Phylogenetic inference from word lists using weighted alignment with empirically determined weights

Gerhard Jäger*
University of Tübingen & Swedish Collegium for Advanced Study, Uppsala
gerhard.jaeger@uni-tuebingen.de

October 22, 2013

Abstract:

Abstract

The paper investigates the task of inferring a phylogenetic tree of languages from the collection of word lists made available by the *Automated Similarity Judgment Project*. This task involves three steps: (1) computing pairwise word distances, (2) aggregating word distances to a distance measure between languages and inferring a phylogenetic tree from these distances, and (3) evaluating the result by comparing it to expert classifications. For the first task, weighted alignment will be used and a method to determine weights empirically will be presented. For the second task a novel method will be developed that attempts to minimize the bias resulting from missing data. For the third task, several methods from the literature will be applied to a large collection of language samples to enable statistical testing. It will be shown that the language distance measure proposed here leads to substantially more accurate phylogenies than a method relying on unweighted Levenshtein distances between words.

Keywords: language phylogenies; automatic language classification; ASJP; weighted string alignment

1 Introduction

Recent years have seen the introduction of many proposals to use phylogenetic inference techniques from bioinformatics in order to extract information about genetic relations from languages. There are essentially two basic approaches being currently employed. *Character based* methods start by defining a set of features — *characters* — to classify languages. The feature values are assumed to be inert in language change. Therefore the number of shared feature values between two languages can be taken as a measure of their relatedness. Methods such as Maximum Parsimony, Maximum Likelihood, and Bayesian phylogenetic inference take a classification of languages according to a list of features as input and produce

*This research was supported by the ERC Advanced Grant 324246 *Language Evolution: The Empirical Turn* (EVOLAEMP).

a phylogenetic tree including changes in feature values along the branches as output (see, for instance, Felsenstein, 2004, for a comprehensive overview). Suitable features may be cognate classes of basic vocabulary items (as used by, for example, Gray and Atkinson, 2003, and Bouckaert et al., 2012) or grammatical features (Dunn et al., 2005, and subsequent work).

The second approach uses *distance based* techniques of phylogenetic inference. These methods start from a matrix of pairwise distances between languages that ideally correspond to the time that has passed since the split of the latest common ancestor of the two languages compared along the two lineages leading to those languages. Phylogenetic inference produces a tree where the path length between two leaf nodes is as close as possible to their pairwise distance. Such methods are suitable when dealing with raw data that are not organized in a feature matrix, such as lists of non-cognate-coded basic vocabulary items.

Extracting phylogenetic information from word list usually proceeds in three steps (see for instance Downey et al., 2008 or Holman et al., 2008): (a) the similarity/distance between words from different languages is determined using some kind of alignment algorithm, (b) these word distances are aggregated to pairwise distances between languages, and (c) a phylogenetic tree is inferred. As for the final step, the *Neighbor Joining* algorithm (Saitou and Nei, 1987) has emerged as the *de facto* standard.

The quality of a phylogeny thus inferred can be assessed by comparing it to expert classifications. How such a comparison is to be performed is an active area of investigation, see Wichmann et al. (2010); Greenhill (2011); Huff and Lonsdale (2011); Pompei et al. (2011) for some recent contributions.

In this study I will propose three innovations pertaining to this research program:

1. A similarity score between words that is computed via weighted alignment, including a procedure to obtain the required weights in a data-driven way,

2. a novel method to aggregate word similarity score into a distances between languages, and

3. a generalization of existing methods for evaluating the quality of distance measures between word lists using expert classifications as gold standard.

The study is carried out using version 15 of the the *Automated Similarity Judgment Project* database (Wichmann et al., 2012), a collection of Swadesh lists for more than 5,800 languages[1] which are phonetically transcribed in a uniform way. Only the 40 most stable Swadesh concepts were used in this paper. After excluding artificial languages, creoles, extinct and reconstructed languages, 5,644 word lists were kept in the database. Attested loan words were not excluded. Diacritics in the phonetic transcriptions were ignored.

As baseline for comparison, I will use the method to compute language distances from ASJP word lists described in Holman et al. (2008).

The structure of the paper is as follows. In Section 2 Holman et al.'s proposal will be reviewed. The novel method for aggregating word similarity scores will be developed in Section 3. Section 4 discusses the issue how to evaluate distance measures between languages and presents a comparative evaluation of the different aggregation methods. Section 5 introduces weighted word alignment and presents the procedure to train the required weights with ASJP

[1] As Søren Wichmann (p.c.) points out, *doculects* would be a more precise term, as the database comprises languages, dialects, and reconstructed word list of proto languages. For simplicity's sake I will use the terms *language* and *doculect* synonymously throughout this paper though.

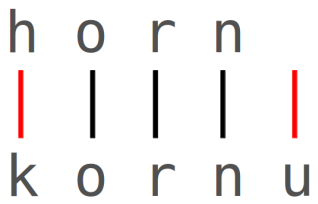

Figure 1: Levenshtein alignment

data. The distance measure obtained this way is thoroughly compared empirically to alternative approaches. In Section 6 the method developed here will be compared to Kondrak's (2002) ALINE system. Section 7 contains some final discussion and conclusions.

2 State of the art: The LDND score

In Holman et al. (2008) a method to compute distances between ASJP word lists based on the *edit distances* between individual words is proposed. The edit distance or *Levenshtein distance* between two words w_1 and w_2 is defined as the minimal number of edit operations (insertion, deletion, replacement) necessary to transform w_1 into w_2. Alternatively, it can be defined as the minimal number of mismatches in an alignment of two words. In the example in Fig. 1 (showing the alignment between the English and the Latin word for *horn*, spelled according to the ASJP transcription system; the ASJP symbols are explained in the Appendix), this value would be 2. To control for varying word length, Holman et al. normalize this measure by dividing it by the length of the longest word. In the example, this amounts to 0.4. By definition, the normalized Levenshtein distance *LDN* takes values between 0 and 1.

The normalized Levenshtein distance provides a distance measure between words. To obtain a distance measure between two word lists, it seems suggestive to simply average over the LDN scores between corresponding words from the languages to be compared. However, if two languages have small and strongly overlapping sound inventories, the number of chance hits is high as compared to a language pair with large and dissimilar sound inventories. On average, the LDN values between unrelated words will be smaller in the former than in the latter case. To control for this effect, the authors propose a method to calibrate the average LDN score between synonymous word pairs to the specific language pair to be compared.

This is best illustrated with an example. Tab. 1 shows the pairwise LDN scores for some English and the Swedish vocabulary items from ASJP.

The average of the values along the diagonal — i.e. between words with identical meanings — for the full matrix is 0.56, while the average of the off-diagonal values — word pairs with different meanings — is 0.91. The authors define the *LDND* score of two languages (Levenshtein Distance Normalized and Divided) as mean LDN score along the diagonal, divided by the mean LDN score off the diagonal. For the comparison of English and Swedish, this amounts to 0.61.

	Ei	**yu**	**wi**	**w3n**	**tu**	**fiS**	...
yog	**1**	0.67	1	1	1	1	
du	1	**0.5**	1	1	0.5	1	
vi	0.5	1	**0.5**	1	1	0.67	
et	1	1	1	**1**	1	1	
tvo	1	1	1	1	**0.67**	1	
fisk	0.75	1	0.75	1	1	**0.5**	
⋮							

Table 1: LDN scores English/Swedish

3 Quantifying the evidence for genetic relatedness of languages

3.1 The Evidence for Relatedness

As spelled out in the previous section, the LDND score aggregates distances between words to distances between languages (i.e.: word lists over a given concept list) L_1 and L_2 by

- computing the distances between all word pairs from L_1 and L_2
- computing the average distance between synonymous and the average distance between non-synonymous words, and
- dividing the former by the latter.

In this section I will propose an alternative method for aggregating a matrix of distances between words from L_1 and L_2 to an overall distance measure between L1 and L2.

To illustrate the underlying intuition, consider again the matrix of LDN scores between English and Swedish words illustrated in Tab. 1. The distribution of off-diagonal scores is shown in Fig. 2. The word pair *fiS/fisK* 'fish' has an LDN score of 0.5. Only 4 off-diagonal entries have a lower score, and 31 entries have the same score. This means that a randomly picked pair of non-synonymous words has only a chance of about 2.2% to be more similar to each other than *fiS/fisk*. Intuitively, the fact that this fraction is so small provides evidence that *fiS* and *fisk* — and therefore English and Swedish — are related. Likewise, each of the other diagonal-entries provide a certain amount of evidence for the languages to be related, depending on their position within the distribution of off-diagonal entries.

To make this precise, let us assume that A and B are the word lists from the languages to be compared. The ith entry of A is denoted by a_i, and likewise for B and b_i. It is assumed that a_i and b_i are pairwise synonymous for all i. $d(i,j)$ is the LDN distance between a_i and b_j.

The ASJP data contain missing entries at many positions. To deal with this issue, we assume that there are N concepts for which both A and B contain entries. If A does not contain an entry for concept i, a_i and $d(i,j)$ are undefined, and likewise for B.

The *rank* of a diagonal entry $d(i,i)$—written as r_i—is the position that $d(i,i)$ would assume if it is added to the set of off-diagonal entries and the resulting set is sorted in increasing order. Formally, we have

$$r_i = |\{(j,k) : j \neq k \text{ and } d(j,k) < d(i,i)\}| + 1$$

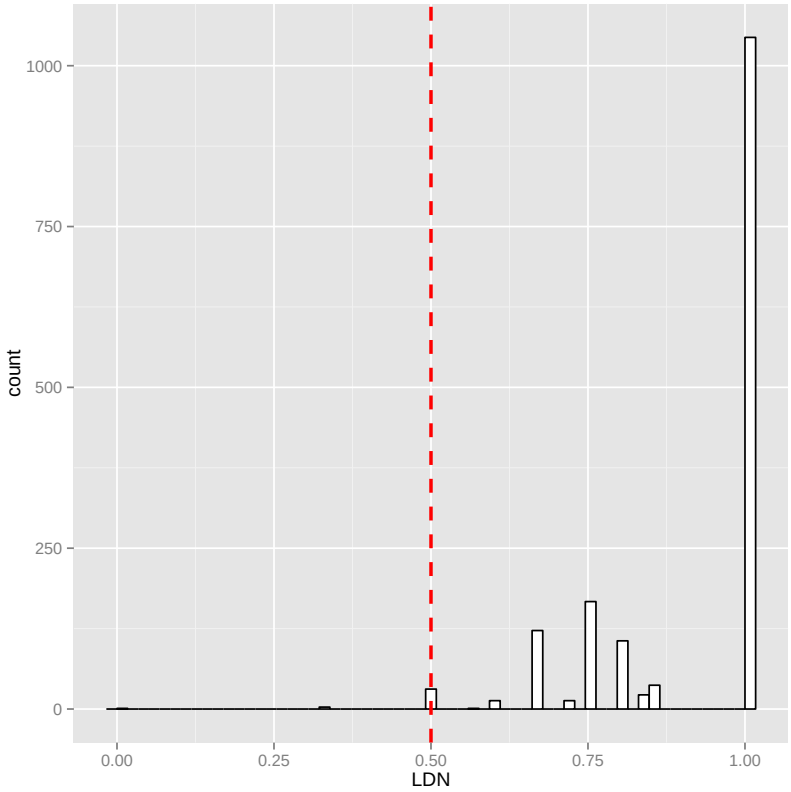

Figure 2: Off-diagonal LDN scores: English vs. Swedish

(It is tacitly assumed that only those pairs (i, j) are counted for which $d(j, k)$ is defined.)

For the time being we assume that there are no ties; this issue will be taken up later on.

As the sizes of word lists may differ between languages, we normalize the rank by dividing it by the maximal possible rank, which is the number of off-diagonal entries +1. This leads to the definition of the *normalized rank* nr_i, which always assumes a value in the interval $(0, 1]$:

$$nr_i = \frac{r_i}{|A| \times |B| - N + 1} \tag{1}$$

If the languages in question are unrelated, the entries along the diagonal are drawn from the same distribution as the off-diagonal entries. Therefore we expect each rank (between 1 and $|A| \times |B| - N$) to be equally likely. However, if the languages are related we expect some diagonal entries to be small in comparison to the off-diagonal entries, i.e. we expect ranks to be small.

This is illustrated in Fig. 3. The left panel shows the distribution of diagonal entries (left boxplot) and off-diagonal entries (right boxplot) for the comparison of English and Swedish. It is clearly visible that the diagonal scores are on average much lower than the off-diagonal scores.

The right panel shows the same data for the comparison of English with Swahili. The two languages are unrelated, and the diagonal entries are similiarly distributed as the off-diagonal entries.

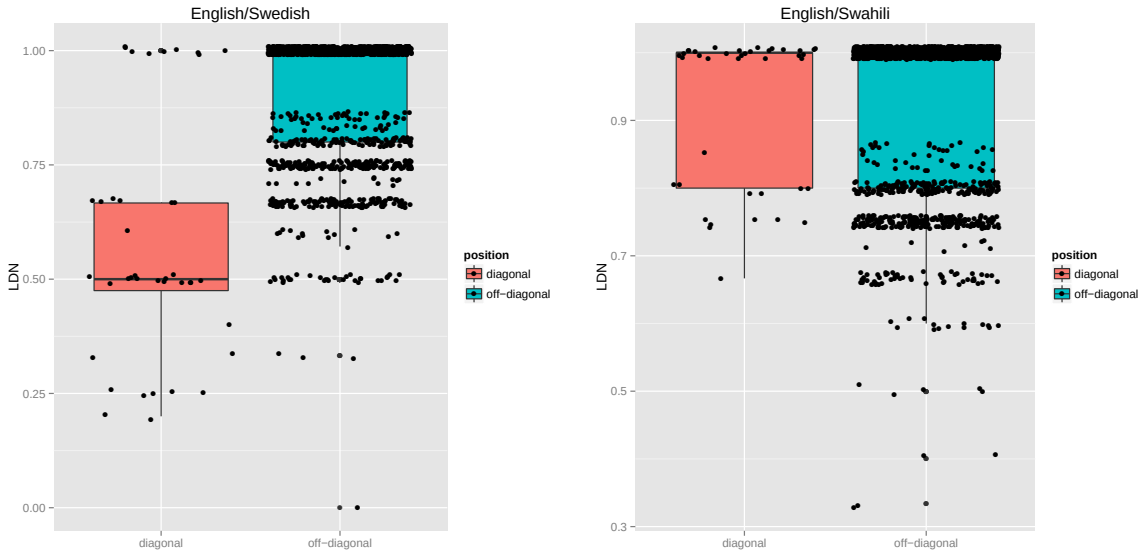

Figure 3: Distribution of diagonal and off-diagonal LDN scores: English/Swedish and English/Swahili

For a pair of unrelated languages we expect the normalized ranks to be uniformly distributed between 0 and 1. If the languages are related, this distribution should be skewed towards small values. To test this, I drew 100,000 ranks from a random selection of ASJP language pairs that belong to the same genus according to WALS, and another sample of 100,000 ranks from language pairs that pairwise belong to different WALS families. The histograms of the two distributions are shown in Fig. 4. As expected, the normalized ranks for pairs of related languages are heavily skewed towards small values, while the values for unrelated languages approximately follow a uniform distribution.[2]

Fig. 5 displays the same data as histograms with logarithmic binning in log-log plot. The values for the related languages lie approximately on a straight line with a negative slope. This indicates that the normalized ranks are distributed according to a *power law* (see for instance Clauset et al., 2009 on power law distributions in empirical data). This mean that there are real numbers $C > 0$ and $\alpha > 1$ such that

$$P(nr_i > x) \approx Cx^{1-\alpha}.$$

We can thus approximate the empirical distribution by a continuous probability density function f_{related} with

$$\begin{aligned} f_{\text{related}}(x) &= Cx^{-\alpha} \\ P(nr_i < x|\text{languages are related}) &\approx \int_0^x f_{related}(x) \end{aligned}$$

[2]The empirical distribution is not entirely uniform; it also has a slight bias towards smaller values. This may be due either to long-distance relationships between languages from different families, similarities due to language contact, or to universal biases in the sound-meaning relationship such as onomatopoeia. These effects are small in comparison to the bias for related languages though, so the uniform distribution is still a good approximation.

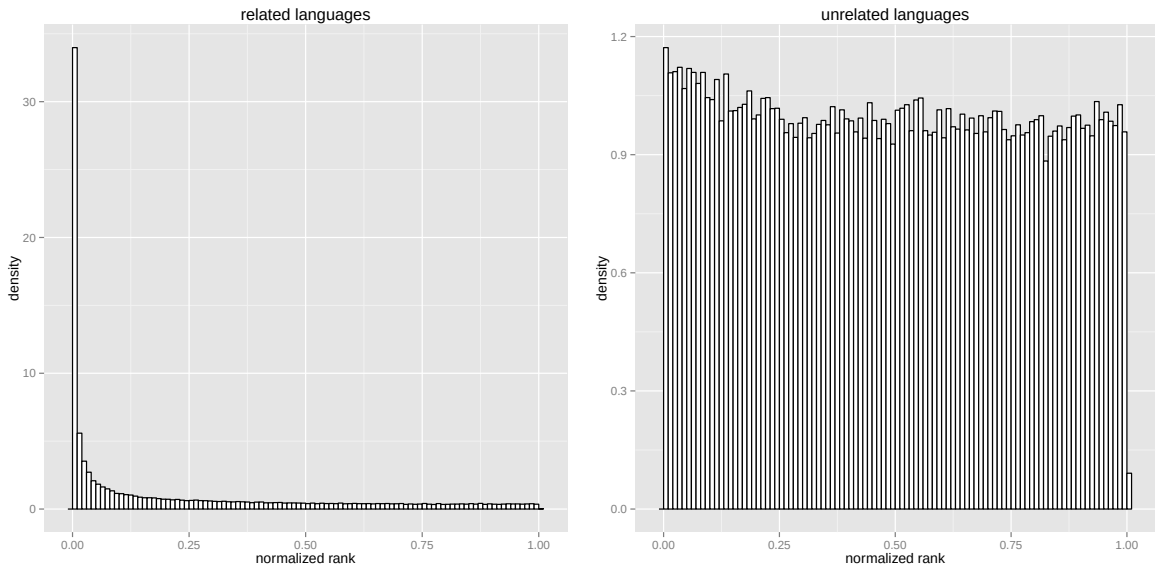

Figure 4: Distribution of normalized ranks from related and from unrelated language pairs

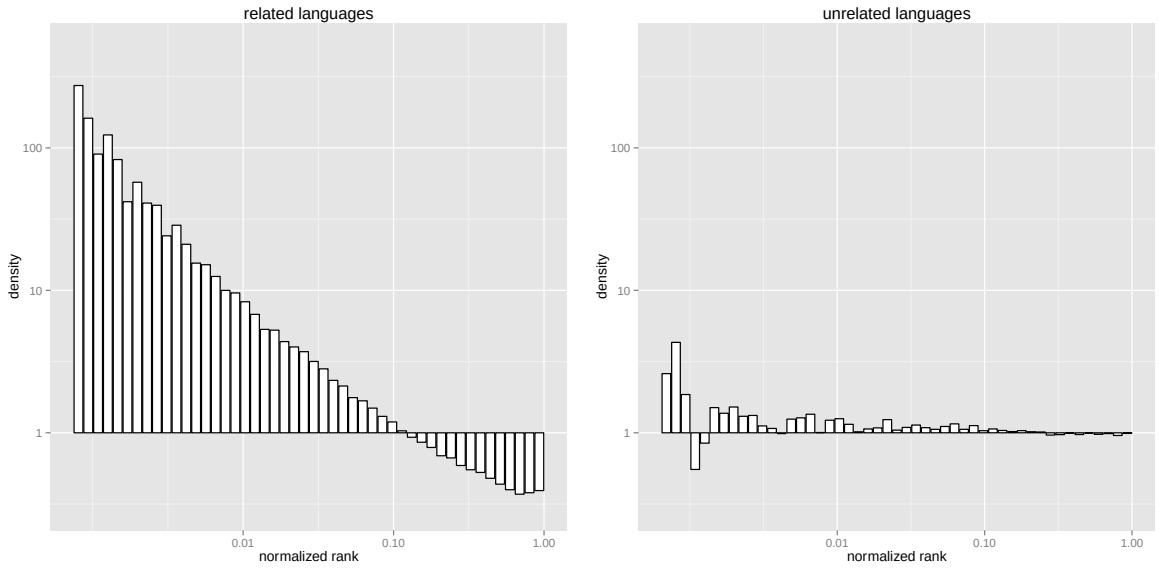

Figure 5: Distribution of normalized ranks from related languages; log-log scale

The distribution of normalized ranks for unrelated languages can be approximated by a constant density function:

$$\begin{aligned} f_{\text{unrelated}}(x) &= 1 \\ P(nr_i < x|\text{languages are unrelated}) &\approx \int_0^x f_{related}(x) \\ &= x \end{aligned}$$

Suppose we have to decide whether or not two languages are related on the basis of normalized ranks of all translation pairs. So we compare two hypotheses: H_0 (languages are unrelated) and H_1 (languages are related). According to Bayes' formula, the posterior odds of the likelihood of the hypotheses is

$$\frac{P(H_1|nr_1,\ldots,nr_N)}{P(H_0|nr_1,\ldots,nr_N)} = \frac{P(nr_1,\ldots,nr_N|H_1)}{P(nr_1,\ldots,nr_N|H_0)}\frac{P(H_1)}{P(H_0)}.$$

If we make the simplifying assumption that the normalized ranks for the individual translation pairs are stochastically independent, this amounts to

$$\begin{aligned} \frac{P(H_1|nr_1,\ldots,nr_N)}{P(H_0|nr_1,\ldots,nr_N)} &= \prod_i \frac{P(nr_i|H_1)}{P(nr_i|H_0)}\frac{P(H_1)}{P(H_0)} \\ &= \prod_i \frac{Cnr_i^{\alpha}}{1}\frac{P(H_1)}{P(H_0)} \end{aligned}$$

The posterior log odds thus come out as

$$\begin{aligned} \log\frac{P(H_1|nr_1,\ldots,nr_N)}{P(H_0|nr_1,\ldots,nr_N)} &= \sum_i(\log C - \alpha\log nr_i) + \log\frac{P(H_1)}{P(H_0)} \\ &= N\log C + \alpha\sum_i -\log nr_i + \log\frac{P(H_1)}{P(H_0)} \end{aligned}$$

While we do not know the prior probabilities $P(H_0)$ and $P(H_1)$, we can determine the term $\sum_i -\log nr_i$ empirically. The posterior log odds are a monotonically increasing linear function of this quantity.

This only holds for a constant N though. Recall that N is the number of concepts for which both word lists to be compared contain an entry. Due to missing data in ASJP, N may assume different values for different language pairs. The maximal number for N is $N_{\max} = 40$. If $N < N_{\max}$, we have to estimate the normalized ranks for the missing entry pairs. The maximum likelihood estimation is that the average value of the missing nr-values equals the average of the known values. Therefore the quantity $\frac{N_{\max}}{N}\sum_i -\log nr_i$ provides the maximum likelihood estimator. As $N_{\max}$ is constant, the estimated posterior log odds are a monotonically increasing function of the quantity

$$\frac{1}{N}\sum_i -\log nr_i.$$

Let us call this quantity the *Evidence for Relatedness* (ER).

To turn this into an operational definition, one further amendment needs to be made.

Recall that there may be ties, i.e. there may be pairs (k, j) with $d(a_k, b_j) = d(a_i, b_i)$. To put it another way, suppose we form the set of distances $\{d(a_k, b_j) : k \neq j\} \cup \{d(a_i, b_i)\}$ and sort it in increasing order. The quantity in the numerator of Eq. (1) is the number of items preceding $d(a_i, b_i)$ is this sequence; adding 1 is $d(a_i, b_i)$'s rank. If there are ties, the rank may not be uniquely defined. In this case we compute the evidence for relatedness for the ith concept for each possible rank and form the average. The possible ranks r_i in the definition below is the set of ranks that $d(a_i, b_i)$ can assume in such a sequence. The normalized rank nr_i is the geometric mean of all possible ranks, divided by the number of off-diagonal entries $+1$. Forming the geometric rather than the arithmetic mean ensures that the logarithm of the normalized rank equals the arithmetic mean of the the logarithms of the possible values of nr.

This leads to the following final definition:

Definition 1 (Evidence for Relatedness)

$$\begin{aligned} r_i &\doteq \{n+1 : |\{(j,k)|j \neq k \text{ and } d(a_j, b_k) < d(a_i, b_i)\}| \leq n \leq \\ &\qquad |\{(j,k) : j \neq k \text{ and } d(a_j, b_k) \leq d(a_i, b_i)\}|\} \\ nr_i &= \prod_{x \in r_i} \left(\frac{x}{|A| \times |B| - N + 1}\right)^{1/|r_i|} \\ ER(A,B) &\doteq \frac{1}{N} \sum_{i=1}^{N} -\log nr_i \end{aligned}$$

It seems reasonable to assume that the Evidence for Relatedness is the stronger the closer two languages are related. ER can thus be considered a similarity measure between languages. It can easily be transformed into a distance measure. ER is maximized if we compare a word list to itself, it contains no homonymies and no missing entries. In this case, all $r_i = 1$. The number of off-diagonal entries equals $40 \times 40 - 40 = 1,560$, so for all i, $nr_i = 1/1561$. Hence the ER score is

$$ER_{\max} = \log 1561 \approx 7.35.$$

The theoretical minimum for the ER score is achieved if all diagonal entries are smaller than all off-diagonal entries. In this scenario all $nr_i = 1$ and hence $ER_{\min} = 0$. The *Distance based on Evidence of Relatedness* (dER) is then defined as follows:

Definition 2 (Distance based on Evidence of Relatedness)

$$dER(A,B) = \frac{ER_{\max} - ER(A,B)}{ER_{\max} - ER_{\min}}$$

The dER score always assume a value between 0 and 1. Note that it does not depend on the values of C and α, so no parameter fitting is necessary.

3.2 Correcting for missing entries

If the word lists to be compared are incomplete, the dER measure relies on a maximum likelihood estimate of the nr scores of the missing entries. As a consequence, the absolute value both of positive and of negative evidence is overestimated in this case. While it seems unproblematic to underestimate the similarity between two languages if the available, gappy

word lists do not provide evidence for relatedness, the error in the opposite direction is potentially more serious. In the case of gappy word lists, chance similarities receive a higher weight than is actually justified. To correct this, high ER scores should be discounted somewhat in proportion to the amount of missing entries.

Suppose the languages A and B are completely unrelated. Then the nr scores are, as a good approximation, drawn from a uniform distribution over the interval $[0, 1]$. The probability density function $f_{-\log}$ for the term $-\log nr_i$ then follows a standard exponential distribution with 1 as its mean and standard deviation.[3]

The ER score is defined as the mean of N (approximately) independent variables that, if the languages are unrelated, are drawn from a distribution with mean and variance = 1. So the ER score is a random variable with mean = 1 and variance = N^{-1} if the languages are unrelated.[4]

The sum (and thus the average) of N exponentially distributed variables follows an *Erlang distribution*. However, this distribution can be approximated by a normal distribution (also with mean = 1 and variance = N^{-1}) if N is sufficiently large. This follows from the Central Limit Theorem. So we can transform the ER score to a variable that is distributed according to a standard normal distribution in the following way:

Definition 3 (Corrected Evidence for Relatedness)

$$ERC(A, B) \quad = \quad \sqrt{N} \times (ER(A, B) - 1)$$

According to this definition, the mean and variance of the ERC scores for unrelated languages do not depend on N, i.e. on the number of missing entries. This enables statistical hypothesis testing for the null hypothesis H_0: "A and B are unrelated." vs. the alternative hypothesis H_1: "A and B are unrelated." The p-value for a given ERC x score is simply the probability that the standard normally distributed variable has a value $> x$ (technically, this is the converse error function of x), regardless of the number of missing entries.[5]

Just like the ER score, the ERC score is a similarity measure between languages. It can be turned into a distance measure analogously to Definition 2.

Definition 4 (Distance based on Corrected Evidence of Relatedness)

$$dERC(A, B) \quad = \quad \frac{ERC_{\max} - ERC(A, B)}{ERC_{\max} - ERC_{\min}}$$

[3]This is an instance of a more general rule. If X is a uniformly distributed random variable over the interval [0,1] and f is a strictly monotonically decreasing function over $[0, 1]$, then $f(X)$ is distributed according to the density function $-\frac{d}{dy}f^{-1}(y)$. Here is the derivation: Let g be the probability density function of $f(X)$.

$g(y)$	=	$\frac{d}{dy}P(f(X) < y)$	(definition of probability density)
	=	$\frac{d}{dy}P(X > f^{-1}(y))$	(because f is monotonically decreasing)
	=	$\frac{d}{dy}(1 - P(X < f^{-1}(y)))$	(laws of probability)
	=	$-\frac{d}{dy}(P(X < f^{-1}(y)))$	(laws of calculus)
	=	$-\frac{d}{dy}f^{-1}(y)$	(because X is uniform over $[0, 1]$)

If $f(x) = -\log x$, then $f^{-1}(y) = e^{-y}$ and $\frac{d}{dy}f^{-1}(y) = -e^{-y}$.

The general form of the exponential distribution is $\lambda e^{-\lambda y}$, where both the mean and the standard deviation equal λ. For our special case, $\lambda = 1$.

[4]This follows from elementary laws of probability theory, i.e. the facts that $\mu(X+Y) = \mu_X + \mu_Y$, $Var(X+Y) = Var(X) + Var(Y)$ if X and Y are independent, and $Var(kX) = k^2 Var(X)$.

[5]The reader may excuse my eclectic usage of both Bayesian and frequentist arguments in this section.

A	B	$dERC/LDN(A,B)$
English	English	0.0812
English	Scots	0.2214
Danish	Swedish	0.3431
English	Swedish	0.4866
English	Frisian	0.4986
English	Dutch	0.5071
Hindi	Farsi	0.6536
English	French	0.7917
English	Hindi	0.8276
Amharic	Vietnamese	0.8401
Swahili	Warlpiri	0.8619
Navajo	Dyirbal	0.8435
Japanese	Haida	0.8547
English	Swahili	0.8687

Table 2: dERC scores

For ASJP data with $N_{\max} = 40$, it holds that $ERC_{\max} = \sqrt{40} \times (\log 1561 - 1) \approx 40.2$ and $ERC_{\min} = \sqrt{40} \times (0 - 1) \approx -6.3$.

To get an idea for the numerical magnitudes, dERC scores for some language pairs are given in Tab. 2.

It might seem counter-intuitive that the dERC of English to itself is larger than 0. This reflects the fact that the ASJP-list for English contains one pair of homonyms: both 'I' and 'eye' are transcribed as *Ei*. Therefore the probability of a chance identity is assessed as positive, and therefore the probability of the two lists being identical despite the languages being unrelated is assessed as positive, if very small.

4 Empirical evaluation

As will become clear later on, the main motivation for developing dERC is that this method of aggregation is also applicable to string distance measure with mathematical properties different from LDN.

A standard way to assess the quality of a distance measure between languages is to relate it to an expert classification. In this paper I will make use of three different expert classifications of languages:[6]

- The two-level classification according to the *World Atlas of Language Structures* Haspelmath et al. (2008), abbreviated as *WALS* in the sequel
- The classification according to *Ethnologue* Lewis (2009), abbreviated as *Ethn*, and
- The classification according to Hammerström (2010), abbreviated as *Hstr*.

I will use three methods to compare a distance matrix to an expert classification:

[6]All three classifications are provided as meta-data in the ASJP database.

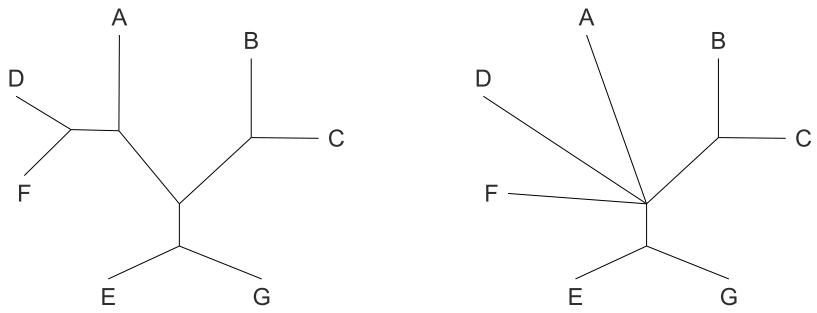

Figure 6: Example trees

1. Triplet distance: This method has been used in Greenhill (2011), and it is closely related to the Goodman-Kruskal Gamma measure used in Wichmann et al. (2010).

   A triplet of languages $(AB|C)$ is *resolved* if and only if the expert tree contains a node that dominates A and B but not C. It is correctly classified by the distance measure if and only if $d(A,B) < \min(d(A,C), d(B,C))$. The triplet distance of the distance measure to the expert tree is the proportion of all resolved triplets that are classified incorrectly.[7]

   The triplet distance (TD) measure has the advantage that it only uses comparisons between distances rather than numerical values. It is therefore invariant under all monotonic transformations of the distance measure, including non-linear ones. Also, it does not rely on a phylogenetic algorithm that may introduce its own bias.

2. Generalized Robinson-Foulds distance: The Robinson-Foulds distance (Robinson and Foulds, 1981) is a standard distance measure between unrooted trees over the same set of leaves. To illustrate it, consider the trees in Fig. 6. The two trees have four and two internal branches respectively. Each internal branch in an unrooted tree induces a bipartition of the set of leaves. The bipartitions induced by the internal branches on the right are identical to the bipartitions in the tree on the left. Additionally, the tree on the left contains two internal branches that have no counterpart in the tree on the right.

   The Robinson-Foulds distance is the number of internal branches in both trees that have no counterpart in the other tree, divided by the total number of internal branches in both trees. In the example, this number is $2/6 \approx 0.33$.

   This number is somewhat misleading though. The tree on the left is binary branching,

[7] The triplet distance is only informative if the languages to be compared exist at the same point in time, i.e. if any two related languages have the same time depth from their common ancestor. If this condition is not met, the triplet distance might be misleading. For instance, it might very well be that Old English and Gothic are closer to each other than Old English is to modern Dutch. Nevertheless the correct classification places Old English and modern Dutch into one group — the West-Germanic languages —, and Gothic into another one, namely East-Germanic. This problem could be avoided by evaluating quartets instead of triplets and induce an unrooted tree. I refrain from doing so here because the number of quartets over a set of languages exceeds the number of triplets by a factor in the order of magnitude of the number of languages. For large datasets, the triplet distance, but not the quartet distance, can still be computed with realistic computational effort. To avoid the mentioned problem, in this article I only use data from languages that are either currently alive or recently extinct.

while the one on the right is not. The tree on the left contains all bipartitions that we find in the tree on the right, so the former approximates the information contained in the latter as closely as is possible for a binary branching tree.

This is a standard situation when comparing a tree that has been constructed by a phylogenetic inference algorithm such as Neighbor Joining — which is necessarily binary branching —, with an expert tree that is not binary branching. To take this asymmetry into account, I follow Pompei et al. (2011) in using the *generalized Robinson-Foulds distance* (GRF). The GRF of a binary branching tree A to another (perhaps non-binary branching) tree B is defined as the proportion of internal branches in B that do not have a counterpart in A. In the example, the distance of the first to the second tree comes out as 0. The GRF is always a number between 0 and 1, with 1 indicating total disagreement and 0 optimal agreement.

3. Generalized quartet distance: Another commonly used distance measure between unrooted trees is the *quartet distance* (Estabrook et al., 1985). Given an unrooted tree and four leaves A, B, C, and D, the tree induces the *butterfly* $(AB|CD)$ if and only if one of the bipartitions that is induced by its internal branches separates AB from CD. If there is no internal branch separating the quartet into two pairs, the tree induces a *star* on the quartet of leaves.

   Given two unrooted trees over the same set of leaves, their quartet distance is the proportion of quartets over their leaves that have different topologies in the two trees. In the example trees in Fig. 2, we have 7 leaves and therefore $\binom{7}{4} = 35$ quartets. Of these 35 quartets, 16 have different topologies in the two trees, so the quartet distance is $16/35 \approx 0.46$.

   Similar to the Generalized Robinson-Foulds distance defined above, I will follow Pompei et al. (2011) in using a generalized version of the quartet distance that takes the asymmetry between binary branching inferred trees and multiply branching expert trees into account. The *generalized quartet distance* (GQD) between an inferred tree and an expert tree is the proportion of butterflies in the expert tree having a different topology in the inferred tree. For the example in Fig. 2, the fit is perfect, i.e. the GQD equals 0.

The quartet measures are less intuitive than the corresponding Robinson-Foulds measures, but they have the advantage of being more tolerant of small errors. For instance, exchanging two leaves in one of two large trees may have a dramatic effect on the GRF while the GQD changes only slightly.

In the sequel I will compare the three distance measures between languages discussed so far, i.e. LDND, dER and dERC. Let us first look at the triplet distances to the three expert classifications mentioned above. The comparison was performed with the full list of $5,481$ ASJP word list that come from living or recently extinct languages and dialects. The results are shown in Tab. 3. For all three expert classifications, we find a slight improvement both from LDND to dER and again from dER to dERC, even though the differences are quite small.

To compute the GRF, for each of the three pairwise distance matrices a phylogenetic tree is computed via the Neighbor Joining algorithm, and those are compared to the three expert classifications both via GRF and via GQD. The results are shown in Tab. 4.

These figures seem to indicate that LDND performs best according to WALS and Ethn, while dER comes out better for the Hstr. These numbers are arguably misleading though.

	LDND	dER	dERC
WALS	0.2180	0.2167	**0.2165**
Ethn	0.2391	0.2378	**0.2376**
Hstr	0.2241	0.2232	**0.2230**

Table 3: Triplet distances for LDND, dER and dERC

	GRF				*GQD*		
	LDND	dER	dERC		LDND	dER	dERC
WALS	**0.3806**	0.3832	0.3911	**WALS**	**0.1034**	0.1141	0.1077
Ethn	0.5107	**0.5068**	0.5163	**Ethn**	**0.1381**	0.1440	0.1417
Hstr	0.5415	**0.5339**	0.5357	**Hstr**	0.1530	**0.1508**	0.1625

Table 4: Generalized Robinson-Foulds distances and Generalized quartet distances for LDND, dER and dERC

The GRF relies on the Neighbor Joining tree, which is quite sensitive to the properties of the specific data set. This can be illustrated with the following little experiment. 10 mutually disjoint subset of ASJP were drawn, each containing 275 word lists. For each of these subsets, the Neighbor Joining trees for LDND, dER and dERC were computed and compared to the WALS classifaction according to GRF and GQD. The results are shown in Tab. 5.

GRF			*GQD*		
LDND	dER	dERC	LDND	dER	dERC
0.3696	**0.3696**	0.3913	**0.1049**	0.1255	0.1382
0.4138	0.3793	**0.3621**	0.1527	**0.0753**	0.1105
0.2778	0.2407	**0.2222**	**0.1022**	0.1109	0.1066
0.4118	**0.3922**	**0.3922**	0.1071	0.1129	**0.0607**
0.4694	0.5306	0.5306	**0.1103**	0.1131	0.1205
0.5333	0.5778	**0.5333**	**0.1064**	0.1086	0.1292
0.3333	**0.2745**	0.3137	0.1042	0.1096	**0.0814**
0.4255	0.4468	0.4468	0.1536	0.1056	**0.0686**
0.3750	0.4167	0.4375	**0.0845**	0.1175	0.1124
0.5306	0.5306	**0.5102**	0.1897	**0.1619**	0.1724
mean 0.41401	0.41588	**0.41399**	*mean* 0.12157	0.11409	**0.11007**

Table 5: Generalized Robinson-Foulds and quartet distances for ten random samples

Both the numerical values of GRF and GQD and the relative ordering of the three distance measures differ widely. For instance, LDND leads to the lowest GRF value five times, dER three times, and dERC five times.

To detect the quality of different distance measures despite the noisyness of phylogenetic inference, I drew 1,000 random sample form the 5,000+ ASJP word lists, each containing 500

	GRF				*GQD*		
	LDND	dER	dERC		LDND	dER	dERC
WALS	0.4346	0.4379	**0.4337**	**WALS**	**0.1894**	0.1938	0.1908
Ethn	0.4888	0.4884	**0.4867**	**Ethn**	**0.2181**	0.2230	0.2201
Hstr	0.4862	0.4844	**0.4798**	**Hstr**	**0.2141**	0.2184	0.2153

Table 6: Generalized Robinson-Foulds and quartet distances for 1,000 random samples

word lists and averaged over the various tree distance measures to the expert classifications.[8] The results are given in Tab. 6 and the distributions are visualized as box plots in Fig. 7. From these data we can conclude that dERC gives slightly better results than dER for all evaluations, so the correction for missing entries does have a positive effect. The comparison between LDND and dERC is equivocal. On avarage dERC is slightly better for GRF and slightly worse for GQD.

As the 1,000 samples used here are not stochastically independent,[9] it is not possible to perform meaningful statistical tests, so it is impossible to say whether there are significant differences between the quality of LDND and dERC. In any event, the differences are very small.

5 Weighted string alignment

5.1 The method

Levenshtein alignment only distinguishes between identical and non-identical sounds. To achieve a better approximation of the etymologically correct alignment of cognate words, a graded notion of similarity between sounds seems more appropriate. The ALINE system from Kondrak (2002), for instance, uses a sophisticated hand-crafted notion of segment similarity that draws on insights from phonology. ALINE will be discussed in more detail in the next section. In this section an alternative approach will be reviewed that has been used in previous work in bioinformatics (see for instance Durbin et al., 1989) and computational dialectometry (cf., among others, Wieling et al., 2009).

The approach emerges from similar considerations that we used in the previous section in the derivation of the ER measure.[10] Suppose we want to compare two strings x and y — such as sequences of DNA bases or of protein molecules, or words that we suspect to be

[8]The choice of exactly 1,000 samples containing exactly 500 languages each is arbitrary. The criterion for choosing these numbers was that the number of samples should be suffiently large to be able to detect trends, and that each sample should be not too small, but small enough to make 1,000 iterations computationally feasible.

[9]It might seem suggestive to evaluate the different distance measures for the individual language families and to average the results because different language families are our best approximation of independent samples when it comes to cross-linguistic data. This protocol has been followed for instance by Pompei et al. (2011). Such a procedure strikes me as misleading though because it only assesses how well the *internal* classification of language families are recoverable based on the different distance measures. It is equally important to take into account, however, how well the competing measures separate different language families. My somewhat pessimistic conclusion is that it is not possible to create sufficiently many independent samples from cross-linguistic data that are both independent from each other and representative for the population as a whole.

[10]The following discusssion draws heavily on Durbin et al. (1989).

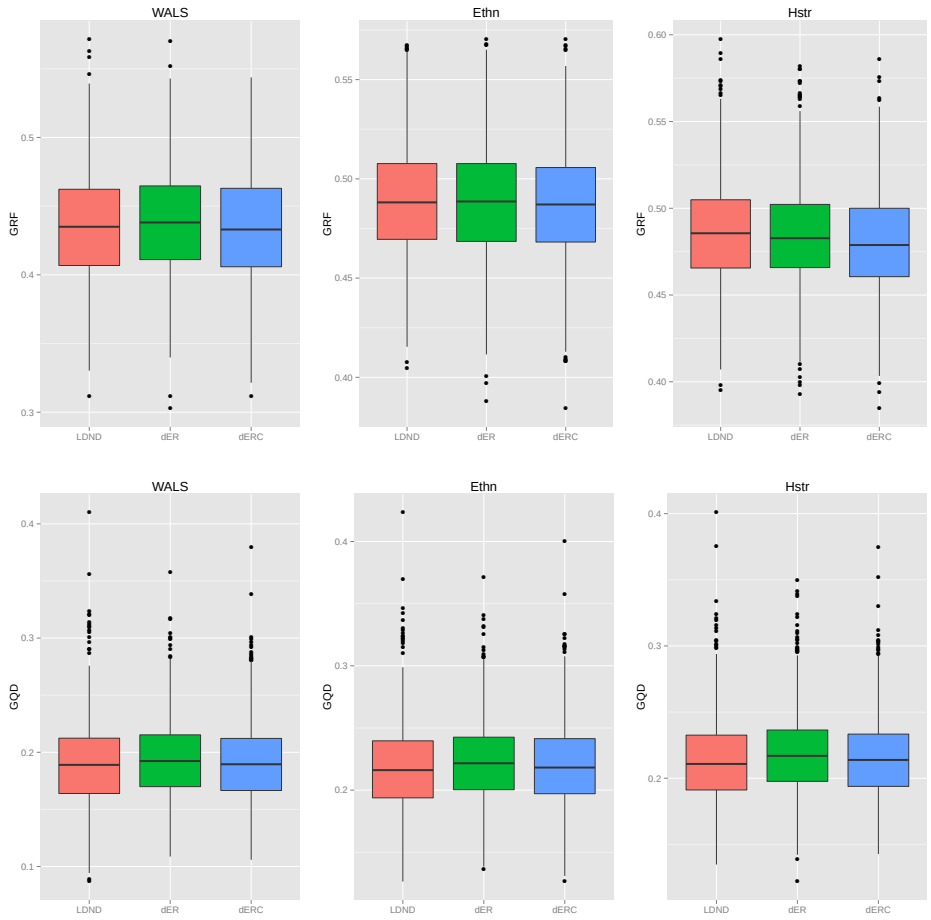

Figure 7: Generalized Robinson-Foulds and quartet distances for 1,000 random samples

cognates — and figure out whether or not they developed from a common ancestor. We have two hypotheses: H_0: "x and y are related." and H_1: "x and y are unrelated.". According to Bayesian logic, it holds that

$$\begin{aligned} \frac{P(H_1|x,y)}{P(H_0|x,y)} &= \frac{P(x,y|H_1)}{P(x,y|H_0)}\frac{P(H_1)}{P(H_0)} \\ \log\frac{P(H_1|x,y)}{P(H_0|x,y)} &= \log\frac{P(x,y|H_1)}{P(x,y|H_0)} + \log\frac{P(H_1)}{P(H_0)} \end{aligned}$$

The prior odds $P(H_1)/P(H_0)$ are unknown, so we focus on the first term on the right hand side, which expresses the strength of the evidence of the particular data point for H_1.

We start with the (unrealistic) assumption that in case H_1 is true, no insertions or deletions of segments have taken place, so x and y have the same length. If H_1 is true, the ith segments of x and y are pairwise historically related. Under the simplifying assumption that point mutations in biological evolution and individual sound shifts in language change are mutually independent, the probability of observing strings x and y given H_1 is the product of the

individual probabilities that x_i and y_i are historically related:

$$\begin{aligned} P(x, y, |H_1) &= \prod_i P(x_i \text{ and } y_i \text{ developed from a common ancestor}) \\ \log P(x, y, |H_1) &= \sum_i \log P(x_i \text{ and } y_i \text{ developed from a common ancestor}) \end{aligned}$$

Let a and b be two segments. The quantity $s(a, b)$ is defined as the probability that a specific segment in some sequence (biomolecule/word) developed into an a along one phylogenetic branch and into a b along another branch. Under this interpretation, s is symmetric, i.e. $s(a, b) = s(b, a)$. If the substitution matrix S, i.e. the value of $s(a, b)$ is known for all segment pairs a, b, we have

$$\log P(x, y, |H_1) = \sum_i \log s(x_i, y_i)$$

If x and y are unrelated, the pairings of x_i and y_i are just randomly picked segments. Let $q(a)$ be the probability of occurrence of a at an arbitrary position within an arbitrary sequence. With the simplifying assumption that the occurrences of segments at different positions within a sequence are independent of each other, i.e. the sequences have no grammar, we have

$$\log(x, y|H_0) = \sum_i \log(q(x_i)q(y_i))$$

Of course the assumptions made here — stochastic independence of positions within a sequence and of evolutionary changes at different positions within a sequence — are wildly unrealistic. Nevertheless this null model leads to workable results, as we will see later on.

Putting the pieces together, we have

$$\log \frac{P(x, y|H_1)}{P(x, y|H_0)} = \log \sum_i \frac{s(x_i, y_i)}{q(x_i)q(y_i)} \tag{2}$$

In the bioinformatics tradition, this quantity is called the *log odds* score. In computational linguistics, it is also known under the name *Partial Mutual Information* (PMI, see Church and Hanks, 1990).[11] I will follow the latter terminology in the sequel. Some notation:

$$\begin{aligned} PMI(a, b) &= \log \frac{s(a, b)}{q(a)q(b)} \\ PMI(x, y) &= \sum_i PMI(x_i, y_i) \end{aligned}$$

We now turn to the issue of insertions and deletions. Suppose we evaluate a specific hypothesis about the historical relation between x and y, which includes assumptions about segments being inserted or deleted. This leads to an *alignment* between the sequences including gaps. This can be illustrated with the comparison of the German and Swedish words for *star*, *Stern/stjärna*, which are *StErn/SEnE* in the ASJP transcription. The etymologically correct alignment is

[11] The PMI score is defined in terms of the binary rather than the natural logarithm. This difference is inessential though because it amounts to a constant factor.

x:	S	t	E	r	n	-
y:	S	-	E	-	n	E

The gap symbol "-" represents a position where either a segment has been deleted or a segment has been added in the other language. x_i and y_i now refer to the ith positions in the *aligned* strings. In the example, y_2 would be the gap symbol.

Following standard practice in bioinformatics, I assume that there is a uniform PMI score for gaps, regardless of the segment the gap is matched with:

$$PMI(a, -) = PMI(-, a) \quad = \quad -d$$

The constant d, which is positive, is referred to as *gap penalty.*[12]

However, both in biological evolution and in language change, insertions and deletions frequently operate on contiguous chunks of segments. For instance, in language comparison we frequently find instances of *partial cognates*, i.e. word pairs where one item is morphologically complex (or is etymologically derived from a morphologically complex word) and the other word is cognate to just one morpheme. For instance, consider the Latin and Italian words for *mountain*, *mons/montagna*, transcribed as *mons/monta5a* in ASJP. The Italian word is probably derived from the Latin *montaneus* 'mountainous', a denominal adjectivization of *mons*. So the correct alignment is

m	o	n	s	-	-	-
m	o	n	t	a	5	a

The three gaps a the end of the upper sequence are the reflex of a single historical process, i.e. suffixation plus semantic change.

Since gaps frequently come in chunks, the penalty for a gap in x at position $i+1$ should be lower if x_i is also a gap than if x_i is a regular segment. This is captured by the notion of *affine gap penalties*. There are two positive constants d (penalty for opening a gap) and e (penalty for extending a gap) with $e \leq d$ such that if $x_i =$ -:

$$PMI(x_i, y_i) \quad = \quad \begin{cases} -e & \text{if } x_{i-1} \text{ is a gap} \\ -d & \text{else} \end{cases}$$

The same applies *mutatis mutandis* to gaps within y.

With these provisos, the PMI score of an alignment of x and y gives an estimate of the strength of evidence that x, y provide for H_1 under a specific alignment. The upper bound thereof is the maximal PMI score for any alignment of x with y. The *Needleman-Wunsch algorithm* (Needleman and Wunsch, 1970) is a simple generalization of the Levenshtein alignment algorithm that for a given substition matrix S and gap penalties d, e, efficiently (i.e. in quadratic time) finds this optimal alignment and its PMI score.

[12]Durbin et al. (1989) give a probabilistic interpretation of gap penalties, according to which $-d$ is the logarithm of the probability of observing a gap. However, this derivation relies on the tacit assumption that sequences are so long that the can be considered as infinite. As words are rather short, this leads to a systematic over-estimation of gap penalties. Therefore gap penalties have no obvious probabilistic interpretation in the context of computational linguistics.

5.2 Parameter estimation

To reliably estimate the PMI scores for all segment pairs, one would ideally need a very large corpus of correctly alined sequence pairs. In bioinformatics such data bases do indeed exist, and several carefully crafted substitution matrices for different domains have been constructed (see Durbin et al., 1989 for details). In dialectometric work (such as Wieling et al., 2012), such data are fairly easy to construct because dialectometric data are organized in cognate sets, and the linguistically correct alignment between cognate words from different dialects of the same language can be reliably constructed with automatic means.

When dealing with cross-linguistic data from a wide variety of languages such as the ASJP data, the situation is more difficult. Sizeable amounts of expert cognacy judgments only exist for a small number of language families (mainly for Indo-European based on the pioneering word of Dyen et al., 1992 and for Austronesian, see Greenhill et al., 2008). Also, it is the ultimate goal of this entire enterprise to do language classification automatically. Therefore information about language family affiliation should not be utilized for parameter training to avoid circularity.

In the sequel a heuristic method is described to extract a large corpus of probable cognate pairs from the ASJP word lists which can be used for parameter training. The method only relies on the word lists themselves; no additional information about cognacy relations or the genetic affiliation of the languages involved is being used.

To avoid the pitfall of overtraining, I split the ASJP database into two sets of about equal size, the *training set* and the *test set*. For training purposes, only the former is used. The resulting model will then be tested against the latter.

To make sure the two sets are really independent — or at least to approximate this ideal as good as possible with cross-linguistic data — the two sets were constructed in such a way that each WALS family either completely belongs to the training set or the test set.[13] To be more specific, the set of WALS families was placed in a random order and languages and language families were added to the training set in this order as long as its size did not exceed half the size of the entire database. The remaining families constitute the test set. The training set contains $2,723$ and the test set $2,758$ word lists. The lists of language families in the two sets are given in the Online Supporting Material.

Relying on the training set only, I used the following procedure for constructing a sufficiently large corpus of probable cognate pairs, which can be used for parameter training:

- All language pairs that have a dERC distance below a given threshold θ_{dERC} are considered to be *probably related*.
- For a pair of probably related languages L_1 and L_2 and a concept c, all entries for c in the ASJP list for L_1 and in the ASJP list for L_2 are considered. The pair of words with the lowest LDN score is considered as a *potential cognate*.
- All pairs of probable cognates are then alined with the Levenshtein algorithm. If there are multiple optimal alignments, only one of them is considered,[14]

This yields a set of aligned sequence pairs. The quantity $s(a, b)$ is estimated as the relative frequency of alignments of a with b (in either direction) within this corpus. The quantity $q(a)$

[13] This was suggested to me by Eric Holman (p.c.).

[14] To be precise: the implementation of the Levenshtein alignment algorithm I used (the Python package *Levenshtein*) only outputs one alignment even if there are others that are equally well.

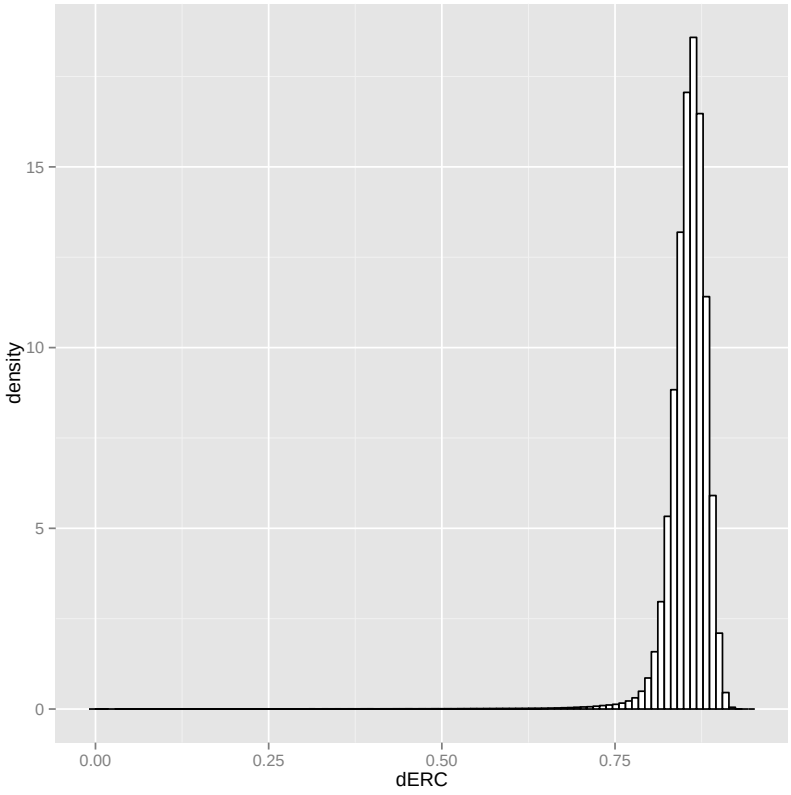

Figure 8: Distribution of dERC scores in the training set

is estimated as the relative frequency of occurrence of the segment type a within the entire ASJP database (or the subset thereof that is used for training purposes). This gives an estimate for

$$PMI(a,b) = \frac{s(a,b)}{q(a)q(b)}$$

for each pair of segment types.

Assuming certain values for the gap penalties (more on this later), as next step the set of potential cognate pairs is aligned with the Needleman-Wunsch algorithm using the estimated parameters.

Additionally I assume a threshold θ_{PMI}. All potential cognate pairs with a PMI score $\geq \theta_{\mathrm{PMI}}$ are treated as *probable cognates*. The set of aligned probable cognates is then used to re-estimate the PMI scores in the way described above.

The re-estimation of parameters is repeated 10 times. Experience shows that the estimated parameter values do not change substanially anymore after that.

It remains to be determined what an appropriate choice for the meta-parameters θ_{dERC} and θ_{PMI} is. As for the former, it is instructive to look at the distribution of dERC distances. It is displayed in Fig. 8 as a histogram.

It can be seen that the distribution is dominated by a bell-shaped curve with the maximum at about 0.85. As pointed out in connection with the derivation of dERC, we do expect that the ERC for unrelated languages approximately follows a standard normal distribution. The

transformation leading from ERC to dERC turns this into a normal distribution with mean at $ERC_{\max}/ERC_{\max}-ERC_{\min} \approx 0.86$, so the histogram shows that the vast majority of language pairs behave under dERC as if they are unrelated. We also see that the distribution is not symmetric — there are more values below than above the predicted mean value. The threshold θ_{dERC} should be chose so that the probability that an unrelated language pair has an dERC$< \theta_{\text{dERC}}$ is very small, while at the same time ensuring that there are still suffiently many language pairs with an dERC $< \theta_{\text{dERC}}$ to make parameter training possible. I picked $\theta_{\text{dERC}} = 0.7$ as a somewhat arbitrary choice which fulfills both requirements. It corresponds to an ERC value of 7.6, and the probability that standard normally distributed variable has a value above that point is about 10^{-14}, which seems to be sufficiently small. On the other hand, there are 20,505 language pairs with a dERC score < 0.7 in the training set, which gives rise to about 7×10^5 pairs of potential cognates. This figure is large enough for parameter estimation.

As pointed out above, there is no straightforward way to estimate gap penalties from a training corpus. Appropriate values for d and e have to be found via optimization. The same holds for θ_{PMI}.

The training procedure supplies a PMI matrix for a given vector θ_{PMI}, d, e, which in turn, together with d and e, define a PMI score for pairs of strings, which is a similarity measure for word pairs. Using the dERC aggregation procedure with LDN scores replaced with negative PMI scores, this gives us a distance measure between languages. Let us call it dERC/PMI.

As a heuristic to assess the quality of a parameter configuration, I sampled 1,000 pairs of probably related languages, i.e. languages with a dERC$< \theta_{\text{dERC}}$. The mean dERC/PMI between these 1,000 language pairs is treated as the target function to be minimized. According to the way dERC/PMI is computed, this amounts to maximizing the ranks of translation pairs, i.e. maximizing the similarity between synonymous word pairs while at the same time minimizing the similarity between non-synonymous word pairs. As we can assume that there are many cognates among translation pairs from probably related languages, minimizing the mean dERC/PMI is tantamount to maximizing similarity between cognates and minimizing similarity between non-cognates.

As even a single evaluation step is computationally quite expensive, advanced methods of optimization such as simulated annealing proved to be impractical. Therefore I performed a simple downhill Nelder-Mead style optimization (cf. Nelder and Mead, 1965), starting from several manually chosen initial positions. The lowest value of the target function was achieved with $d \approx -2.4930$, $e \approx -1.7057$, and $\theta_{\text{PMI}} \approx 4.4451$.[15]

The PMI scores for a selection of sounds are shown in Tab. 7. (The full matrix is given in the Online Supporting Material.) Not surprisingly, the entries along the diagonal are all positive, i.e. alignment of two identical elements provides strongest evidence for relatedness. Additionally, we find positive PMI scores for several sound pairs that are known to be frequently historically related via sound shifts, such as *p/b*, *d/t*, *d/8* and *s/h*. The latter case is especially interesting because the two sounds are articulatorily dissimilar, but the sound shift from *s* to *h* is known to be quite common (see for instance Ferguson, 1990).

Fig. 9 displays a hierarchical clustering of the ASJP sound symbols according to their PMI scores.[16] We find a primary split between vowels and consonants. The consonants are

[15]The value of the target function at this point is 0.5225, while the baseline, i.e. the mean dERC/LDN is 0.5532

[16]To perform the clustering, PMI scores were transformed into distances by subtracting them from the maximal PMI score. For the hierarchical clustering, Ward's method was used; see Ward (1963).

	a	**e**	**i**	**o**	**u**	**p**	**b**	**d**	**t**	**8**	**s**	**h**
a	**1.88**	−1.35	−2.35	−1.66	−2.54	−8.49	−8.82	−7.07	−7.03	−4.64	−8.78	−8.40
e	−1.35	**2.40**	−0.48	−1.52	−2.88	−7.47	−7.80	−7.66	−6.01	−5.01	−7.76	−7.38
i	−2.35	−0.48	**2.37**	−2.81	−1.32	−6.75	−8.46	−8.33	−8.98	−3.48	−7.04	−6.66
o	−1.66	−1.52	−2.81	**2.48**	−0.27	−7.08	−8.10	−7.96	−8.61	−5.31	−8.06	−7.68
u	−2.54	−2.88	−1.32	−0.27	**2.76**	−6.62	−8.05	−7.91	−8.56	−5.26	−8.01	−7.63
p	−8.49	−7.47	−6.75	−7.08	−6.62	**3.69**	**0.36**	−6.59	−4.30	−3.94	−2.70	−0.49
b	−8.82	−7.80	−8.46	−8.10	−8.05	**0.36**	**3.62**	−4.84	−5.09	−3.58	−5.63	−3.24
d	−7.07	−7.66	−8.33	−7.96	−7.91	−6.59	−4.84	**3.41**	−0.10	**2.52**	−2.29	−2.81
t	−7.03	−6.01	−8.98	−8.61	−8.56	−4.30	−5.09	−0.10	**3.15**	**2.11**	−1.67	−1.76
8	−4.64	−5.01	−3.48	−5.31	−5.26	−3.94	−3.58	**2.52**	**2.11**	**5.49**	**1.92**	−0.85
s	−8.78	−7.76	−7.04	−8.06	−8.01	−2.70	−5.63	−2.29	−1.67	**1.92**	**3.50**	**0.26**
h	−8.40	−7.38	−6.66	−7.68	−7.63	−0.49	−3.24	−2.81	−1.76	−0.85	**0.26**	**3.50**

Table 7: PMI scores

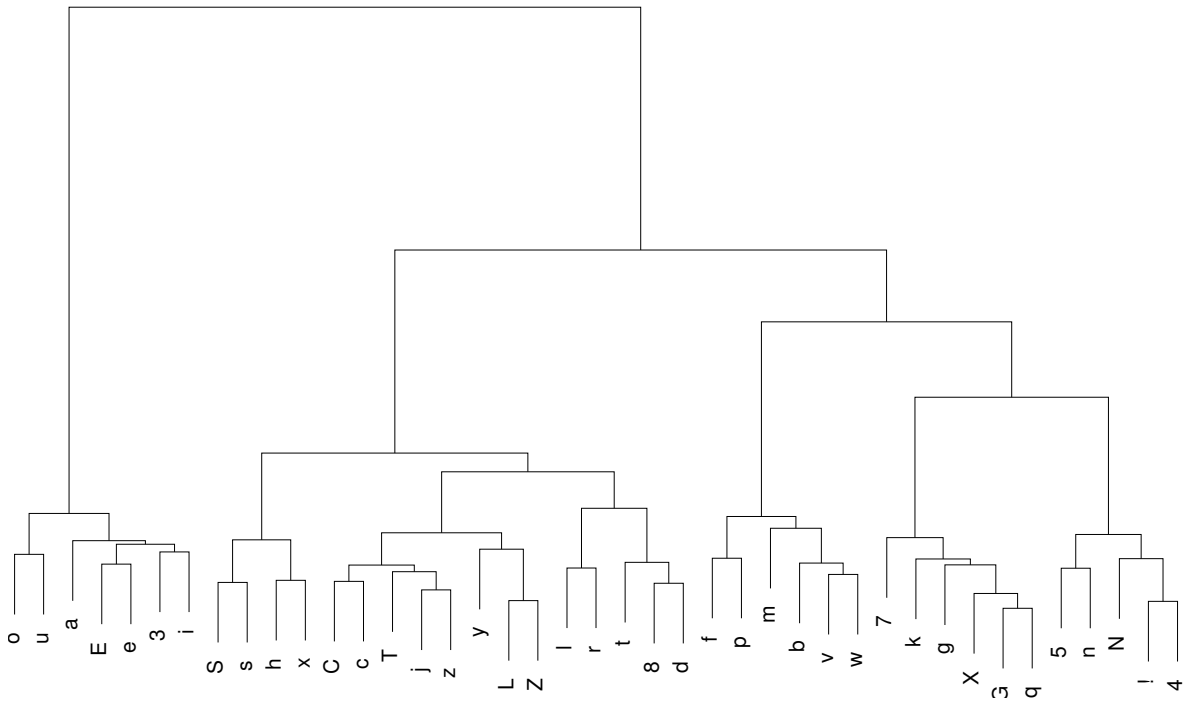

Figure 9: PMI scores: hierarchical clustering

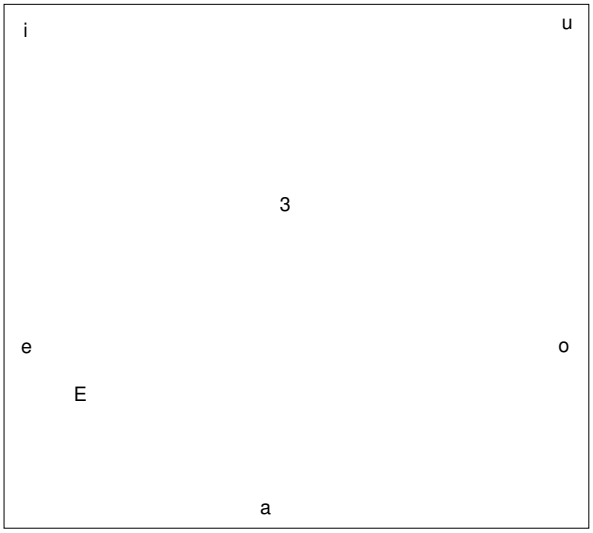

Figure 10: Vowel PMI scores: multidimensional scaling

further divided into three large groups, which largely correspond to the dental, the labial, and the velar/uvular sounds. The only exception to this pattern according to place of articulation is the position of h and x (the voicedless and voiced velar fricatives), which are clustered together with the s-sounds within the larger cluster of dental sounds. This is probably a reflex of the already mentioned diachronic cline from s to h.

Following the example of Wieling et al. (2012) — who obtained PMI scores essentially in the same way but using data from different dialects of the same language —, I performed non-metric multidimensional scaling with the PMI scores among the vowels. The result is displayed in Fig. 10. We find that the articulatory vowel triangle is reproduced to a good approximation, with the schwa (ASJP-symbol *3*) in the center.

Brown et al. (2013) also use the ASJP data to estimate the probability of different sound correspondences across the languages of the world. Their method is quite different from the one developed here, so a comparison of the results provides a certain validity check.

The mentioned authors use a highly conservative heuristics to identify regular sound correspondences. According to this method, a pair of languages A, B exhibits a regular correspondence between the segments x and y if and only if:

- L_1 and L_2 belong to the same genus, and
- there are at least two concepts c such that the ASJP entry for c from A can be transformed into its translation to B by replacing all occurrences of x by y (and vice versa).

To use the running example of the English/Swedish comparison again, there are only two regular correspondences that can be detected from the 40-item word lists: *o-e* (*bon/ ben* 'bone' and *ston/ sten* 'stone'; *liv3r/ lev3r* 'liver' and *si/se* 'see').

A certain genus is *available* for a correspondence *x-y* if both segments, x and y, occur in at least one language within this genus. The *PG* score ("percentage of available genera") of a correspondence is the relative frequency (expressed in percent) of genera exhibiting the correspondence at least once among all genera that are available for that correspondence.

Fig. 11 plots the PG scores of all consonant pairs that have a positive PG score given in the supporting online material from Brown et al., 2013) against their corresponding PMI score.

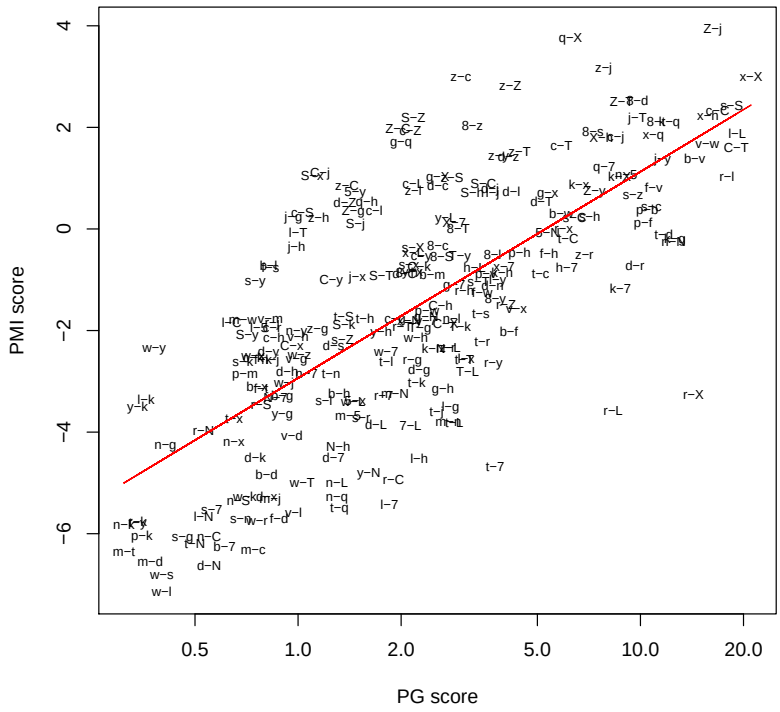

Figure 11: PG scores vs. PMI scores

The x-axis is logarithmically transformed while the y-axis is linear. We see a strong positive relation (indicated by the regression line) between the two measures. In fact, the correlation between the logarithms of the positive PG scores and the corresponding PMI scores is 0.59. This underscores that the empirically induced PMI scores do in fact reflect genuine patterns of regular sound correspondences.

5.3 Aggregation

In Tab. 1 the LDN scores for several English/Swedish word pairs were given. Tab. 8 gives the corresponding PMI scores.

It can be seen that the PMI notion of string similarity is more fine-grained than LDN. For instance, while both *du/yu* and *vi/wi* receive a positive score (i.e. they are more likely to be related than not), the absolute value for the latter is much higher. This reflects the fact that a corresondence between v and w is more likely than one between d and y. The pair *fisk/fiS* has en even higher PMI score because (a) the words are longer than *vi/wi*, i.e. the evidence they provide is stronger, and (b) the correspondence *s/S* is very likely. This is counterbalanced only by a single gap penalty.

	Ei	yu	wi	w3n	tu	fiS	...
yog	**−7.77**	0.75	−7.68	−7.90	−8.57	−10.50	
du	−7.62	**0.33**	−5.71	−7.41	2.66	−8.57	
vi	−2.72	−2.83	**4.04**	−1.34	−6.45	0.70	
et	−5.47	−7.87	−5.47	**−6.43**	−1.83	−4.70	
tvo	−7.91	−4.27	−3.64	−4.57	**0.39**	−6.98	
fisk	−7.45	−11.2	−3.07	−9.97	−8.66	**7.58**	
⋮							

Table 8: PMI scores English/Swedish

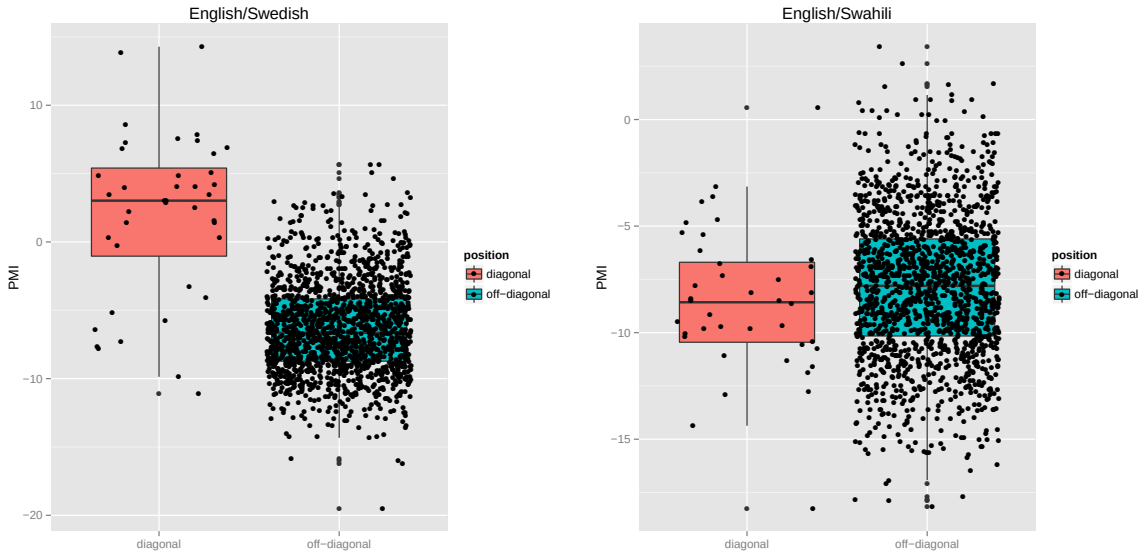

Figure 12: Distribution of diagonal and off-diagonal PMI scores: English/Swedish and English/Swahili

The distribution of PMI scores for the English/Swedish comparison on the diagonal and off the diagonal are shown in the left panel of Fig. 12. The right panel shows the same data for the comparison English/Swahili.

In comparison to the corresponding plots for LDN, the PMI values are much more spread out. Apart from that, we find a similar qualitative pattern (apart from the inessential difference that LDN is a distance and PMI a similarity measure). For a pair of related languages, the diagonal entries are mostly much higher than the off-diagonal entries, while boths collections appear to be drawn from the same distribution for a pair of unrelated languages.

The *normalized ranks* of PMI scores are now computed according to the definition given in Section 3, with LDN scores replaced by PMI scores and $\leq$ / $<$ replaced by $\geq$ / $>$. As shown in Figure 13, the PMI based normalized ranks from related languages follow approximately a power law distribution and those from unrelated languages a uniform distribution, just as the LDN based normalized ranks. Therefore the theoretical justification for the dERC style

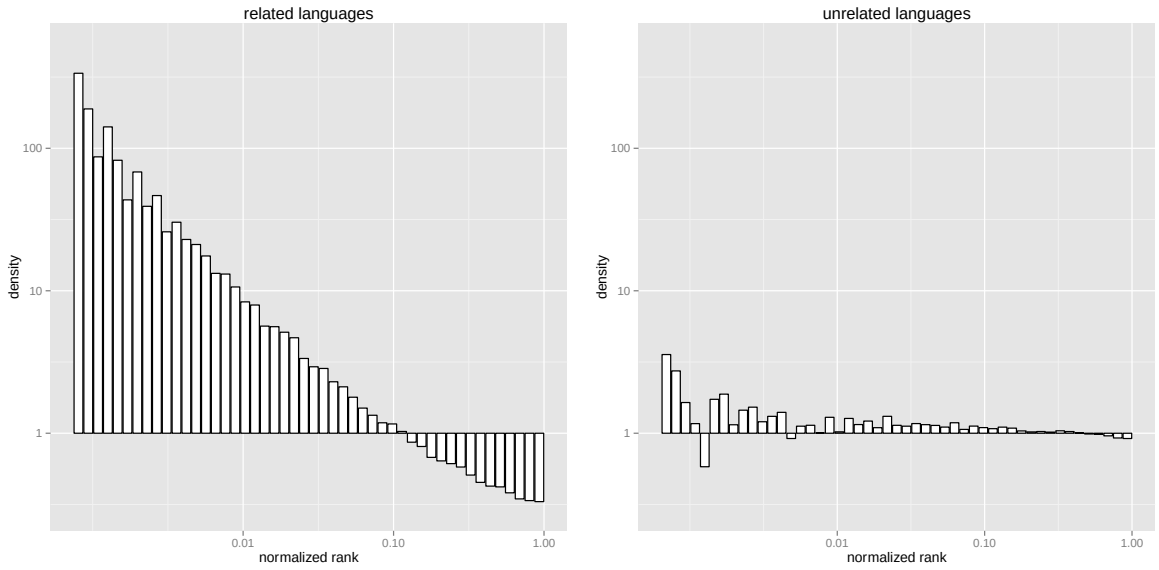

Figure 13: Distribution of normalized PMI ranks from related languages; log-log scale

aggregation of normalized ranks to a distance measure between languages also applies to PMI scores.

In Tab. 9 the dERC/LDN scores and dERC/PMI scores for the language pairs from Tab. 2 are compared.

These numbers convey the impression that the dERC/PMI scores for related languages are generally lower than the corresponding dERC/LDN scores, while the scores for unrelated languages are randomly distributed around 0.86 for both measures.

5.4 Empirical evaluation

The methods described in Section 4 to compare different distance measure will now be used to evaluate the quality of the dERC/PMI against LDND and the LDN-based version of dERC (referrred to as dERC/LDN henceforth). Only the word lists from the test set will be used for this comparison.

The triplet distances to the three expert classifications are given in Tab. 10 and visualized in Fig. 14. We find a slight improvement from LDND to dERC/LDN and a more substantial improvement from dERC/PMI.[17]

From the distances matrices for the test set for LDND, dERC/LDN and dERC/PMI, the corresponding phylogenetic trees where computed with Neighbor Joining. The generalized Robinson-Foulds distances and quartet distances are given in Tab. 11.

[17]It might be surprising that the triplet distances given in Tab. 3—that were calculated for the entire ASJP—are in the 20% range, while the values for the test set are in the $10-15\%$ range. This reflects the fact that the task of automatically classifying a given set of word lists has something like an inherent level of difficulty. The low scores for the test set might have something to do with the fact that almost one third of it are Austronesian languages. Therefore a substantial proportion of triplets to be evaluated consist of two Austronesian and one non-Austronesian language, and the signal distinguishing Austronesian from the rest of the world's languages is fairly strong.

A	B	$dERC/LDN(A,B)$	$dERC/PMI(A,B)$
English	English	0.0812	0.0078
English	Scots	0.2214	0.2139
Danish	Swedish	0.3431	0.2773
English	Swedish	0.4866	0.3981
English	Frisian	0.4986	0.4215
English	Dutch	0.5071	0.4040
Hindi	Farsi	0.6536	0.6231
English	French	0.7917	0.7720
English	Hindi	0.8276	0.7735
Amharic	Vietnamese	0.8401	0.8566
Swahili	Warlpiri	0.8619	0.8573
Navajo	Dyirbal	0.8435	0.8436
Japanese	Haida	0.8547	0.8504
English	Swahili	0.8687	0.8901

Table 9: dERC scores

	LDND	dERC/LDN	$dERC/PMI$
WALS	0.1218	0.1171	**0.0898**
Ethn	0.1521	0.1485	**0.1235**
Hstr	0.1423	0.1398	**0.1202**

Table 10: Triplet distances for LDND, dERC/LDN and dERC/PMI

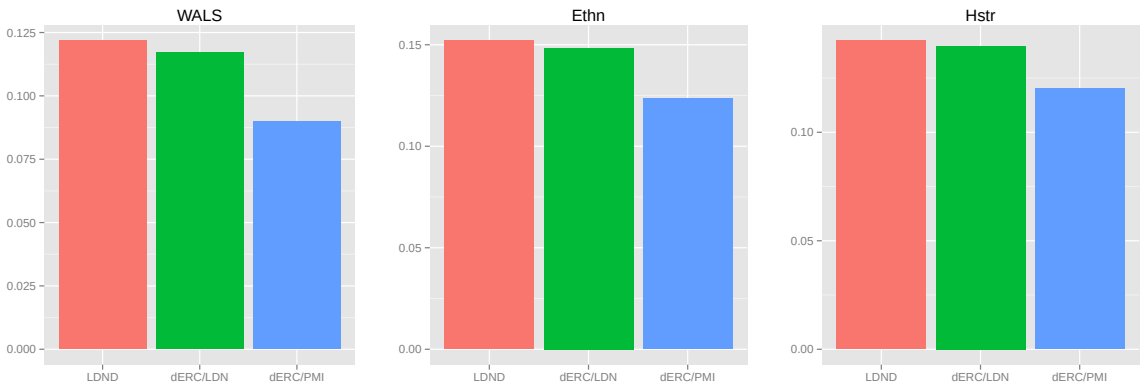

Figure 14: Triplet distances for LDN, dERC/LDN and dERC/PMI

GRF

	LDND	dERC/LDN	dERC/PMI
WALS	0.3503	0.3559	**0.3390**
Ethn	0.5317	0.5469	**0.5293**
Hstr	0.5547	0.5622	**0.5460**

GQD

	LDND	dERC/LDN	dERC/PMI
WALS	0.0830	**0.0490**	0.0915
Ethn	0.1454	**0.1177**	0.1443
Hstr	0.1659	**0.1346**	0.1631

Table 11: Generalized Robinson-Foulds distances and Generalized quartet distances for LDND, dERC/LDN and dERC/PMI

GRF

	LDND	dERC/PMI	dERC/PMI
WALS	0.3929	0.3938	**0.3612**
Ethn	0.4773	0.4805	**0.4667**
Hstr	0.4939	0.4881	**0.4784**

GQD

	LDND	dERC/LDN	dERC/PMI
WALS	0.1032	0.1004	**0.0842**
Ethn	0.1538	0.1558	**0.1422**
Hstr	0.1646	0.1669	**0.1577**

Table 12: Generalized Robinson-Foulds and quartet distances for 1,000 random samples

The results are not decisive, with dERC/PMI giving the lowest GRF scores and dERC/LDN the lowest GQD scores. However, as discussed in Section 4, evaluating different Neighbor-Joining trees for a single data set can be highly misleading. Therefore the same procedure as above is applied here: 1,000 random samples of word lists from the test set, each comprising 500 doculects are generated, Neighbor Joining trees for LDND, dERC/LDN and dERC/PMI are computed and all three trees are compared to the three expert trees both regarding GRF and GQD. The results are depicted in Fig. 15 and the mean values are given in Tab. 12.

The mean values for the 1,000 samples display a similar pattern as the triplet distances: LDND and dERC/LDN perform about equally well (with a slight advantage for the latter), while dERC/PMI leads to lower distance scores. As the aggregation method for dERC/LDN and dERC/PMI is identical, we can conclude that the PMI based method of measuring string similarities leads to better phylogenetic inference than (normalized) Levenshtein distance.

As a further test, I performed a version of *cross-validation*.[18] In general, k-fold cross-

[18]This was suggested by an anonymous reviewer.

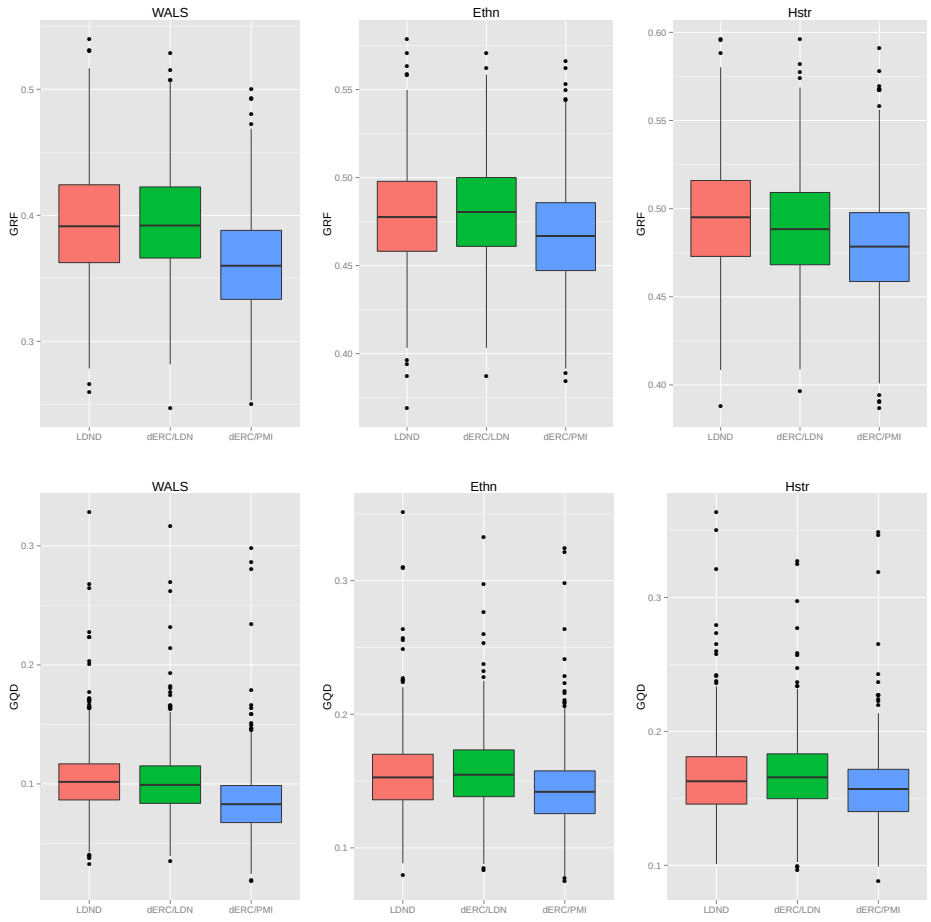

Figure 15: Evaluation results: distribution of fit measures for 100 random samples of 1,000 languages each

validation means that a data set is split into k subsets of equal size. Then on subset is singled out as test set. The remaining data are used for training, and the model thus obtained is tested against the test set. This is repeated for each subset.

Cross-validation requires the individual subsets to be independent from each other. As discussed above, obtaining mutually independent subsamples of a cross-linguistic database such as ASJP that are representative for the data set as a whole is a non-trivial issue. As an approximation, I performed 4-fold cross-validation, where the subsets correspond to the four continental areas Africa (including all Afro-Asiatic languages), Eurasia, the Indo-Pacific region (including Australia), and America.

For each continental area c, PMI scores where trained with the languages outside c following the method described in the previous section. For the threshold θ_{PMI} and the gap penalties d and e, the values obtained from the original training set were used. Using the thus induced PMI scores, the pairwise dERC/PMI-values for the languages in c where computed and the triplet distance to WALS, Ethn and Hstr where determined and compared to the corresponding values for LDND and dERC/LDN. The results are given in Tab. 13 and displayed in Fig. 16.

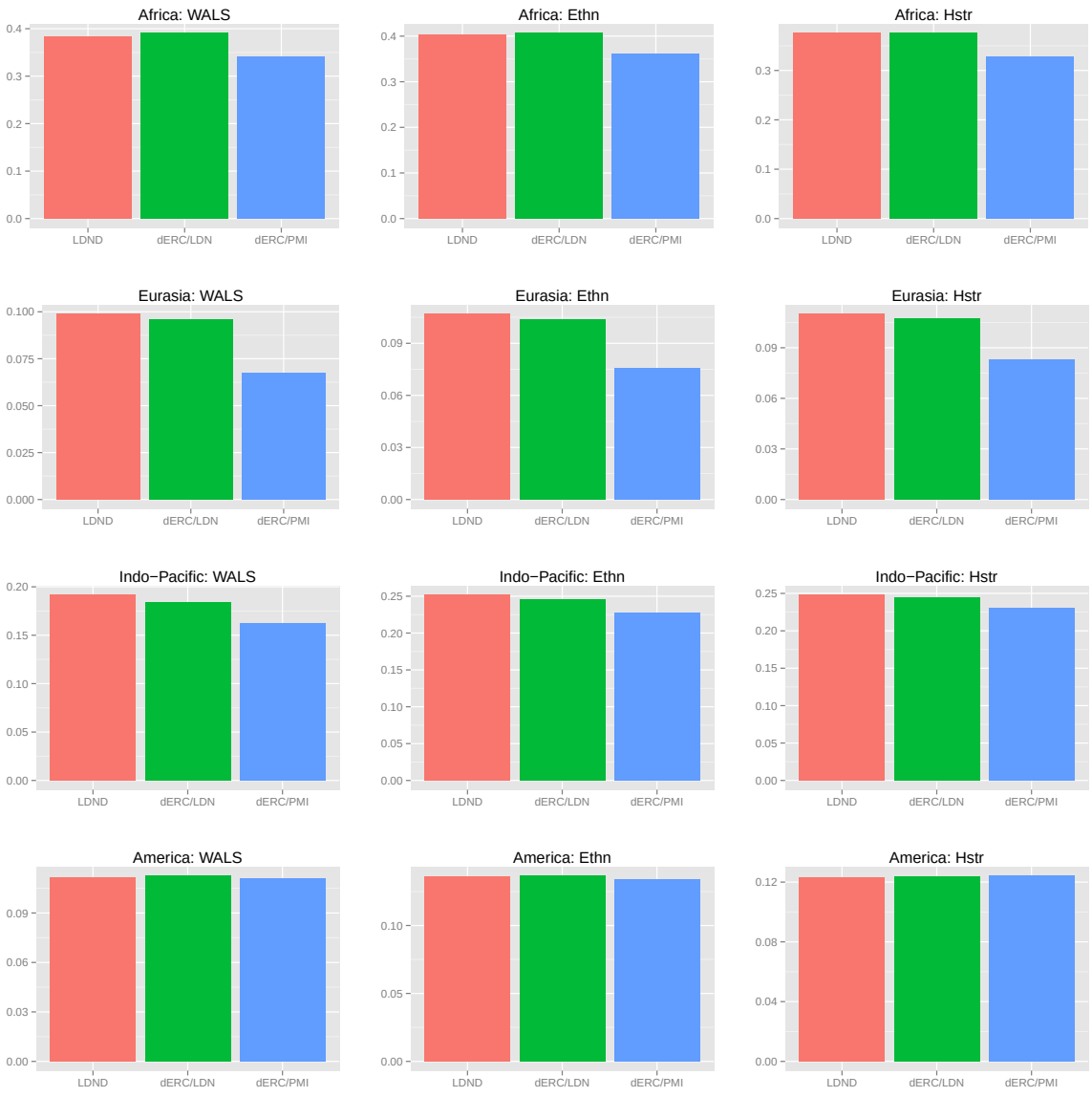

Figure 16: Triplet distances for LDN, dERC/LDN and dERC/PMI: Continental areas

	Africa			*Eurasia*		
	LDND	dERC/LDN	dERC/PMI	LDND	dERC/LDN	dERC/PMI
WALS	0.3843	0.3909	**0.3414**	0.0988	0.0958	**0.0675**
Ethn	0.4026	0.4068	**0.3613**	0.1069	0.1040	**0.0757**
Hstr	0.3758	0.3757	**0.3276**	0.1100	0.1073	**0.0832**

	Indo-Pacific			*America*		
	LDND	dERC/LDN	dERC/PMI	LDND	dERC/LDN	dERC/PMI
WALS	0.1916	0.1844	**0.1624**	0.1117	0.1125	**0.1107**
Ethn	0.2518	0.2462	**0.2271**	0.1358	0.1365	**0.1342**
Hstr	0.2480	0.2451	**0.2306**	**0.1233**	0.1238	0.1242

Table 13: Triplet distances for LDND, dERC/LDN and dERC/PMI: Continental areas

In 11 out of 12 cases, dERC/PMI provides the best results (the exception being the Hstr classification for America, where LDND is slightly better). The general pattern for Africa, Eurasia and the Indo-Pacific is similar to the test set above: LDND and dERC/LDN are about equally good, while dERC/PMI is about 2%-3% better. For America, all three distance measures perform roughly equally well.

The average correlation of the PMI-matrices obtained during cross-validation with the PMI-matrix obtained from the training set is 0.89, and the average correlation between the four PMI-matrices from cross-validation is 0.92. This indicates that the patterns of regular sound correspondences across different samples of language families are highly similar.

5.5 Discussion

A possible objection against the general approach developed here concerns the risk of circularity. As an anonymous reviewer points out, it might be problematic to perform automatic language classification on the basis of parameters that are trained with data from a database "which was [...] obtained through some other type of (manual) analysis". Let us therefore carefully review what kind of information goes into the training procedure and what kind of information we get out of it.

The construction of the training corpus of word pairs relied on guessing a value for θ_{dERC}. The guess of $\theta_{\mathrm{dERC}} = 0.7$ is to some degree arbitrary, but it was motivated by a visual inspection of the distribution of dERC/LDN scores.

dERC/LDN scores are determined on the basis of pairwise LDN scores for words from the word lists to be compared. No further information about the genetic affiliation of the languages involved is being used here, and LDN scores are obtained from Levenshtein-distances, a general-purpose string comparison method that does not rely on any specifically linguistic information.

Once the training corpus is constructed, initial PMI scores are estimated using Levenshtein alignment. In subsequent steps, Needleman-Wunsch alignment is performed ten times, each time using the PMI scores estimates from the previous step. For given values of θ_{PMI} and the gap penalties d and e, the PMI scores thus obtained define a string similarity score which is in turn fed into the dERC aggregation scheme to yield a distance measure between word

lists. $\theta_{\mathrm{PMI}}, d, e$ are estimated via optimization in a way that minimizes the average distance between a sample of language pairs. This sample in turn was collected using only dERC/LDN scores and θ_{dERC}. So the only information that enters the entire procedure are ultimately the plain word lists. No knowledge about the languages involved is used anywhere in parameter training. In the terminology of machine learning, PMI scores are obtained via *unsupervised learning*.

The test procedures in turn use the aggregate distance between word lists thus obtained to do phylogenetic inference and to compare the results to expert classifications. (Triplet distance relies on classifying triplets of languages, so this also involves a kind of phylogenetic inference). So the information that is obtained from the parameterized model — language classification — is of an entirely different nature than the information that went into it, namely word lists.

Another potential objection concerns the fact that the overall gain in accuracy — about 3% for triplet distances and 1%-2% for GRF and GQD — may still appear small. However, three considerations should be kept in mind here:

- Both for the triplet distance and for GQD, the baseline of completely randomly distributed distances is not 1.0 but 0.67 because there are only three rooted binary trees for a triplet and three butterflies for a quartet of languages.

- Even very crude distance measures achieve a much higher accuracy as suggested by these base lines. To illustrate this point, I defined such a crude measure: for each word list, the vector of relative frequencies of occurrence of sounds are computed. The *cosine distance* between to languages is then defined as 1− the cosine of these vectors. So this distance measure only quantifies how much the frequency patterns of unigrams differ between word lists, without any reference to the meaning of the words. The Neighbor Joining tree derived from these distances for the entire ASJP database already achieves a GQD of 0.32 to WALS, 0.35 to Ethn and 0.40 to Hstr.

- The practically achievable minimum GQD (and likewise for triplet distance and GRF) is arguably somewhat above 0. First, the expert classifications contain controversial units (such as Altaic, Australian, Niger-Congo and Trans-New Guinea in WALS), which may partially be wrong. In this case it would not be a defect of an automatic classification if those units are not detected. Second, the 40-item Swadesh lists arguably do not always contain the information that human experts would need to establish a genetic relationship between a group of languages.

To make a rough guess, the maximum GQD (achievable by a simple-minded distance measure such as the cosine distance) for a given data set may be around $35\% - 40\%$ and the minimum GQD that can possibly be attained by automatic methods from 40-item Swadesh lists may by around 3%. Each gain in accuracy of a certain percentage thus actually amounts to a much higher proportion (by a factor of about 3) of this range.

6 Comparison to ALINE

The PMI scores for word similarities used here are obtained via weighted string alignment. There have been several proposals in the literature on computational historical linguistics and computational dialectometry to employ weighted alignment for this purpose. Some of them

use empirically determined log-odds scores as weights like the present propocal (cf. Wieling et al., 2012), while others (see for instance Covington, 1996, Somers, 1998, Heeringa, 2004, among others) assume linguistically motivated hand-crafted substitution weights for segment pairs. The most sophisticated approach along the latter lines is perhaps the ALINE system by Kondrak (2002). A detailed discussion of ALINE would go beyond the scope of this article, so I will just mention the essential features.

In ALINE, each sound is represented by a vector of phonetic features, such as *syllabic*, *back*, *place* etc. These features have real numbers as values. The similarity between two segments is computed from their differences in feature values, weighted by the salience of these features.

Additionally, ALINE captures *compressions* and *expansions*, i.e. alignments of a single segment in one word with two adjacent segments in the other word. Kondrak uses the cognate pair Latin *factum*/ Spanish *hecho* 'fact' to illustrate this point. In the etymologically correct alignment the Spanish affricate [ʧ] should be matched with the [t] and the [k] in the Latin word simultaneously. ALINE defines weights for aligning a single sound with a consecutive sequence of two sounds as well.

The present proposal uses the Needleman-Wunsch algorithm for string alignment. This algorithm finds the optimal *global* alignment, i.e. an alignment of the full sequences. ALINE instead uses *half-global* alignment. This means that in both strings to be compared, final subsequences can be ignored if this leads to a better alignment score. Half-global alignment is motivated by the observation that the left periphery of words is especially unstable in language change.

In Huff (2010) and Huff and Lonsdale (2011), the system PyAline is described, a freely available Python implementation of ALINE that includes substitution scores of ASJP sound classes. PyAline also contains an implementation of Downey et al.'s (2008) method to aggregate ALINE alignment scores to distances between languages. This facilitates a comparison with the distance measures defined here. In Huff and Lonsdale (2011) such a comparison with LDND is discussed. The authors conclude that both measures perform about equally well in phylogenetic inference.

Downey et al.'s aggregation method differs in two essential ways from ERC. First, word similarities are *normalized*. Given alignment scores (which are similarity scores), the normalized ALINE distance between two words is defined as

$$d_{\text{ALINE}}(x, y) = 1 - \frac{2 \times \text{similarity}(x, y)}{\text{similarity}(x, x) + \text{similarity}(y, y)}$$

Second, Downey et al. (2008) define the distance between two languages as the average normalized ALINE distance between translation pairs. This amounts to taking the average of the diagonal matrix of individual word distances, while the off-diagonal entries are not taken into account. Let us call this distance measure d_{Dow}.

These differences in detail make a comparison to dERC difficult, because it has to be factored out whether possible differences in performance are due to the different alignment weights, the different alignment algorithm, the normalization step or the difference in the aggregation scheme. As an additional complication, PyAline's alignment algorithm is implemented in plain Python, which makes it comparatively slow. There are highly efficient Python libraries for the Needleman-Wunsch algorithm used for the computation of PMI scores, which makes the computation of a pairwise dERC/PMI distance matrix for several thousands of

	LDND	dERC/PMI	d_{Dow}	dERC/ALINE
WALS	0.1221	**0.0879**	0.1385	0.1068

Table 14: Estimated triplet distances

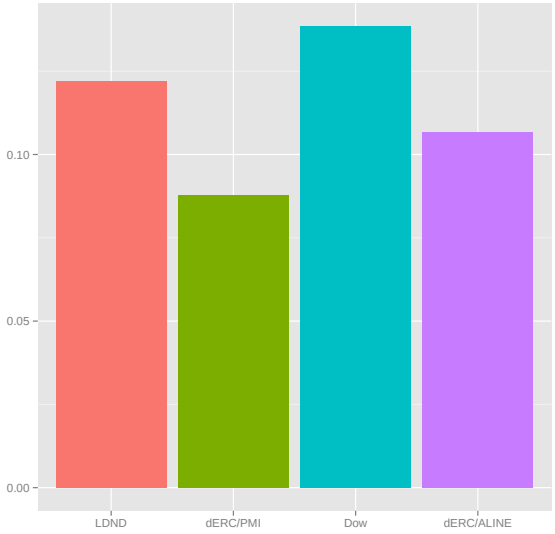

Figure 17: Estimated triplet distances

word lists feasible. For PyAline this is not realistic.[19]

For these reasons, I will defer a detailed comparison of the present proposal with ALINE to another occasion and only report the results of a small pilot study here that could be carried out with moderate computational effort.

From the test set, 10,000 triplets were sampled that are resolved according to WALS. They were used to estimate the triplet distance to WALS for (a) LDND, (b) dERC/PMI, (c) d_{Dow}, and (d) dERC/ALINE. The latter measure uses the normalized ALINE distances between words and aggregates them according to the dERC scheme.

The results are given in Tab. 14 and displayed in Fig. 17.

The estimates for LDND and dERC/PMI are 0.1221 and 0.0879, while the correct values are 0.1218 and 0.0898 respectively (see Tab. 10). This suggests that the estimates are actually quite accurate.

The results indicate that d_{Dow} is substantially worse than LDND. This is arguably due to the aggregation method rather than the ALINE method of computing word distances though. Combining ALINE word distances with dERC-style aggregation gives results that are better than LDND but still worse than dERC/PMI.

With the proviso that these results are still preliminary, they seem to suggest (a) that weighted alignment improves the accuracy of phylogenetic inference in comparison to plain

[19] On the hardware currently at my disposal, computing the distance matrix for the full test set with PyAline would take more than a week.

Levenshtein-style alignment, and (b) that empirically determined PMI scores are superior to hand-crafted weighting schemes.

7 Conclusion

This paper aims at making three contributions to the current discussion in the field of computational historical linguistics: (1) it argues for the usage of weighted alignment using empirically obtained weights for determining word distances, (2) it proposes a novel method to aggregate word similarities/distances to distances between languages, and (3) it presents several protocols for evaluating automatically generated phylogenies that extend existing proposals.

The results from the previous sections show that weighted alignment improves the accuracy of language distance measures when compared to Levenshtein distance methods. The method used here — the Needleman-Wunsch algorithm using log-odds scores and affine gap penalties — was developed in the context of bioinformatics and is justified by the properties of biomolecular evolution. The model assumptions that underly its mathematical foundations are actually not met in the case of sound change. It rests on the simplifying assumptions that mutations at different positions are stochastically independent and that mutation probabilities are constant across lineages. The latter assumption, especially, is highly problematic when applied to sound change since specific sound changes are known to be historically contingent events that apply to the entire lexicon of a language. Therefore a more adequate model would have to use a different substitution matrix for each pair of related languages which captures the history of sound changes along the two lineages from the latest common ancestor. It is in principle possible to obtain these substitution matrices empirically, but this would arguably require much larger word lists than the commonly used Swadesh lists.

Also, work on automatic cognate recognition (see, for instance, List, 2012) has shown that the quality of word alignments improves considerably if *multiple sequence alignment* is used. It is to be expected that language distance measures using multiple alignments will also lead to more accurate phylogenetic inference. An additional advantage of using multiple sequence alignments is that they can be used for character-based methods, which are known to be more accurate than distance-based methods.

A further direction that may lead to higher accuracy is the usage of resampling methods such as bootstrapping and jackknifing, which can be used at various points in the inference process. In this paper, individual word alignment scores were calibrated by comparing them to the distribution of alignment scores across all pairs of non-synonymous word pairs from the two languages to be compared. Sampling a large number of these scores with replacement will arguably lead to a more accurate estimate of this distribution. Furthermore, sampling 40 Swadesh concepts with replacement a large number ($\geq 1,000$) of times and doing phylogenetic inference with each sample individually will result in a large number of slightly different inferred trees. These can be used to generate a consensus tree and to quantify the confidence in the language grouping thus obtained.[20]

Regarding the evaluation described in the previous section, the main innovation presented here is the use of a large collection of random samples of languages to assess the quality of

[20]This kind of *bootstrapping* is standardly being used in character based phylogenetic inference, including work in historical linguistics such as Gray and Atkinson (2003).

a distance measure. According to my own experience, results obtained in this way are much more robust and informative than evaluation results for a single collection of languages.

Acknowledgments

The work being described in this article benefited considerably from discussions with Johann-Mattis List, Taraka Rama, Søren Wichmann and Martijn Wieling, which is gratefully acknowledged. Kate Bellamy, Michael Dunn, Eric Holman, Søren Wichmann and three anonymound reviewers from LDC pointed out various mistakes in a previous version of this article. Thanks also to Thomas Zastrow for setting up the hardware which made this work possible.

Software used

All word alignments and distance measure computations were performed using (Numeric) Python. Levenshtein alignment and Needleman-Wunsch alignment were done using the *Levenshtein* package and the *pairwise2* module of the *Biopython* package (Cock et al., 2009; `http://biopython.org`) respectively.

For the Neighbor Joining algorithm, Joseph Felsenstein's *Phylip* package (Felsenstein, 1989; `http://evolution.genetics.washington.edu/phylip/`) was used. Quartet fits were computed with Christian Pedersen's *qdist* package (`http://birc.au.dk/software/qdist/`). Thanks to its author and to Thomas Mailand for their help in finding and installing this software.

For manipulating and visualizing phylogenetic trees as well as for computing Robinson-Foulds distances, the Python toolkit *ETE* (`http://ete.cgenomics.org/`) and Daniel Huson's *Dendroscope* software (`http://ab.inf.uni-tuebingen.de/software/dendroscope/`) proved highly useful.

Online Supporting Material

Descriptions of the training set and the test set, as well as the PMI scores obtained in the way described in Subsectino 5.2, are contained in an online document that can be downloaded from `http://www.sfs.uni-tuebingen.de/~gjaeger/publications/ldcBenchmarkingSI.pdf`.

References

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337: 957–960.

Brown, Cecil H., Eric Holman, and Søren Wichmann. 2013. Sound correspondences in the world's languages. *Language* 89: 4–29.

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16: 22–29.

Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review* 51: 661–703.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25: 1422–1423. doi:10.1093/bioinformatics/btp163.

Covington, Michael A. 1996. An algorithm to align words for historical comparison. *Computational linguistics* 22: 481–496.

Downey, Sean S., Brian Hallmar, Murray P. Cox, Peter Norquest, and J. Stephen Lansing. 2008. Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction. *Journal of Quantitative Linguistics* 15: 340–369.

Dunn, Michael, Angela Terrill, Ger Ressink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309: 2072–2075.

Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1989. *Biological Sequence Analysis*. Cambridge, UK: Cambridge University Press.

Dyen, Isidore, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society* 82: 1–132.

Estabrook, George F., F. R. McMorris, and Christopher A. Meacham. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Biology* 34: 193–200.

Felsenstein, Joseph. 1989. Phylip — Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164–166.

——. 2004. *Inferring Phylogenies*. Sunderland: Sinauer Inc. Publishers.

Ferguson, Charles A. 1990. From esses to aitches: identifying pathways of diachronic change. In Croft, William A., Suzanne Kemmer, and Keith Denning (eds.) *Studies in Typology and Diachrony: Papers presented to Joseph H. Greenberg on his 75th birthday*, Philadelphia: John Benjamins, 59–78.

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426: 435–439.

Greenhill, Simon J. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics* 37: 689–698.

Greenhill, Simon J., Robert Blust, and Russell D. Gray. 2008. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4: 271–283.

Hammerström, Harald. 2010. A full-scale test of the language farming dispersal hypothesis. *Diachronica* 27: 197–213.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie. 2008. The World Atlas of Language Structures online. Max Planck Digital Library, Munich. http://wals.info/.

Heeringa, Wilbert Jan. 2004. *Measuring Dialect Pronunciation Difference using Levenshtein Distance*. PhD dissertation, University of Groningen.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Advances in automated language classification. In Arppe, Antti, Kaius Sinnemäki, and Urpu Nikanne (eds.) *Quantitative Investigations in Theoretical Linguistics*, University of Helsinki, 40–43.

Huff, Paul. 2010. *PyAline: Automatically Growing Language Family Trees Using The ALINE Distance*. PhD dissertation, Brigham Young University.

Huff, Paul and Deryle Lonsdale. 2011. Positing language relationships using ALINE. *Language*

Dynamics and Change 1: 128–162.

Kondrak, Grzegorz. 2002. *Algorithms for Language Reconstruction*. PhD dissertation, University of Toronto.

Lewis, M. Paul (ed.) . 2009. *Ethnologue: Languages of the World*. SIL International, Sixteenth ed. Online version: http://www.ethnologue.com.

List, Johann-Mattis. 2012. *Sequence Comparison in Historical Linguistics*. PhD dissertation, University of Düsseldorf.

Needleman, Saul B. and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48: 443–453.

Nelder, John A. and Roger Mead. 1965. A simplex method for function minimization. *The computer journal* 7: 308–313.

Pompei, Simone, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PLoS ONE* 6: e20109.

Robinson, D. F. and Leslie R. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53: 131–147.

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

Somers, Harold L. 1998. Similarity metrics for aligning children's articulation data. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 2, Association for Computational Linguistics, 1227–1232.

Ward, Joe H., Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58: 236–244.

Wichmann, Søren, Eric W. Holman, Dik Bakker, and Cecil H. Brown. 2010. Evaluating linguistic distance measures. *Physica A: Statistical Mechanics and its Applications* 389: 3632–3639.

Wichmann, Søren, André Müller, Viveka Velupillai, Annkathrin Wett, Cecil H. Brown, Zarina Molochieva, Julia Bishoffberger, Eric W. Holman, Sebastian Sauppe, Pamela Brown, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Oleg Belyaev, Matthias Urban, Harald Hammarström, Agustina Carrizo, Robert Mailhammer, Helen Geyer, David Beck, Evgenia Korovina, Pattie Epps, Pilar Valenzuela, and Anthony Grant. 2012. The ASJP Database (version 15). http://email.eva.mpg.de/∼wichmann/ASJPHomePage.htm.

Wieling, Martijn, Eliza Margaretha, and John Nerbonne. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics* 40: 307–314.

Wieling, Martijn, Jelena Prokić, and John Nerbonne. 2009. Evaluating the pairwise string alignment of pronunciations. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, Association for Computational Linguistics, 26–34.

Appendix: ASJP transcription code

Tab. 15 and Tab. 16 contain the description of the ASJP code (quoted verbatim from Brown et al., 2013).

ASJP code symbol	*Description*	*IPA symbols*
p	voiceless bilabial stop and fricative	p,ɸ
b	voiced bilabial stop and fricative	b, β
f	voiceless labiodental fricative	f
v	voiced labiodental fricative	v
m	bilabial nasal	m
w	voiced bilabial-velar approximant	w
8	voiceless and voiced dental fricative	θ, ð
4	dental nasal	n̪
t	voiceless alveolar stop	t
d	voiced alveolar stop	d
s	voiceless alveolar fricative	s
z	voiced alveolar fricative	z
c	voiceless and voiced alveolar affricate	ts, ʣ
n	alveolar nasal	n
r	voiced apico-alveolar flap and all other varieties of "r-sounds"	ɾ, r, ʀ, ɻ
l	voiced alveolar lateral approximant	l
S	voiceless post-alveolar fricative	ʃ
Z	voiced post-alveolar fricative	ʒ
C	voiceless palato-alveolar affricate	ʧ
j	voiced palato-alveolar affricate	ʤ
T	voiceless and voiced palatal stop	c, ɟ
5	palatal nasal	ɲ
y	palatal approximant	j
k	voiceless velar stop	k
g	voiced velar stop	g
x	voiceless and voiced velar fricative	x, ɣ
N	velar nasal	ŋ
q	voiceless uvular stop	q
G	voiced uvular stop	ɢ
X	voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative	χ, ʁ, ħ, ʕ
h	voiceless and voiced glottal fricative	h, ɦ
7	voiceless glottal stop	ʔ
L	all other laterals	ʟ, ɭ, ʎ
!	all varieties of "click-sounds"	!, |, ‖, ǂ

Table 15: ASJP transcription code: consonants

ASJP code symbol	*Description*	*IPA symbols*
i	high front vowel, rounded and unrounded	i, ɪ, y, ʏ
e	mid front vowel, rounded and unrounded	e, ø
E	low front vowel, rounded and unrounded	æ, ɛ, œ, ɶ
3	high and mid central vowel, rounded and unrounded	ɨ, ɘ, ə, ɜ, ʉ, ɵ, ɞ
a	low central vowel, unrounded	a, ɐ
u	high back vowel, rounded and unrounded	ɯ, u
o	mid and low back vowel, rounded and unrounded	ɤ, ʌ, ɑ, o, ɔ, ɒ

Table 16: ASJP transcription code: vowels