PRZ	$-0.43(69)$	$\mathbf{0.83}(72)$	$-0.4(68)$
PRZ_loss	$-0.6(67)$	$\mathbf{0.51}(76)$	$0.03(71)$
HR	$-0.12(73)$	$\mathbf{0.29}(74)$	$-0.17(73)$

Table 2: Pearson correlations between ASD values and modal parameters, strongest per row highlighted. In brackets, ASD value when modality was used exclusively (weighting factor set to 1 and all other weighting factors to 0, average over combinations of letter case confusion matrices).

value introduced by Roos and Heikkilä (2009) as:

$$u(A,B,C) = 1 - \frac{1}{2}|sign(d(A,B) - d(A,C)) - \\ sign(d'(A,B) - d'(A,C))|$$

A, B and C are nodes present in both the true and the estimated stemma, $d(A,B)$ is the distance of the two nodes in the true stemma defined as the number of edges on the shortest path between them, $d'(A,B)$ the same distance for the estimated tree. $sign(d(A,B) - d(A,C))$ returns so to speak only the sign, discarding length, thus -1 if $d(A,B) < d(A,C)$, 1 in the opposite case and 0 if both are equal. The index equals 1 if both stemmata agree and 0 if they differ ($\frac{1}{2}$ if they partly agree, for details see the formula or (Roos and Heikkilä, 2009)) and is computed and turned into a proportion for all such triples. ASD is the to date most used evaluation metric for stemmata on the artificial data sets used for instance in (Lai and O'Sullivan, 2010; Roos and Zou, 2011; Hoenen, 2015a).[16]

4.1. Experiment and Results

For each of the artificial traditions, we tested 102 combinations of uppercase and lowercase confusion matrices and for each such combination, we tested 66 different parameter settings, including such where the weight for any one parameter was 1.0. In all, these were $6,732$ combinations per tradition, thus roughly $27,000$ results. Since this is far too much to be displayed in a simple table, we give results in several different ways. Table 1 contains the *best achieved results* of the MMD for each tradition (including the further not focussed loss scenario). Ranges between the best and worst results in the whole grid of $6,732$ configurations were roughly 24% ASD for NB with the worst result 45, 5 for PRZ (worst:67) and 26 for HR (worst:48).

As what regards the combinations of uppercase and lowercase matrices, it must be said that those transitions truly involving only uppercase letters were rare and those involving mixed case still very infrequent (in NB roughly 10% and for PRZ roughly 14%) in respect to lowercase to lowercase transitions. All matrix combinations (1 uppercase, 1 lowercase confusion matrix) witnessed parameter settings for which the respective best results were produced. Averages per matrix (over all parameter settings) produced roughly similar results and no matrix combination was an extreme outlier.

As for the values of the modalities, we looked at the weighted average contributions of the parameters, that is for each modality: $\sum_{i=1}^{6732} \omega * ASD[i]$, where all 6,732 ASD values are in one array and each position of the array is conditioned by four parameters: the matrix combination, the weight for vision, acoustics and motorics and where ω is the corresponding weight for the modality under investigation.

Values were almost identical for the modalities and a desired significant difference was not visible. Looking only at those results, where one of the parameters had been set to 1, only for NB some significant pattern emerged: vision (contributing 39%) worked slightly better than audition (contributing 34%) while motorics performed a little worse (contributing 27%) and deviated from the mean significantly (t-test, significance level 0.01). Getting a deeper insight, while this was not possible for the matrix combinations due to the categorical character of these data points, for the weighting factor values additionally a Pearson correlation analysis with the ASD array could be conducted, which yielded more interesting results, see Table 2.

There is a strong positive correlation of the acoustic modality with the result for PRZ and NB and a weak one for HR. It must be said however that these correlations are to be understood as on the conjunction of parameter settings. This means that a larger negative value does not automatically mean that there is a bad effect of this modality but solely that in conjunction with at least one better performing modality, the contribution to the overall result was moderate. In other words, the reason why ASD values suffer if the negatively correlated modality gets stronger may be the result of the more effective ones getting weaker not necessarily because of a bad fit of the modality itself. This is corroborated by the values where the single modalities were used exclusively and by the fact that the best overall results were often only reached in settings where the modalities had been combined.

We conducted all the above analyses in the same way also for the so called TASCFE corpus (Hoenen, 2015a) which has some special characteristics such as being based on 4 different initial versions (entailing multiple roots or 4 clusters) and can be used as a testset to robustness of a stemma generating algorithm. Secondly and most importantly, TASCFE is written in Persian making all letter distance matrices from Müller and Weidemann (2011) unuseful. Not least because of the small length (the alignment

[16] A python and a C++ implementation are available from (Roos and Heikkilä, 2009) through the stemmatology challenge website.

features only 137 positions) this data set is the most challenging and produced not unexpectedly the worst results. Best ASDs were roughly between 56 and 63 (still far from chance) on the 4 complete single subsets. For the MMD, visual confusion resulted in only 1 matrix which had been modelled based on the similarity of letter features in Wiley et al. (2016), motoric similarity came from the same source but used the reported stroke similarity between the letters. Acoustic to graphemic mappings were deducted from the International Phonetic Alphabet (International-Phonetic-Association, 1999).

5. Discussion

Results are generally negative in that they did not outperform the best of some previously reported values although they are in the range of many of the there presented approaches.[17] However, to a certain extent this was expectable, given that only a subset of the innovations found which occurred along the textual transmission are estimated to be the direct result of simple modal or multi-modal confusion on the token level. Consequently, we first looked at the data in more detail to find out how many of the deviations were possibly such captured by the MMD. We conducted a tentative and surely partly subjective classification to this end. In order to better be able to interpret this data, the analysis was confined to NB (French) and PRZ (English). Some of the confusions occurred many times in different copy processes not always of the same source manuscript. Some of them, in the French case, are presumably reverts of a dictation copy where a non-native speaker had misrecorded some silent endings. A large number of cases involved deletions or insertions of letters, which possibly in part explain why the results of the binary distance and the MMD are not differing for NB and PRZ. Those are not subject to the MMD weighting schema but will make token distance coincide with the Hamming value. Another reason can be that some distances in the matrices (for some matrices, the majority of distances) were so small that they could hardly make a big difference as compared to the Hamming distance.

Overall, we found roughly a third of the differences to be applicable to a visual (127 of 422) or acoustic (115) weighting. Motoric confusion was deemed possible for roughly a sixth (56) of the cases. In conjunction roughly half (206) of the variation was subject to MMD weighting.

As for the matrices, generally their performance was not extremely different with some interesting observations. Müller and Weidemann (2011) comparing 11 of the matrices found a mean correlation of 0.68 to the generated average matrix (p.30), which well aligns with our observation. The matrices were all qualitatively roughly similar but some would have a large range between smallest and largest values, some would give differing values for self similarity. In cases with no self similarity reported, we had assigned unity regardless of the magnitude of differences with and in the rest of the matrix. Overall, the matrix of (Geyer, 1977) performed best by a very small margin. (Courrieu and De Falco, 1989)'s matrix although on average best performer for English clearly performed worst for French. The data had been gained from the confusions of preschoolers and showed a presumably acoustically decoupled confusability component (<p> with <q>). Such relations might have influenced the distance matrix in a way as to overwrite some genealogical relations for French so that NJ, which is a greedy algorithm found some tree quite different from the others. Generally, the information from the differences which are not measurable by MMD may additionally crucially determine the schema of information reduction from distance matrix to stemma and thus obscure the fit.

Comparing the different metrics, interestingly, the Levenshtein distance was clearly outperformed for NB and PRZ by the binary distance despite Levenshteins ability to measure the degree of difference between two tokens. For HR however, this was not the case. Furthermore, for HR MMD was outperforming the binary distance. This result may be due to the writing systems of the languages involved. More specifically, Katz and Frost (1992) introduce the notion of *orthographic depth*. English and French in this sense are deep orthographies, that is their g2p and p2g relations contain many n:m relationships, whereas Finnish is a shallow system (Joshi and Aaron, 2013). Illustrating the difference between a deep (English) and a shallow (Finnish) orthography, it may suffice to look at the following two examples:
P2G: /k/ → {<c>, <k>, <ck>}^{EN}, {<k>}^{FIN}
G2P: <a>→ {/ɑ/, /ɑː /, /ɒ/, /æ/, /eɪ/}^{EN},{/ɑ/}^{FIN}
The values from our confusion matrices in the MMD cover 1:1 letter confusion values. If now confusion took place also on some levels of graphemic units, these would not be captured by the visual and motoric confusion values, albeit by the acoustical ones. For instance the confusion of <their> and <there> could entail such a larger-unit-based confusion, where not one letter is cofused with one other letter. Acoustic distances as modelled however take into account such units since there is the above-described mapping between phonemes and graphemic units. In fact, the positive correlations of the acoustic weighting factors seem to support such an interpretation. Moreover, since Levenshtein may assign too large a value to confusions which involve n:m relations that correspond to just one confusion it may introduce noise, especially for deeper orthographies, so much so that its overall result becomes worse than the binary distance.[18] For the same reason, MMD is distinguishable from the binary distance for Finnish. In this vein, results all seem to be most consistent with an interpretation which suggests that the proposed method currently works best for texts written in languages with a shallow writing system (e.g. Latin). Confusion matrices for more com-

[17] Note, that our results on the binary distance combined with NJ achieved slightly worse results than those obtained in the challenge using the same data, metric and algorithm, for NB (theirs:76.2, ours:69.35), PRZ (theirs:81.5, ours:76.04) . This may be due to the fact that NJ is a greedy algorithm and different implementations and/or manuscript text orderings may output different trees.

[18] A similar explanation may hold for the observation of Spencer et al. (2004b) who found that subjective weights had made few difference. Here, weights might have accidentally obscured the genealogical information although the weighting, quite like Levenshtein may not have been unreasonable in itself.

plex orthogaphic units could be useful.[19] It also suggests that the level of graphemic units could be quite important in analysing confusion (on token level). This interpretation would be consistent also with the Persian data, where taking into account abstract letter identities[20] produced some better values than the graphemic distances.

However, utter caution must be taken since the data sets are by all means small and not representative of historical data as such. Their size entails a grave danger of overfitting, which is why using methods of machine learning to optimize the weights may be dangerous and surely much more effective on larger data sets. Additionally, our model of an interplay of the modalities is not the only possibility and ideally each position of a manuscript would require some different weighting input or an entirely different model (for instance if not modal confusion on the token level but contextual priming effects paired with some degree of visual similarity cause miscopying (Hoenen, 2015b)). There are confusions, where one single modality is to be held responsible. When Spencer et al. (2004a) mention the exapmle of <cl> and <d>, it is unlikely that the reason for the confusion lie in any other modality than vision. Thus modelling each modality separately and summing them, apart from having neurological correlates, is not unreasonable but the presented approach is surely just a first step to investigate a complex and data sparse object.

6. Conclusion

We presented an approach to weighted stemma generation from pairwise manuscript text distance matrices. In the approach, external data in the form of psycholinguistically generated letter and phoneme distance matrices in the visual, acoustic and motoric modalities was used to model weights for a weighted version of the Levenshtein distance. We tested and evaluated the approach producing stemmata from manuscript pair distances of three artificial data sets with known ground truth. Results were not outperforming the best results reported in Roos and Heikkilä (2009), but in all cases were better than many other approaches. Which external input matrix to choose was found not to be crucial in our setting and all combinations of matrices performed very similarly. Regarding the contribution of the single modalities, acoustics as modelled performed very well, but best results were often only achieved when the modalities were combined in a weighting schema. We additionally found that most likely orthographic depth was the reason why MMD outperformed the binary distance only for Finnish and why the unweighted Levenshtein distance was outperformed by the binary distance for the French tradition NB and the English tradition PRZ. The main contribution of the paper is thus in corroborating an argument in the discourse. That argument is that weighting beyond the word level may make sense, but weights must be carefully elicited and theoretically grounded for instance using psycholinguistically derived confusability matrices. Approaches to weighting which are not confined to the comparison of manuscript and token pairs, but which take into account additional distributional information of each variant, such as the one presented by Roelli and Bachmann (2010) could improve results of weighted approaches in another vein, which a quantitative assessment for instance against the benchmark datasets could reveal. We conclude with a word of caution, that all results have been obtained on relatively small data sets.

7. Acknowledgements

The author would like to thank Prof. Dr. Alexander Mehler, who proposed the design of formula (1) and made many other helpful suggestions and provided infrastructure.

8. Bibliographical References

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

Banister, H. (1927). Block capital letters as tests of visual acuity. *The British journal of ophthalmology*, 11(2):49.

Baret, P., Macé, C., and P. Robinson, e. (2004). Testing methods on an artificially created textual tradition. In *Linguistica Computationale XXIV-XXV*, volume XXIV-XXV, pages 255–281, Pisa-Roma. Instituti Editoriali e Poligrafici Internationali.

Blough, D. (1985). Discrimination of letters and random dot patterns by pigeons and humans. *Journal of Experimental Psychology: Animal Behavior Processes*, 11(2):261–280.

Boles, D. B. and Clifford, J. E. (1989). An upper- and lower case alphabetic similarity matrix, with derived generation similarity values. *Behavior Research Methods,Instruments, & Computers*, 21:597–586.

Coffin, S. (1978). Spatial frequency analysis of block letters does not predict experimental confusions. *Perception & Psychophysics*, 23(1):69–74.

Courrieu, P. and De Falco, S. (1989). Segmental vs. dynamic analysis of letter shape by preschool children. *Cahiers de psychologie cognitive*, 9(2):189–198.

Cutler, A., W., A., Smits, R., and Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, 116:3668 —-3678.

Matthew S. Dryer et al., editors. (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ellison, J. W. (1957). *The use of electronic computers in the study of the Greek New Testament text*. Harvard University.

Flight, C. (1994). A complete theoretical framework for stemmatic analysis. *Manuscripta*, 38(2):95–115.

Geyer, L. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, 22(5):487–490.

Gilmore, G., Hersh, H., Caramazza, A., and Griffin, J. (1979). Multidimensional letter similarity derived

[19] To this end, some experiments with OCR error data on n:m confusions showed positive effects.

[20] An abstract letter identity is a cognitive entity which connects different elements of a writing system behaving in the same way, for instance the lowercase 'incarnation' and the uppercase 'incarnation' of a letter $\{a, A\}$.

from recognition errors. *Perception & Psychophysics*, 25(5):425–431.

Guex, A. and Pithon, M. (1975). *Manuel de phonétique française*. Ecole de français moderne de l'Université, Lausanne.

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29(2):147–160.

Hoenen, A., Eger, S., and Gehrke, R. (2017). How many stemmata with root degree k? In *Proceedings of the 15th Meeting on the Mathematics of Language*, pages 11–21.

Hoenen, A. (2014). Simulation of scribal letter substitution. In T. L Andrews et al., editors, *lectio 1*, pages 119–139. Brepols, Turnhout.

Hoenen, A. (2015a). Das artifizielle Manuskriptkorpus TASCFE. In *DHd 2015 - Von Daten zu Erkenntnissen - Book of abstracts*. DHd.

Hoenen, A. (2015b). Simulating misreading. In *Proceedings of the 20th International Conference on Applications of Natural Language to Information Systems (NLDB)*.

Hoenen, A. (2017). Beyond the tree - a theoretical model of contamination and a software to generate multilingual stemmata. In *Book of Abstracts*, pages 155–159. AIUCD.

International-Phonetic-Association. (1999). *Handbook of the International Phonetic Association*. Cambridge University Press.

Joshi, R. M. and Aaron, P. (2013). *Handbook of Orthography and Literacy*. Routledge.

Katz, L. and Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. *Haskins Laboratories Status Report on Speech Research*, SR-111:147–160.

Keren, G. and Baggen, S. (1981). Recognition models of alphanumeric characters. *Perception & Psychophysics*, pages 234–246.

Lai, P.-H. and O'Sullivan, J. A. (2010). Mdl hierarchical clustering with incomplete data. In *Information Theory and Applications Workshop (ITA), 2010*, pages 1–5. IEEE.

Lehtonen, A., (2013). *Handbook of Orthography and Literacy*, chapter Sources of Information Children Use in Learning to Spell: The Case of Finnish Geminates, pages 63–80. Routlegde.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848.

Lord, A. B. (1960). *The Singer of Tales*. Harvard University Press.

Lyytinen, H., Aro, M., Holopainen, L., Leiwo, M., Lyytinen, P., and Tolvanen, A., (2013). *Handbook of Orthography and Literacy*, chapter 4, pages 47–62. Routlegde.

Miozzo, M. and Bastiani, P. D. (2002). The organization of letter-form representations in written spelling: Evidence from acquired dysgraphia. *Brain and Language*, 80(3):366 – 392.

Mooney, L. R., Barbrook, A. C., Howe, C. J., and Spencer, M. (2003). Stemmatic analysis of lydgate's "kings of england" : a test case for the application of software developed for evolutionary biology to manuscript stemmatics. *Revue d'histoire des textes*, 31(2001):275–297.

Müller, S. and Weidemann, C. (2011). Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta Psychologica*.

O'Grady, W., Dobrovolsky, M., and Katamba, F. (1997). *Contemporary Linguistics*. Longman, St. Martin's.

O'Hara, R. J. (1996). Trees of history in systematics and philology. *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano*, 27(1):81–88.

Pasquali, G. and Pieraccioni, D. (1952). *Storia della tradizione e critica del testo*. Le Monnier.

Reynolds, L. and Wilson, N. (2013). *Scribes and Scholars, A Guide to the Transmission of Greek & Roman literatures*. Oxford University Press.

Robinson, P. (2015). Four rules for the application of phylogenetics in the analysis of textual traditions. *Digital Scholarship in the Humanities*.

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

Roelli, P. and Macé, C. (2015). Parvum lexicon stemmatologicum. a brief lexicon of stemmatology.

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

Roos, T. and Zou, Y. (2011). Analysis of textual variation by latent tree structures. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, pages 567–576.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Spencer, M., Davidson, E. A., Barbrook, A. C., and Howe, C. J. (2004a). Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology*, 227:503–511.

Spencer, M., Mooney, L., Barbrook, A., Bordalejo, B., Howe, C., and Robinson, P. (2004b). The effects of weighting kinds of variants. In P. van Reenen, et al., editors, *Studies in Stemmatology II*, pages 227–240. John Benjamins.

Uttal, W. R. (1969). Masking of alphabetic character recognition by dynamic visual noise (dvn). *Perception & Psychophysics*, 6(2):121–128.

Van Berkel, A., (2005). *Second Language Writing Systems*, chapter The Role of Phonological Strategy in Learning to Spell, pages 97–121. Multilingual Matters.

Vanek, K. (2007). *Ars corrigendi in der frühen Neuzeit: Studien zur Geschichte der Textkritik*, volume 4. Walter de Gruyter.

Wiik, K. (1965). *Finnish and English vowels*. Turun Yliopisto.

Wiley, R. W., Wilson, C., and Rapp, B. (2016). The effects of alphabet and expertise on letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, 42(8):1186–1203.