Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

LINGUISTIC PHYLOGENETIC INFERENCE BY PAM-LIKE MATRICES

Antonella Delmestri and Nello Cristianini

November 2010

Technical Report # DISI-10-058

Linguistic Phylogenetic Inference by PAM-like Matrices

Antonella Delmestri[1] and Nello Cristianini[2]
1. Department of Information Engineering and Computer Science, University of Trento, Italy
2. Intelligent Systems Laboratory, University of Bristol, U.K.

Abstract

We apply to the task of linguistic phylogenetic inference a successful cognate identification learning model based on PAM-like matrices. We train our system and we employ the learned parameters for measuring the lexical distance between languages. We estimate phylogenetic trees using distance-based methods on an Indo-European database. Our results reproduce correctly all the established major language groups present in the dataset, are compatible with the Indo-European benchmark tree and include also some of the supported higher-level structures. We review and compare other studies reported in the literature with respect to recognised aspects of Indo-European history.

Keywords: phylogenetic inference, distance-based methods, PAM-like matrices.

1. Introduction

Languages which originate from a common ancestor are genetically related and there are two main methodologies used in historical linguistics to study their relationships: the comparative method [1] and the multilateral comparison method [28]. The *comparative method*, developed over the last two centuries, in order to detect connections among languages studies sound recurrent correspondences in *cognates*, which are words sharing the same etymological origin. The *multilateral comparison method*, developed by Greenberg, supported by Ruhlen [50] but very controversial, proposes language classification based on the number of surface similarities between groups of semantically connected words to determine the level of genetic relatedness between languages. Related to the comparative method, *lexicostatistic* [58,59] is a mathematical measure based approach that uses lists of core lexical data to determine language relationships on the basis of the percentage of shared cognates between languages. *Glottochronology* [21] is an extension of this method and aims to estimate divergence times under the assumption of a constant rate of language evolution. Surprisingly, also textual statistical analysis is able to identify relationships among languages by studying their shared statistical properties [60].
Because language evolution presents a close analogy with species evolution [12,3,56], its study may be successfully approached by techniques developed in evolutionary biology, as it happens more and more frequently. Language evolution may involve phonological, lexical and morphological changes, when sound, word or grammar changes are respectively considered. These changes are generally represented as a set of features or characters whose choice and coding are crucial for the outcome of the phylogenetic analysis. Common lexical characters used in studying language evolution are cognates which originate from a “vertical” transmission and do not include *borrowings* that are words loaned from other languages through a “horizontal” transmission.
Cognate words may be grouped and coded by the same character state or they may be used as states themselves when the characters considered are meaning lists. Cognate words have the advantage of being non-homoplastic by definition because they can exhibit the same character state only as a result of an evolutionary relationship and not as a result of parallel developments or back mutations. For this reason, the study of cognate words provides evidence of historical relationship between languages and may be used to identify genetic relationships between speech varieties and to infer phylogenies. The synergy between cognate identification and phylogenetic inference, both representing very promising applications of computational techniques to historical linguistics, may contribute to the tracing of language evolution and to the investigation of the origin of language.

2. Phylogenetic Inference

Phylogenies are evolutionary trees and phylogenetic inference aims to estimate the genetic relationships between taxa, which in principle may be species, languages or other entities. In linguistics a phylogenetic tree represents an estimation or hypothesis about the evolutionary relationships among groups of languages, based upon similarities and differences in their characters. The languages appear as leaves in the tree and are joined together when they are supposed to descend from a common ancestor. Internal nodes represent intermediate, not documented languages and tree branch lengths may signify language distances or divergence time accordingly with the methodology employed. Phylogenetic trees are generally binary and they are *unrooted*, when they only represent relationship between languages, or they are *rooted*, when they also identify a common ancestor. Any unrooted tree can be rooted on any of the internodes in the tree and it is compatible with all the rooted trees that can be built in this way. The number of unrooted binary trees for n languages is equal to $3*5*7*\ldots*(2n-5)$ and because for each unrooted tree there are $2n-3$ possible rooted trees the number of rooted trees is $3*5*7*\ldots*(2n-3)$ [23]. A common way to root an unrooted tree is to utilise an *outgroup* which serves as a reference for determination of the evolutionary relationship among the other nodes. It should be a language considered related to the other languages in the set, but less closely related to any language in the group than they are to each other.
Phylogenetic networks are rooted direct graphs that may be used when, together with evolutionary relationships, more complex interactions need to be represented which may include borrowing, creolization or language mixture. Due to space limitations, they are not covered in this paper.

2.1 Methods for Phylogenetic Inference

Methods for linguistics phylogenetic inference estimate the evolutionary history of languages using the information available about them. This information is generally coded in a matrix that may be a distance matrix or a character matrix. Depending on this, methods are classified as *distance-based methods* or *character-based methods* and most of the methods are guaranteed to reproduce the true evolutionary tree under certain conditions. When a method returns more than one tree having the same best score a *consensus tree* has to be calculated [23].
Some methods not only aim to infer phylogenetic tree topologies, but also to estimate the dating of language divergence times that depend on the original character data and on various assumptions. The scholars are divided as to whether or not the currently available statistical methodology for dating purposes may be accepted with any degree of confidence in historical linguistics [22,2,29,4].

2.1.1 Distance-based methods

Distance-based methods represent a major family of phylogenetic methods where the initial character matrix is used to statistically calculate a pairwise distance matrix which is then used to estimate a phylogenetic tree. It has been proved that the amount of information about the phylogeny that is lost in this process is remarkably small and that the estimates of the phylogenies produced by distance-based methods are quite accurate [23]. Distances may be considered estimates of the branch lengths separating pairs of languages, where different branches may have different rates of evolution.
A famous class of distance-based methods consists of *clustering algorithms* that apply an algorithm to a distance matrix in order to produce a phylogenetic tree. These methods are very fast and, under certain assumptions, they are guaranteed to perform well. However, their statistical properties are not clear because they do not optimise an explicit criterion. Two standard clustering algorithms extensively used in computational biology are UPGMA [55] and Neighbour-Joining [52].

2.1.1.1 UPGMA

UPGMA (Unweighted Pair-Group Method with Arithmetic mean) [55] is guaranteed to perform well under the molecular clock hypothesis, which implies that the input distances represent languages that have evolved with a constant rate of evolution, following the glottochronology approach. This is a reasonable assumption only if the entities are closely related. At each step the algorithm combines together the nearest two clusters into a new cluster and calculates the distance between the new cluster and the others as the mean distance between the elements of each cluster. The computational cost is

$O(n^2)$, where n is the distance matrix dimension. Clocklike trees are rooted and have an equal total branch length from the root to any leaf [23].

2.1.1.2 Neighbor-Joining

Neighbor-Joining [52,57] is a clustering algorithm that does not assume a molecular clock and is guaranteed to reconstruct phylogenetic trees perfectly when the pairwise distances are the exact reflection of a tree. It assumes the minimum evolution criterion for phylogenetic trees and, at each iteration, it chooses the topology that minimises the total branch length. It produces an unrooted tree that can be rooted by using an outgroup. The computational cost is $O(n^3)$, where n is the distance matrix dimension.

2.1.2 Character-based methods

A language may be described by a vector of character states and a group of languages may be represented by a matrix, where each row symbolises a language and each column signifies a character. Character-based methods use this character matrix to evaluate a phylogeny.

2.1.2.1 Maximum Parsimony

Maximum Parsimony (MP) [23] is a non-parametric statistical method whose target is to find an unrooted tree that requires the minimum number of evolutionary changes to describe the observed data and may find several trees with the same best score. MP does not guarantee to produce the true tree because of the "long branch attraction" that occurs when the rates of evolution are very different on different branches of the true tree. In this case MP considers lineages that evolve rapidly as to be closely related, regardless of their true evolutionary relationships. MP may be weighted, when different weights are assigned to different characters, or unweighted. Finding a MP tree is an NP-complete problem [25] and for this reason MP analyses are frequently performed using heuristics. These generally may find only local optima, rather then global optima, and anyway be very time consuming.

2.1.2.2 Maximum Compatibility

Maximum Compatibility (MC) [23] is a non-parametric method which aims at finding an unrooted tree that presents the maximum number of compatible characters to illustrate the observed data, where being compatible here means evolving without any homoplasy, i.e. without back mutation or parallel evolution. When a tree has all the characters compatible it is called a *perfect phylogeny*. MC may be weighted or unweighted and may find several trees with the same best score. The problem is NP-complete [9] and there are no heuristics available that are highly accurate. However, if the maximum number of states per character is bounded then it is possible to find a solution in $O(2^{2r}\, n\, k^2)$ where r is the maximum number of states per character, n is the number of leaves, and k is the number of characters [35].

2.1.2.3 Maximum Likelihood

Maximum Likelihood (ML) methods [23] are based on explicit parametric models of character evolution and they aim to estimate the tree and the parameters that maximize the likelihood of the observed data under the chosen evolutionary model. ML is statistically consistent and generally produces very good estimates of the tree phylogeny, but it is NP-hard [11].

2.1.2.4 Bayesian Inference

Bayesian methods [23] are also based on explicit parametric models of character evolution. Their objective is to estimate a consensus tree, or sometimes the maximum posterior probability tree, of a posterior probability distribution on the space of the model trees, calculated from an initial tree and the observed data. Bayesian methods generally produce very good estimates of the phylogeny, but their computational time is extremely expensive. Markov Chain Monte Carlo (MCMC) algorithms [30] are frequently used to calculate an approximate posterior distribution of the trees instead. Initial priors may allow the inclusion in the evolutionary analysis of evidence available from other fields, like genetics, anthropology and archaeology. However, the results should be examined considering

both their sensitivity to the priors used and the reliability of the MCMC approximation of the trees probabilities [33].

2.2 Evaluation of Phylogenetic Inference

The evaluation of phylogenetic estimations is very difficult because the true evolutionary history is not generally fully known even for the best understood language families. The choice of both phylogenetic inferring methodology and data significantly impact the phylogenetic estimation. The following criteria, proposed by Nichols and Warnow [45], should be a necessary and crucial requirement of any phylogenetic estimation when using data from a well-known language family.

The *"Compatible resolution"* criterion requires that the inferred tree is compatible with the benchmark tree, meaning that the established subgroups should not be mixed, but they may be not completely resolved. That may happen when the data are not sufficient to provide a complete resolution or when a consensus tree is used.

The *"No missing subgroups"* criterion requires that the evaluated tree includes all the established subgroups and it is strictly stronger than the first criteria, and for this reason is considered desirable, but not essential.

The "*Calibration*" criterion is essential for models that estimate dating. It requires that a method is tested on one or more datasets and if the inferred dates are not close enough to the established dates, the method has to be calibrated on the known dates.

3. State of the Art

The last decade has seen a large number of studies developing and employing phylogenetic techniques to investigate the evolution of language. We shall review some of the more interesting results especially regarding the Indo-European language family, which is the most intensively studied, but we shall mention studies involving other language families as well. Investigations comparing different methods include Nakhleh et al. [41], Barbançon et al. [7], Wichmann and Saunders [61].

3.1 Distance-based methods

Dyen et al. [19] collected an Indo-European dataset described in Section 4.1, made a lexicostatistical classification of the 84 languages included and calculated the percentage of cognates shared by each language pair creating an 84-by-84 distance matrix. They developed a non-standard clustering algorithm belonging to the family of pair-group methods, like UPGMA, adapted to deal properly with lexicostatistical percentages. The tree proposed was not completely compatible with the benchmark tree of the Indo-European language family, even if it reproduced all the established major Indo-European branches with the exclusion of the Indo-Iranian clade.

Ellison and Kirby [20], calculated a word similarity within each of the 95 languages extracted from the Indo-European dataset by Dyen et al. [19]. They called it lexical metric and they defined it as a distribution of confusion probabilities, based on the Levenshtein distance [38] normalised by the average length of the words. The divergence between two languages was defined as the divergence of their lexical metrics and calculated as the geometric path between the two distributions, creating a distance matrix. They used Neighbor-Joining to build a phylogenetic tree and rooted it with a random outgroup. This tree was not compatible with the benchmark tree of the Indo-European language family, even if it showed a correct subgrouping for many languages.

Serva and Petroni [53] applied the Levenshtein distance [38] normalised by the length of the longer word, to 50 language pairs extracted from the Indo-European dataset by Dyen et al. [19]. For each language pair they compared 200 word pairs with the same meaning and they computed the average of these edit distances in order to create a 50-by-50 matrix of language distances. Serva and Petroni transformed this distance matrix into a time distance matrix following the glottochronology approach and imposed some well-known priors to the system in the aim of providing a phylogenetic tree topology with absolute time scales. Finally, they inferred a rooted phylogenetic tree using UPGMA [55]. The proposed tree topology satisfied the "No missing subgroups" criterion but violated some compatibility requirements for phylogenetic estimation [45]. The same methodology was applied by *Petroni and Serva* [46] to the Austronesian language family and was expanded by *Blanchard et al.* [8]

to represent geometrically the relationships between languages belonging to both the Indo-European and the Austronesian language families.

Brown et al. [10] developed ASJP (Automated Similarity Judgment Program) aiming to perform a large-scale classification of languages by calculating their lexical similarity following a lexicostatistical approach. They used Swadesh 100-word lists [59] from 245 globally distributed languages with the objective to expand their database to all the world's languages. They used the Neighbor-Joining [52] algorithm to generate the phylogenetic trees. The list dimension was subsequently reduced to 40 more stable lexical elements for the achievement of better results, and the database was expanded to 900 languages [32]. The algorithm used to determine whether or not words were likely to be cognate was changed by *Bakker et al.* [6]. They employed the Levenshtein distance [38], as proposed by Serva and Petroni [53] but with a double normalisation, first dividing it by the length of the longer word and then dividing this quantity by the averaged normalised Levenshtein distance among the words with different meaning. ASJP presented non uniform performance, passing both evaluation tests for some language families and failing both for others [45].

Downey et al. [18] estimated phylogenies for the Sumbanese language family using ALINE [37] to produce a distance matrix that was then processed by distance-based methods. ALINE is an algorithm developed by Kondrak for sequence alignment that evaluates the phonetic similarity between word pairs using incorporated linguistic knowledge. Downey and co-workers, in order to control the bias due to different string length, normalised the aligner score by the arithmetic mean of the rate given by ALINE applied to align each string with itself. Downey et al. utilised both UPGMA [55] and Neighbor-Joining [52] to estimate phylogenetic trees. The proposed phylogenetic trees were close to the historical reconstruction, especially the phylogeny built by UPGMA [55], which satisfied the "No missing subgroups" criterion, but violated some compatibility requirements for phylogenetic estimation.

3.2 Maximum Parsimony

Gray and Jordan [27] made one of the first attempts to apply biological phylogenetic methods to historical linguistics. They encoded the presence or absence of 5,185 lexical characters of cognacy for 77 Austronesian languages in a binary matrix and they employed a maximum parsimony analysis that produced a single most-parsimonious tree. The topology of this tree supported the express-train model of Austronesian expansion [16] and showed considerable agreement with traditional linguistic groupings, even if the tree violated the compatible resolution criterion [45].

Rexová et al. [48] used Maximum Parsimony and greedy consensus trees on the comparative Indo-European database by Dyen et al. [19] focussing on the impact of the character encoding. They employed three different methods of character encoding creating a standard multi-state matrix, an altered multi-state matrix and a binary matrix. The study showed substantial dissimilarities between the two multi-state matrices and the binary matrix, including different tree rooting, suggesting that the binary encoded data matrix produced less reliable trees than those created employing the multi-state matrices.

3.3 Maximum Compatibility

Ringe et al. [49] used Maximum Compatibility to estimate the phylogenetic tree of the Indo-European language family. They utilised lexical, morphological and phonological characters from 24 Indo-European languages and the Kannan and Warnow algorithm [35] which runs in polynomial time. They assigned weights to characters, which made the model very dependent on the linguistics choice. The Ringe et al. method rooted the tree by hand, after examination of the unrooted tree produced by MC and passed the two minimal evaluation criteria of phylogenetic inference [41].

3.4 Bayesian Analysis

Gray and Atkinson [26] estimated the language-tree divergence times for the Indo-European language family suggesting a root age of Indo-European of between 7,800 and 9,800 BP, consistent with the Anatolian theory of Indo-European origin. In order to help the evaluation of older language relationships, they added to the Indo-European dataset by Dyen et al. [19] three extinct Indo-European languages, Hittite, Tocharian A and Tocharian B, reaching a total of 87 languages. Based on cognate

judgments from this extended corpus, they produced a binary matrix of 2,449 lexical characters indicating the presence or absence of words in each cognate group. This binary matrix was then examined using maximum-likelihood models, Bayesian MCMC analysis and rate-smoothing algorithms to produce a majority-rule consensus tree. This model allowed homoplasy and supported polymorphism. The proposed tree topology satisfied the two criteria required by Nichols and Warnow [45] for phylogenetic estimation, while the dating failed the calibration criterion. The Gray and Atkinson method was subsequently extended [5,2,4] and applied also to study the Bantu language family [47,31].

Nicholls and Gray [44,43] applied to language evolution a stochastic model first introduced by Huson and Steel [34], and dated the Indo-European language family at about 8,000-9,000 BP. The model implemented Dollo parsimony principles, used Bayesian phylogenetic inference and MCMC algorithms [30] to generate a sample distribution of trees, and screened them using constraints before producing a consensus tree. Nicholls and Gray encoded the multi-state lexical characters from the extended Dyen et al. corpus [19,26] and from the Ringe et al. dataset [49]. They ran several analyses considering 6 subsets of the first dataset and 3 subsets of the second. They found that age estimations of the root were uniform across all the analysis whereas the topologies were not reliable. This model did not allow homoplasy and supported polymorphism. This model could not handle missing data and so the analysis had to be limited only to those characters shown in all speech varieties, dropping some languages which presented too much missing data.

Ryder and Nicholls [51] extended the Nicholls and Gray [43] method to handle missing data. They used binary encoding of cognate classes as lexical traits from the Ringe et al. dataset [49] and they gave also an analysis of the extended Dyen et al. corpus [19,26] in the paper supplement. They estimated the date of the Proto-Indo-European language around 7,100-9,800 BP.

4. The Indo-European Language Family

The Indo‑European language family is one of the most intensively studied and it is significant to the field of historical linguistics as it possesses one of the longest recorded histories. However, its origin still represents one of the most recalcitrant problems of historical linguistics [17] and its higher-order subgrouping remains controversial. All languages are supposed to be descendants of a common ancestor, the Proto-Indo-European, and the basic subgroups are very well established. They include the extinct Anatolian and Tocharian and the contemporary Albanian, Armenian, Celtic, Germanic, Greek, Italic, Baltic and Slavic, grouped together into a Balto-Slavic clade, Indo-Aryan and Iranian, grouped together to form an Indo-Iranian clade. Even if the higher structure does not find agreement, the initial split into Anatolian versus all the others is linguistically well sustained. Moreover, some phylogenies have more support than others, including a radial phylogeny, one where Celtic departs very early, one that groups Balto-Slavic and Indo-Iranian together, or Armenian with Greek or Celtic with Italic [45].

4.1 Indo-European Datasets

The corpora prepared by Dyen et al. [19] and by Ringe et al. [49] for the Indo-European language family are recommended datasets for linguistics studies [45]. They differ in many aspects including the number of languages considered, their dating and the types of characters reported.

The *Comparative Indo-European Database by Dyen et al.* [19] provides lexical data in the form of 200-word Swadesh lists [58] from 84 contemporary Indo-European speech varieties. In it, each word is presented in orthographic format without diacritics, using the 26 letters of the Roman alphabet. The data are grouped by meaning and cognateness, which is reported as certain or doubtful. It is considered the less accurate of the two and may benefit from a revision. The dataset digital version covers 95 languages, of which only 84 were considered accurate enough to be included in the monograph.

The *Dataset of Ringe et al.* [49] is provided in two versions, unscreened and screened, both containing phonological, morphological and lexical characters for 20 extinct and 4 existing Indo-European languages, chosen as the oldest well attested languages in each branch of the Indo-European family. The screened dataset is produced from the unscreened version by removing all characters that clearly

exhibited homoplasy. The data are provided in character matrices, each corresponding to a character type.

5. Methodology

We have developed a model to estimate phylogenies based on a learning system introduced in [14,15] for cognate identification which outperformed previously presented proposals. We have focussed on the tree topology and we have avoided modelling a dating scheme. The learning system, described in Section 5.1, trains PAM-like substitution matrices from a sensibly aligned dataset and uses them together with a family of string similarity measures to align word pairs in order to calculate their similarity scores. We have utilised these similarity scores between word pairs to calculate similarity scores between language pairs. We have employed Swadesh lists of words with the same meaning in different languages and, for each language pair, we have calculated and averaged the word pair similarity rates, producing a similarity matrix for each PAM-like matrix and similarity measure employed. We have then converted these similarity matrices into distance matrices as explained in Section 5.2 and we have computed a weighted average of them to reach a consensus [40,23]. We have then employed distance-based methods to estimate phylogenetic trees.

5.1 The Learning System

The learning system proposed by Delmestri and Cristianini [14,15] is inspired by biological sequence analysis and consists of three main modules.

The first component is a global pairwise aligner [42] that aligns sensibly cognate pairs and prepares a meaningful training dataset, guided by a linguistic-inspired substitution matrix [15]. This 26-by-26 matrix aims to represent the a-priori likelihood of transformation between each character of the Roman alphabet into another and tries to code well known systematic sound changes left in written Indo-European languages.

The second component of the learning system is a generator of PAM-like substitution matrices that uses a technique similar to the PAM method developed by Margaret Dayhoff et al. [13] and widely used for amino acid sequence analysis. The PAM approach aims to learn substitution parameters from global alignments between closely related sequences and then to extrapolate from these data longer evolutionary divergences, assuming a constant rate of evolution.

The third component of this system is a pairwise aligner that benefits from the PAM-like matrices and from a family of parameterised string similarity measures [14] to rate the word pairs. The similarity measures derive from different normalisations of a generic scoring algorithm and take into account the similarity of each string with itself with the aim of eliminating, or at least reducing, the bias due to different string length. Possible scoring algorithms include the Needleman-Wunsch algorithm [42] for global alignment and the Smith-Waterman [54] for local alignment.

5.2 From Similarities to Distances

In order to convert our similarity matrices into distance matrices, we have experimented with three different methods in the aim of studying possible differences in the resulting phylogenetic trees. Because the similarity measures employed are defined through different normalisations of a generic similarity rating algorithm score [15], the similarity scores fall in the range [1..0], where 1 means maximum similarity and 0 means no similarity. Given a generic similarity matrix S, in the first case we have taken the more obvious approach subtracting S to the identity matrix so that the distance rates drop into the range [0..1]. In the second case we have calculated the negative natural logarithm of S, which ranges from 0 to ∞. In the third case we have first divided the identity matrix by S and then subtracted the identity matrix from it, which produces rates always equal or greater than zero. These last two methods were investigated by Feng and Doolittle [24] for measuring evolutionary times.

$$D_1 = 1 - S$$
$$D_2 = -\ln S$$
$$D_3 = (1/S) - 1$$

6. Dataset

We have employed the Comparative Indo-European Database by Dyen et al. [19] considering the 84 languages documented in the monograph. In the absence of two large homogeneous linguistics datasets to be used as training and test dataset without intersection, we have split the Dyen et al. corpus [19] into two groups of meanings identified by odd and even ordinal numbers. Firstly, we have created a training dataset from the odd meanings and prepared a test dataset from the even meanings, called test_even. Secondly we have done the opposite, using the even meanings as training dataset and the odd meanings as test dataset, named test_odd.
For the training datasets we have used only the word pairs reported by Dyen et al. [19] as certain cognates with each other and we have included only the first cognate pair, if more words were provided for the same meaning in the same language. We have then aligned the two training groups of word pairs using a linguistic-inspired matrix [14] obtaining two separate training datasets, called respectively training_odd and training_even.
For the test datasets we have considered all the word pairs reported by Dyen et al. [19] as certain or uncertain cognates, but we have excluded those words classified as not acceptable or not cognate with any other, as they may include borrowings which we wanted to discard from our analysis. We have corrected a few evident errors.
In order to root our Indo-European phylogenetic tree, we have included the Turkish language as outgroup because it belongs to the Altaic language family [39] which is not too distantly related to the Indo-European language family. We have utilised the Turkish list provided by Kessler [36] and we have added the 9 words in which this list differs from the 200-word Swadesh list by Dyen et al. [19] checking multiple sources to ensure reliability. We have extended test_odd and test_even respectively with the odd and even meanings from this Turkish list reaching a total of 85 languages. Having two training datasets and two test datasets has avoided any data overlap, thereby ensuring that independent analyses have been conducted and their results subsequently averaged, as explained in Section 7.
We would have liked to include also Hittite, Tocharian A and Tocharian B provided by Gray and Atkinson [26], but studying the data we found them inappropriate for our analysis. On several occasions the same meaning for the same language has been classified more than once with different Cognate Class Numbers (CCN), which is not the case for the rest of the original dataset by Dyen et al. [19]. This would have biased the learning procedure towards Hittite, Tocharian A and Tocharian B that would have given more contributions to the PAM-like matrices than the other languages.

7. Experiments

We have designed our experiments aiming to evaluate a phylogenetic tree which may reflect lexical similarity between languages. PAM-like substitution matrices can be seen as an indicator of the relative evolutionary interval since the languages diverged. Given that languages evolve at broadly changing rates, there is no simple connection between PAM distance and evolutionary time. However, for an analysis of a specific language family across multiple speech varieties, the corresponding PAM-like matrices will provide a relative evolutionary distance between the languages and allow accurate phylogenetic inference.
We have employed the two training datasets training_odd and training_even to learn two families of PAM-like matrices based on the Roman alphabet extended with gap, as it has been proved in [15,14] that learning gap penalties increases the effectiveness of the system. We have called these two matrix families respectively DAY_84b_odd and DAY_84b_even. In order to choose which matrices and which similarity measures to use to reach the best phylogenetic estimation, we have tested the performance of these matrix families in the task of cognate identification on the English, German, Latin, French and Albanian lists provided by Kessler [36]. PAM3, PAM4, PAM5 and PAM6 from both the families DAY_84b_odd and DAY_84b_even have reached the best accuracy when used with sim_1, sim_3, sim_5 and sim_6 [15] based on the Smith-Waterman (SW) algorithm [54]. We have used these PAM-like matrices and these similarity measures based on SW to calculate the language similarity between each of the 85 speech varieties in test_even and test_odd respectively, as the average similarity between the word pairs belonging to the language pair and having the same

meaning. We have supported polymorphism and if one or both languages presented more than one word for a meaning, the maximum similarity between the different pairs has been considered in the average calculation. We have obtained two 85-by-85 similarity matrices and we have calculated their average scores reaching a single 85-by-85 similarity matrix for each PAM-like pair (odd, even) and for each similarity measure employed. Finally, we have transformed these similarity matrices into distance matrices as described in Section 5.2 and we have calculated their weighted average to reach a consensus [40,23]. We have then applied to this distance matrix UPGMA [55] and Neighbor-Joining [52] to estimate phylogenies.

8. Results

Figure 1 shows the 85-by-85 pairwise distance matrices produced using D_1, D_2 and D_3 presented in Section 5.2 with the outgroup in first position. This visual representation highlights the subsets of sequences that are more closely related to each other, represented by the darker tones in the central clusters. All three matrices display the major Indo-European branches following the order of the Dyen et al. dataset classification [19] from top-left to bottom-right: Celtic, Italic, Germanic, Balto-Slavic, Indo-Aryan, Greek, Armenian, Iranian and Albanian. The first matrix presents a clearer distinction between the central clusters and the rest of the data.

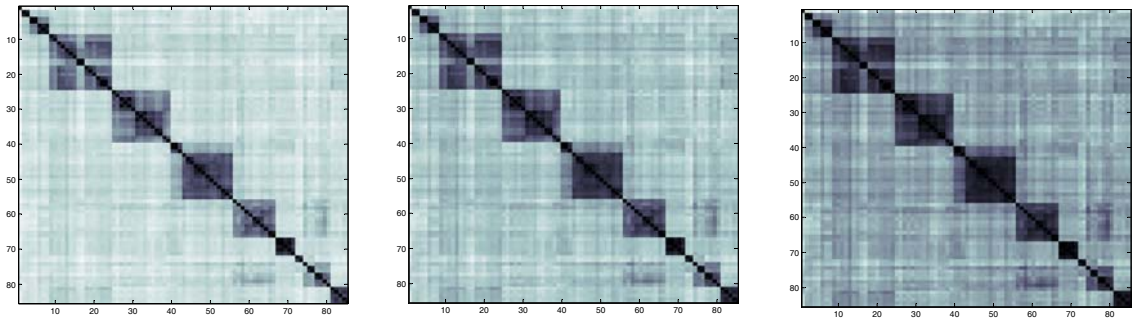

Figure 1 - Visual representation of the distance matrices

Figure 2 displays the topology of the consensus tree reached using UPGMA [55] with D_1, as defined in Section 5.2, that shows an identical canonical form to the tree built with D_2. The tree evaluated using D_3 gives only a slight variation within two subgroupings (Albanian and Iranian) and it is not reported. The algorithm has produced a tree rooted on the Turkish language, that is the outgroup we have added to the Dyen at al. dataset [19]. The confidence of the consensus tree is 100% for 77% of the branches and the uncertainty derives only from the internal Albanian and Iranian subgroupings. The tree estimated is compatible with the benchmark tree as documented in Ethnologue [39] and reproduces all the established major Indo-European groups present in the dataset. The position of the French Creole speech varieties, which are not even considered as Indo-European languages, is justified by the nature of creolization which requires network models of evolution [45]. The tree topology shows also some of the higher-level supported structures such as Celtic departing early and Balto-Slavic grouping with Indo-Iranian. Italic groups with Celtic, but after forming a clade with Albanian. UPGMA has worked particularly well here because the PAM-like matrices utilised to calculate the language similarities assume a constant rate of evolution. This is the prerequisite for UPGMA to infer phylogenies accurately [23] and in this case it is also a reasonable assumption because the languages considered belong to the same family and are quite closely related.

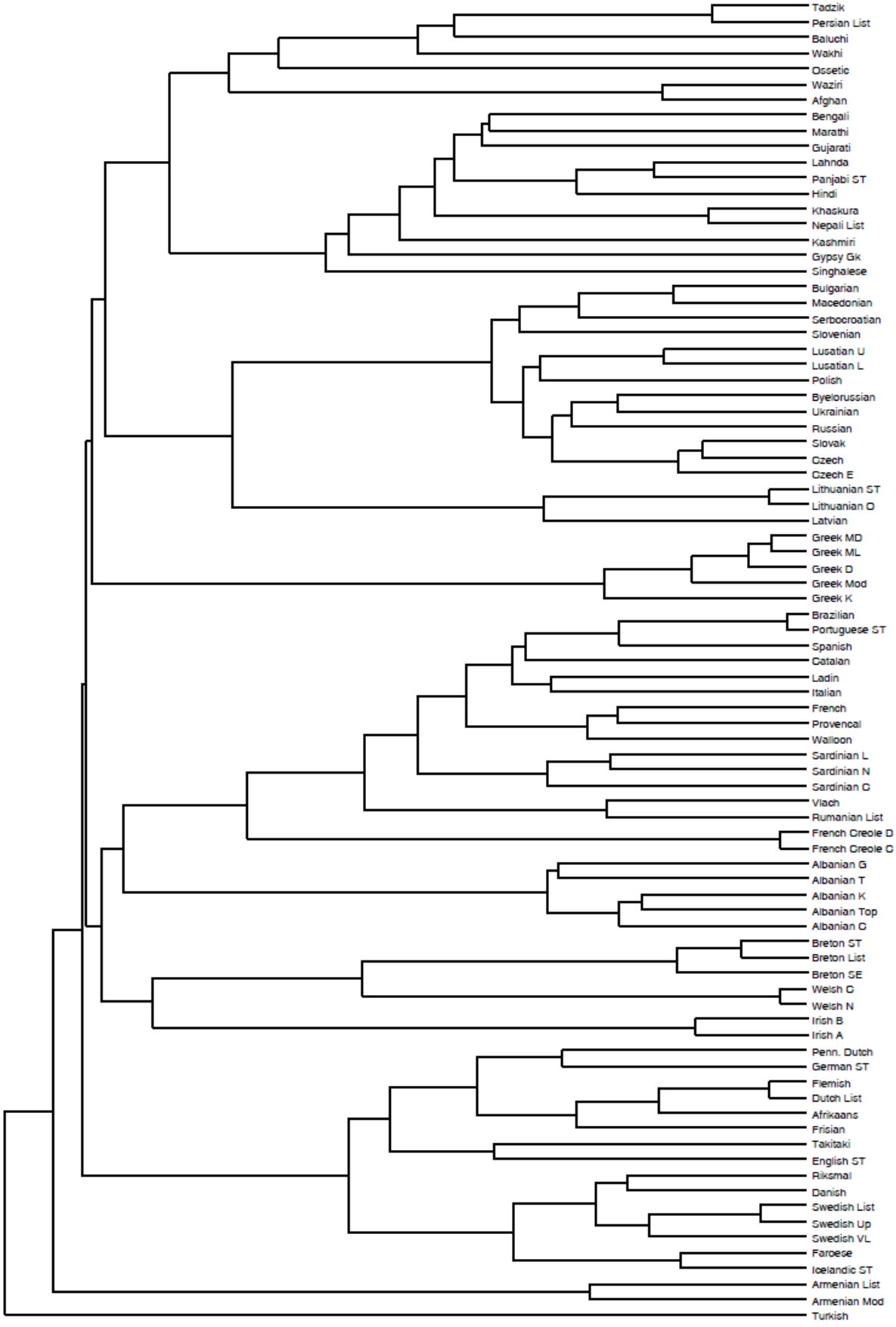

Figure 2 – Indo-European phylogenetic tree produced by UPGMA

The topologies of the unrooted consensus trees reached using Neighbor-Joining [52] with the distances calculated by D_1, D_2 and D_3 defined in Section 5.2 show different canonical forms. The three trees estimated reproduce all the established major Indo-European groups present in the dataset, but with some differences in the subgroupings. For example, the trees evaluated using D_2 and D_3 show accurately the French Creole speech varieties joined to the Gallo-Romance branch, but fail in grouping correctly the East Slavic branch.

Figure 3 shows the unrooted consensus tree calculated by D_1 that has proved to be the more accurate. The confidence of this consensus tree is 100% for 55% of the branches and the uncertainty spreads across the tree with the exclusion of the Armanian, Greek, Italic and Baltic groups.

The estimated tree is compatible with the benchmark tree as documented in Ethnologue [39] and reproduces all the established major Indo-European groups present in the dataset. The position of the French Creole speech varieties is not precise as in the case of UPGMA, however these languages are not even classified as Indo-European because of their creolization [39].

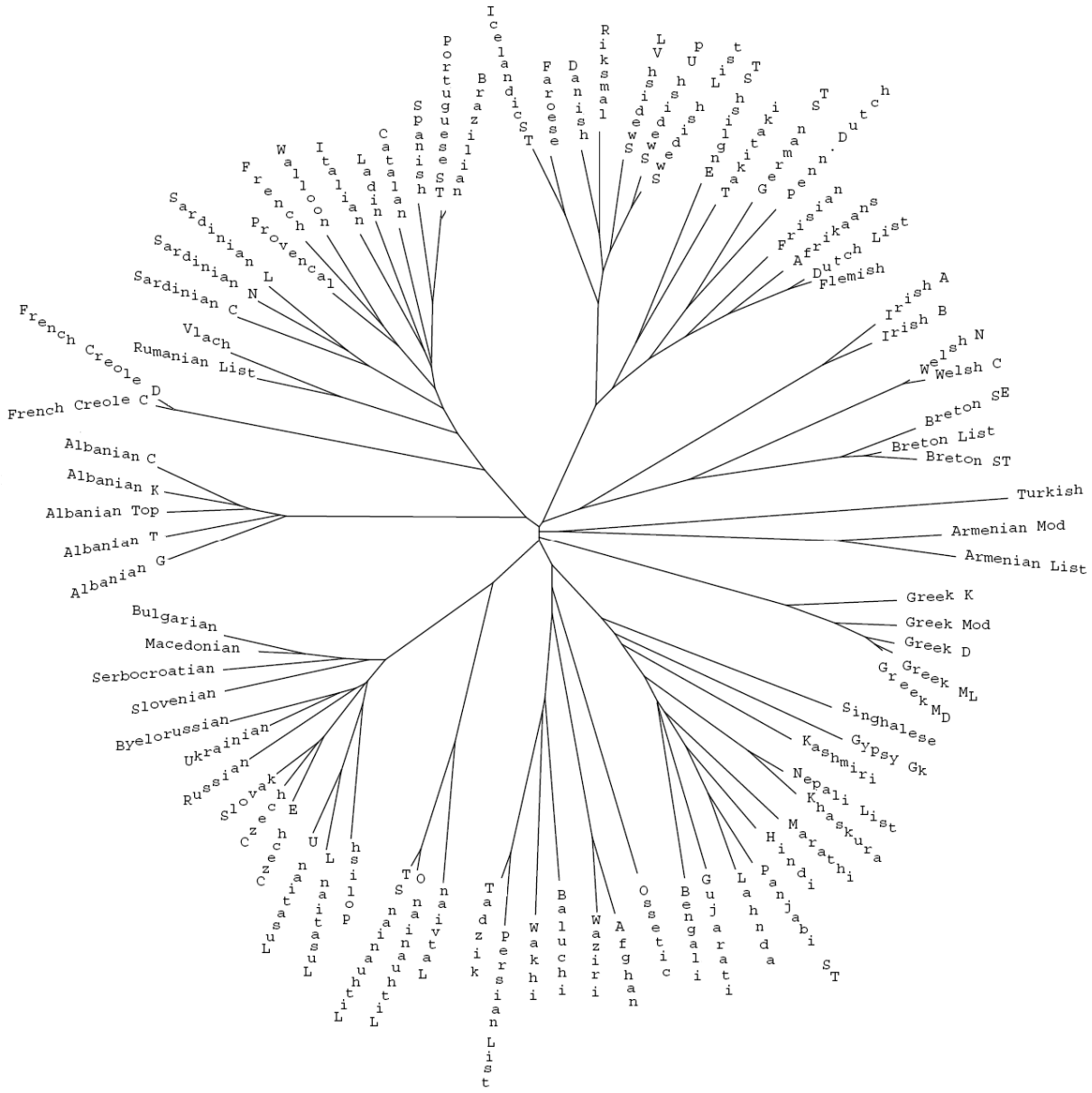

Figure 3 - Unrooted Indo-European phylogenetic tree produced by Neighbor-Joining

8.1 Related works

Serva and Petroni [53,46], *Blanchard et al.* [8], *Bakker et al.* [6] and *Downey et al.* [18] used distance-based methods to infer phylogenies as reported in Section 3.1. In all these cases language distance was calculated by averaging the distance of word pairs having the same meaning in compared languages. In order to compute word distance, the first three scholar groups utilised the Levenshtein distance [38] choosing different normalisation, and the fourth group employed ALINE [37], normalised as well. *Serva and Petroni* [53,46] together with *Blanchard et al.* [8] considered only 50 languages from the Dyen at al. dataset [19] reducing enormously the complexity of the phylogeny. *Bakker et al.* [6] developed their own dataset and *Downey et al.* [18] applied their method to the Sumbanese language family. Because of these differences, a specific comparison of our results with theirs is not possible. However, it has been shown in [14] that our cognate identification system produces an average accuracy approximately 28% higher that the Levenshtein distance normalised by the length of the longer word, and 18% higher than ALINE, as reported in the literature. This would suggest that our methodology may infer phylogenies more accurately than the methods reported.

9. Conclusion

We have applied to the task of phylogenetic inference a cognate identification system that had already shown an outstanding performance in previous studies. The capacity of our methodology to detect word and language similarity has proved to be successful, allowing the inference of phylogenetic trees compatible with the Indo-European benchmark tree and the reproduction of all the established major groups present in the dataset. This would suggest that our method satisfies all the crucial evaluation criteria of phylogenetic estimation. This result is very promising considering that our model uses lexical characters in the form of words and that the database used is known not to be completely accurate. This would imply that our system is quite resistant to data noise and that it may even improve its performance further, when cleaner data were available.

Our future objectives include the application of the methodology presented in this study to phylogenetic inference using the Indo-European dataset by Ringe et al. [49], whose accuracy is well documented. Another step forward would be the development of BLOSUM-like matrices to be employed in both the fields of cognate identification and phylogenetic inference.

10. Acknowledgments

We thank Quentin Atkinson for supplying and commenting the Hittite and Tocharian lists and Geoff Nicholls for providing some of his papers and datasets.

11. References

[1] R. Anttila, *An Introduction to Historical and Comparative Linguistics*. New York, U.S.A.: Macmillan Publishing Co., Inc., 1972.

[2] Q. D. Atkinson, R. D. Gray, "Are accurate dates an intractable problem for historical linguistics?," in *Mapping our Ancestry: Phylogenetic Methods in Anthropology and Prehistory*, C. Lipo et al., Eds. Chicago, Illinois: Aldine, 2006, pp. 269–96.

[3] Q. D. Atkinson, R. D. Gray, "Curious Parallels and Curious Connections - Phylogenetic Thinking in Biology and Historical Linguistics," *Systematic Biology*, vol. 54, no. 4, pp. 513-526, 2005.

[4] Q. D. Atkinson, R. D. Gray, "How old is the Indo-European language family? Illumination or more moths to the flame?," in *Phylogenetic methods and the prehistory of languages*, P. Forster and C. Renfrew, Eds. Cambridge, U.K.: MacDonald Institute Press, University of Cambridge, 2006, ch. 8, pp. 91–109.

[5] Q. D. Atkinson, G. Nicholls, D. Welch, R. D. Gray, "From words to dates: water into wine, mathemagic or phylogenetic inference?," *Transactions of the Philological Society*, vol. 103, no. 2, pp. 193–219, 2005.

[6] D. Bakker et al., "Adding typology to lexicostatistics: a combined approach to language classification," *Linguistic Typology*, vol. 13, pp. 167-179, 2009.

[7] F. Barbançon, T. Warnow, S. N. Evans, "An experimental study comparing linguistic phylogenetic reconstruction," in *Proceedings of the Conference on Language and Genes, University of California*, Santa Barbara, September 2006.

[8] P. Blanchard, F. Petroni, M. Serva, D. Volchenkov, "Geometric representations of language taxonomies," *Computer Speech and Language (In press)*, 2010.

[9] H. L. Bodlaender, M. R. Fellows, T. J. Warnow, "Two strikes against perfect phylogeny," in *Automata, Languages and Programming. Lecture Notes in Computer Science*, W. Kuich, Ed. Berlin, Germany: Springer Verlag, 1992, vol. 623, pp. 273-283.

[10] C. H. Brown, E. W. Holman, S. Wichmann, V. Vilupillai, "Automated classification of the World's languages: A description of the method and preliminary results," *STUF – Language Typology and Universals*, vol. 61, no. 4, pp. 285-308, 2008.

[11] B. Chor, T. Tuller, "Finding a maximum likelihood tree is hard," *Journal of the ACM (JACM)*, vol. 53, no. 5, pp. 722-744, 2006.

[12] C. R. Darwin, *The descent of man, and selection in relation to sex*. London: John Murray, 1871.

[13] M. O. Dayhoff, R. V. Eck, C. M. Park, "A Model of Evolutionary Change in Proteins," *Atlas of Protein Sequence and Structure*, vol. 5, pp. 89-99, 1972.

[14] A. Delmestri, N. Cristianini, "Robustness and Statistical Significance of PAM-like Matrices for Cognate Identification," *Journal of Communication and Computer (In press)*, vol. 7, no. 12, 2010.

[15] A. Delmestri, N. Cristianini, "String Similarity Measures and PAM-like Matrices for Cognate Identification," *Bucharest Working Papers in Linguistics (In press)*, vol. XII, no. 2, 2010.

[16] J. M. Diamond, "Express train to Polynesia," *Nature*, vol. 336, no. 6197, pp. 307-308, 1988.

[17] J. M. Diamond, P. Bellwood, "Farmers and Their Languages: The First Expansions," *Science*, vol. 300, no. 5619, pp. 597-603, April 2003.

[18] S. S. Downey, B. Hallmark, M. P. Cox, P. Norquest, S. J. Lansing, "Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction," *Journal of Quantitative Linguistics*, vol. 15, no. 4, pp. 340–369, 2008.

[19] I. Dyen, J. B. Kruskal, P. Black, "An Indoeuropean classification: A lexicostatistical experiment," *Transactions of the American Philosophical Society*, vol. 82, no. 5, 1992.

[20] M. T. Ellison, S. Kirby, "Measuring language divergence by intra-lexical comparison," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Sydney, Australia, 2006, pp. 273-280.

[21] S. M. Embleton, *Statistics in Historical Linguistics, Series: Quantitative linguistics, vol. 30*. Bochum: Studienverlag Brockmeyer, 1986.

[22] S. N. Evans, D. A. Ringe, T. Warnow, "Inference of divergence times as a statistical inverse problem," in *Phylogenetic Methods and the Prehistory of Languages*, P. Forster and C. Renfrew, Eds. Cambridge, U.K.: McDonald Institute for Archaeological Research, July 2004.

[23] J. Felsenstein, *Inferring Phylogenies*. Sunderland, Massachusetts, U.S.A.: Sinauer Associates Inc. Publishers, 2004.

[24] D-F. Feng, R. F. Doolittle, "Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships," *Journal of Molecular Evolution*, vol. 44, no. 4, pp. 361-370, April 1997.

[25] L. R. Foulds, R. L. Graham, "The steiner problem in phylogeny is NP-complete," *Advances in Applied Mathematics*, vol. 3, no. 1, pp. 43-49, March 1982.

[26] R. D. Gray, Q. D. Atkinson, "Language-tree divergence times support the Anatolian theory of Indo-European origin," *Nature*, vol. 426, pp. 435-439, 27 November 2003.

[27] R. D. Gray, F. M. Jordan, "Language trees support the express-train sequence of Austronesian expansion," *Nature*, vol. 405, pp. 1052-1055, June 2000.

[28] J. H. Greenberg, *Essays in Linguistics*. Chicago, Illinois: University of Chicago Press, 1957.

[29] S. J. Greenhill, Q. D. Atkinson, A. Meade, R. D. Gray, "The shape and tempo of language evolution," *Proceedings of the Royal Society, B: Biological Science (In press)*, 2010.

[30] W. K. Hastings, "Monte Carlo sampling methods using Markov Chains and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.

[31] C. J. Holden, R. D. Gray, "Rapid radiation, borrowing and dialect continua in the Bantu languages," in *Phylogenetic methods and the prehistory of languages*, P. Forster and C. Renfrew, Eds. Cambridge, U.K.: McDonald Institute for Archaeological Research, 2006, pp. 19-31.

[32] E. W. Holman et al., "Explorations in automated language classification," *Folia Linguistica*, vol. 42, no. 2, pp. 331-354, 2008.

[33] J. P. Huelsenbeck, B. Larget, R. E. Miller, F. Ronquist, "Potential Applications and Pitfalls of Bayesian Inference of Phylogeny," *Systematic Biology*, vol. 51, no. 5, pp. 673–688, 2002.

[34] D. H. Huson, M. Steel, "Phylogenetic trees based on gene content," *Bioinformatics*, vol. 20, no. 13, pp. 2044–2049, 2004.

[35] S. Kannan, T. Warnow, "A fast algorithm for the computation and enumeration of perfect phylogenies when the number of character states is fixed," in *Proceedings of the 6th annual ACM-SIAM symposium on Discrete algorithms*, San Francisco, California, U.S.A., 1995, pp. 595-603.

[36] B. Kessler, *The Significance of Word Lists*. Stanford, California, U.S.A.: CSLI Publications, 2001.

[37] G. Kondrak, "A New Algorithm for the Alignment of Phonetic Sequences," in *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL 2000)*, vol. 4, Seattle, Washington, U.S.A., 2000, pp. 288-295.

[38] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707-710, 1966.

[39] M. P. Lewis, Ed., *Ethnologue: Languages of the World*, 16th ed. Dallas, Texas: SIL International, 2009.

[40] Matlab. Analyzing the Origin of the Human Immunodeficiency Virus http://www.mathworks.com/computational-biology/demos.html?file=/products/demos/shipping/bioinfo/hivdemo.html

[41] L. Nakhleh, T. Warnow, D. A. Ringe, S. N. Evans, "A comparison of phylogenetic reconstruction methods on an Indo-European dataset," *Transactions of the Philological Society*, vol. 103, no. 2, pp. 171-192, 2005.

[42] S. B. Needleman, C. D. Wunsch, "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443-453, March 1970.

[43] G. K. Nicholls, R. D. Gray, "Dated ancestral trees from binary trait data and their application to the diversification of languages," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 70, no. 3, pp. 545–566, July 2008.

[44] G. K. Nicholls, R. D. Gray, "Quantifying uncertainty in a stochastic model of vocabulary evolution," in *Phylogenetic methods and the prehistory of languages*, P. Forster and C. Renfrew, Eds. Cambridge, U.K.: McDonald Institute for Archaeological Research, 2006, pp. 161–171.

[45] J. Nichols, T. Warnow, "Tutorial on computational linguistic phylogeny," *Language and Linguistics Compass*, vol. 2, no. 5, pp. 760-820, 2008.

[46] F. Petroni, M. Serva, "Language distance and tree reconstruction," *Journal of Statistical Mechanics: Theory and Experiment*, vol. P08012, pp. 1-15, August 2008.

[47] K. Rexová, Y. Bastin, D. Frynta, "Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data," *Naturwissenschaften*, vol. 93, pp. 189–194, 2006.

[48] K. Rexová, D. Frynta, J. Zrzavý, "Cladistic analysis of languages: Indo-European classification based on lexicostatistical data," *Cladistics*, vol. 19, pp. 120–127, 2003.

[49] D. Ringe, T. Warnow, A. Taylor, "Indo-European and Computational Cladistics," *Transactions of the Philological Society*, vol. 100, no. 1, pp. 59-129, March 2002.

[50] M. Ruhlen, *The Origin of Language*.: John Wiley & Sons Inc, 1994.

[51] R. J. Ryder, G. K. Nicholls, "Missing data in a stochastic Dollo model for cognate data, and its application to the dating of Proto-Indo-European," *Journal of the Royal Statistical Society, Series C: Applied Statistics (In press)*.

[52] N. Saitou, M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406-425, 1987.

[53] M. Serva, F. Petroni, "Indo-Iuropean languages tree by Levenshtein distance," *EPL (Europhysics Letters)*, vol. 81, no. 6, pp. 68005-p1:p5, March 2008.

[54] T. F. Smith, M. S. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195-197, March 1981.

[55] R. R. Sokal, C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, vol. 38, pp. 1409- 1438, 1958.

[56] L. Steels, "Analogies between Genome and Language Evolution," in *Artificial Life IX: Proceedings of the 9th International Conference on the Simulation and Synthesis of Living Systems*, J. B. Pollack et al., Eds. Cambridge, Massachusetts, U.S.A.: The MIT Press, 2004.

[57] J. A. Studier, K. J. Keppler, "A Note on the Neighbor-Joining Algorithm of Saitou and Neil," *Journal of Molecular Biology and Evolution*, vol. 5, no. 6, pp. 729-731, 1988.

[58] M. Swadesh, "Lexico-Statistic Dating of Prehistoric Ethnic Contacts," *Proceedings of the American Philosophical Society*, vol. 96, no. 4, pp. 452-463, August 1952.

[59] M. Swadesh, "Towards Greater Accuracy in Lexicostatistics Dating," *International Journal of American Linguistics*, vol. 21, no. 2, pp. 121-137, April 1955.

[60] M. Turchi, N. Cristianini, "A Statistical Analysis of Language Evolution," in *The Evolution of Language: Proceedings of the 6th Internationl Conference (EVOLANG6)*, Rome, Italy, 2006, pp. 348-355.

[61] S. Wichmann, A. Saunders, "How to use typological databases in historical linguistic research," *Diachronica*, vol. 24, no. 2, pp. 373-404, 2007.