Classification of the Indo-European languages using a phylogenetic network approach

Alix Boc, Anna Maria Di Sciullo and Vladimir Makarenkov

Université du Québec à Montréal
Case postale 8888, succursale Centre-ville Montréal (Québec) H3C 3P8 Canada
boc.alix@courrier.uqam.ca, di_sciullo.anne-marie@uqam.ca and
makarenkov.vladimir@uqam.ca

Summary. Discovering the origin of the Indo-European (IE) language family is one of the most intensively studied problems in historical linguistics. Gray and Atkinson [6] inferred a phylogenetic tree (i.e., additive tree or X-tree [2]) of the IE family, using bayesian inference and rate-smoothing algorithms, based on the 87 Indo-European language data set collected by Dyen et al. [5]. When conducting their classification study, Gray and Atkinson assumed that the evolution of languages was strictly divergent and the frequency of borrowing (i.e., horizontal transmission of individual words) was very low. As consequence, their results suggested a predominantly tree-like pattern of the IE language evolution. In our opinion, only a network model can adequately represent the evolution of the IE languages. We propose to apply a method of horizontal gene transfer (HGT) detection [8] to reconstruct phylogenetic network depicting the evolution of the IE language family.

Key words: biolinguistics, historical linguistics, horizontal gene transfer, language classification, phylogenetic network, phylogenetic tree

1 Introduction

A number of curious parallels between the processes of historical linguistics and species evolution have been observed [1, 6, 11]. The evolutionary biologists and historical linguists often look for answering similar questions and face similar problems [1]. Recently, the theory and methodology of the two fields have evolved in remarkably similar ways. A number of important studies have considered the applications of phylogenetic methods to process language data (e.g., [1, 6, 11]). For instance, one of the most intensively studied topics is the evolution of the Indo-European (IE) language family ([4]). Gray and Atkinson [6] inferred a consensus phylogenetic tree of the IE language family using maximal likelihood models of lexical evolution, bayesian inference and rate-smoothing algorithms; the 87 Indo-European language data set collected by Dyen et al. [5] was analyzed in [6]. On the other hand, Rexová et al. [11]

also reconstructed a phylogeny of the IE languages when applying a cladistic methodology to study the same lexicostatistical data set [5]. The results obtained in [11] were very similar to those found in [6]. However, to reconstruct their phylogenies Gray and Atkinson, as well as Rexová et al., were constrained to assume that the evolution of languages was strictly divergent, each language was transmitted as a whole, and the frequency of borrowing (i.e., horizontal transmission of individual words) between languages was low. As consequence, the obtained results suggested a predominantly tree-like pattern of the IE language evolution with little borrowing of individual words.

In our opinion, only a phylogenetic network can adequately represent the evolution of this language family. A network model can incorporate the borrowing and homoplasy (i.e., evolutionary convergence) processes that influenced the evolution of the Indo-European languages. For example, although English is a Germanic language, it has borrowed around 50% of its total lexicon from French and Latin [10].

We propose to apply the methods of horizontal gene transfer (HGT) detection, which are becoming very popular among molecular biologists, in order to reconstruct the evolutionary network of the IE language family. The most frequent *horizontal word transfers*, representing borrowing events, will be added to the phylogenetic tree inferred by Gray and Atkinson (Fig. 1 in [6]) to represent the most important word exchanges which occurred during the evolution of the IE languages. In particular, a HGT detection algorithm ([8]) will be applied to build the evolutionary network of the IE languages.

In this article, we first outline the data in hand and then describe the new features of the HGT detection algorithm used to identify the word borrowing events. In the Results and discussion section, we present the obtained results for the 12 most important groups of the IE languages and report the words borrowing statistics. The most important word exchanges characterizing the evolution of this language family will be brought to light and discussed.

2 Description of the Dyen database

The database developed by Dyen et al. [5] includes the 200 words of the Swadesh list [14]. The Swadesh list is one of several lists of vocabulary with basic meanings, developed by Morris Swadesh in the 1940-50s [14], which is widely used in lexicostatistics (quantitative language relatedness assessment) and glottochronology (language divergence dating). Dyen et al. [5] built a database that provides cognation data among 95 Indo-European speech varieties. For each word meaning in the list of 200 basic meanings (chosen by Swadesh in 1952), the database contains the forms (e.g., word) used in the 95 speech varieties and the cognation decisions among the speech varieties made by Isidore Dyen in the 1960s. For each meaning, the forms were examined and cognation judgments were made [5]. The cognation judgments were made only between forms having the same meaning. The cognation judgments were recorded in classes of forms such that the forms in each class were "cognate" or "doubtfully cognate" with each other. Two forms, in two different speech varieties, were identified as "cognate" if within both of the varieties they had an unbroken history of descent from a common ancestral form. For example, since the English word FRUIT and French word FRUIT are known to be related by borrowing, they have been assigned different Cognate Classification Numbers (CCN) in the Dyen database [5]. Forms believed to be related by borrowing or by accidental similarity were thus not treated as cognate. In a small number of cases it was difficult to distinguish cognates from borrowing or accidental similarities; in this case they were treated as "doubtfully cognate" [5]. The cognate content information was used by Gray and Atkinson [6] to reconstruct the evolutionary tree of IE languages. In our study we also subdivided the 200 words of the Swadesh list into two broad categories : lexical (nouns and verbs; 138 words in total) and functional (adjectives + pronouns, conjunctions and determiners; 62 words in total) in order to see whether the rate of borrowing differs for these two broad categories.

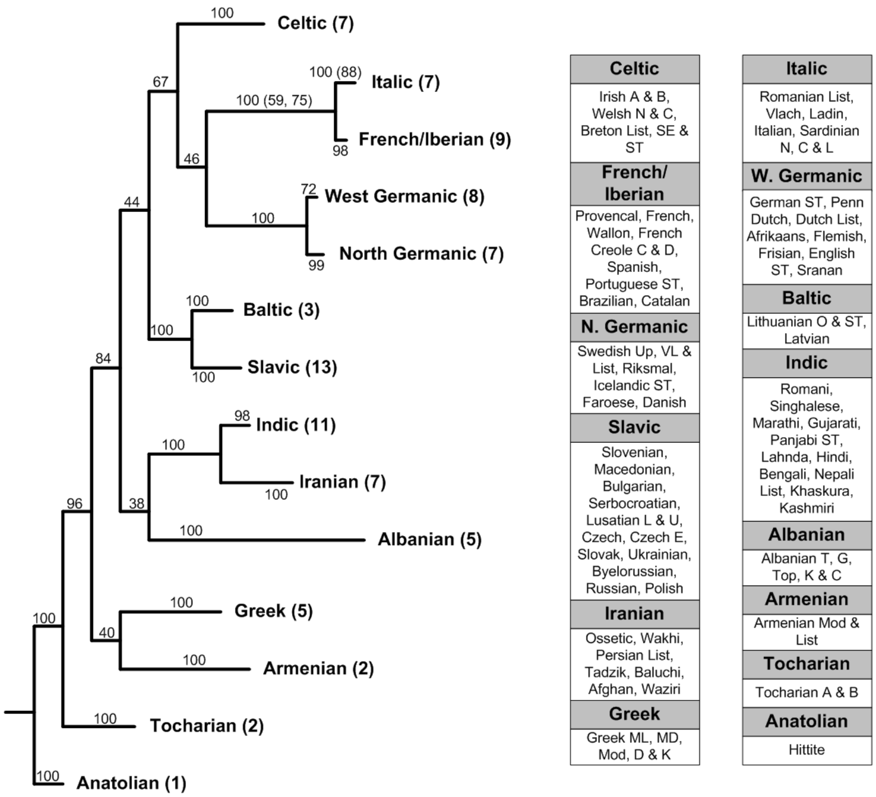
3 Materials and methods

In this section, we describe the new features of the HGT detection algorithm [8], applied here in a biolinguistics context, to infer a phylogenetic network of the IE languages family. When applied in a biological context, this algorithm identifies horizontal gene transfers (HGT) of a given gene for a given set of species thus reconciliating the species and gene phylogenetic trees. At each step of the reconciliation process, a HGT event is inferred. In this study, we draw a parallel between the HGT detection and the word borrowing detection processes. In our model, the IE languages tree (Fig. 1 and Fig. 4 in [6]) plays the role of the species tree and the word tree, representing the evolution of a given word (a given translation in all 87 considered languages), plays the role of the gene tree. The algorithmic procedure includes the three main steps, which are as follows:

Step 1. Let L be the rooted tree of 87 IE languages inferred by [6]. Fig. 1 shows a representation of this tree by groups (the group content is reported on the right). We also considered the 200 words of the Swadesh list [14] and their translations into 87 IE languages [5]. For each word of this list, we computed a distance matrix, $\mathbf{D}_i$ (87 x 87), i=1,...,200, between its translations using a normalized Levenshtein distance (Equation 1, [7]).

$$d(i,j) = \frac{Levenshtein\_distance(i,j)}{length(i) + length(j)}. \quad (1)$$

For each such matrix, we inferred the word phylogenetic tree W_i, using the Neighbor Joining method [13]. Fig. 2 shows the Robinson and Foulds (RF) topological distance [12] (normalized by its maximal value of $2n-6$ for two binary trees with n leaves) between each of the 200 word trees W_i and the



Celtic
Irish A & B, Welsh N & C, Breton List, SE & ST

French/ Iberian
Provencal, French, Wallon, French Creole C & D, Spanish, Portuguese ST, Brazilian, Catalan

N. Germanic
Swedish Up, VL & List, Riksmal, Icelandic ST, Faroese, Danish

Slavic
Slovenian, Macedonian, Bulgarian, Serbocroatian, Lusatian L & U, Czech, Czech E, Slovak, Ukrainian, Byelorussian, Russian, Polish

Iranian
Ossetic, Wakhi, Persian List, Tadzik, Baluchi, Afghan, Waziri

Greek
Greek ML, MD, Mod, D & K

Italic
Romanian List, Vlach, Ladin, Italian, Sardinian N, C & L

W. Germanic
German ST, Penn Dutch, Dutch List, Afrikaans, Flemish, Frisian, English ST, Sranan

Baltic
Lithuanian O & ST, Latvian

Indic
Romani, Singhalese, Marathi, Gujarati, Panjabi ST, Lahnda, Hindi, Bengali, Nepali List, Khaskura, Kashmiri

Albanian
Albanian T, G, Top, K & C

Armenian
Armenian Mod & List

Tocharian
Tocharian A & B

Anatolian
Hittite

Fig. 1. Gray and Atkinson (Fig. 1 in [6]) IE language evolutionary tree for 14 main language groups. The group content is indicated on the right. The numbers on the tree branches are the tree bootstrap scores; the number of languages for each group is indicated between parentheses.

language tree L. The average value of the normalized RF distance was 82 %. Such a high value suggests an important overall discrepancy between the language tree L and the word trees W_i (i=1,...,200).

Step 2. We applied the HGT detection algorithm ([8]) to infer word borrowings, considering, in turn, the language tree L and each of the 200 word trees W_i. Therefore, 200 different scenarios of tree reconciliation were computed. As the Dyen database [5] did not comprise any translation for the Hittite and Tocharian A and B languages, belonging to the Anatolian and Tocharian groups respectively, these languages were not considered in our analysis.

Step 3. We combined all results from the obtained transfer scenarios to compute the borrowing statistics. Intra-group transfers were ruled out in our computations because of the high risk of accidental similarity among the words from the same language group. First, we assessed the total numbers of transfers (i.e., number of word borrowings) between each pair of groups, and then

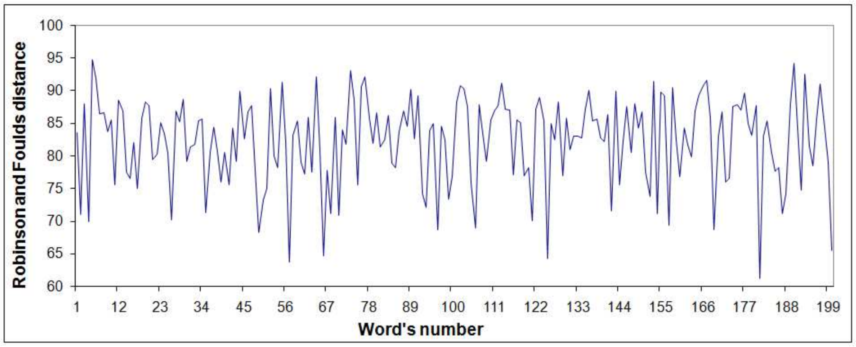


Fig. 2. Normalized Robinson and Foulds topological distance [12] between each of the word trees and the IE language tree in Fig. 1.

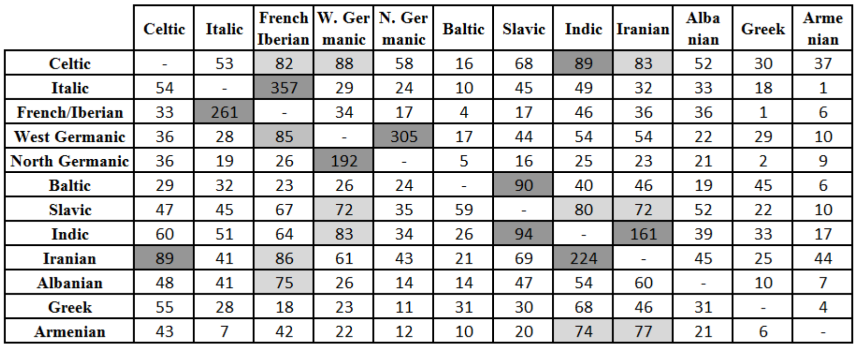
the percentages of words affected by these transfers in each group. The 10 most important transfers were mapped into the IE language tree (see Figs. 3, 4 and 5). These computations were carried out, first, for all 200 words and then, separately, for the words from the lexical and functional categories.

4 Results and discussion

Fig. 3 shows the total numbers of borrowed words found for each pair of groups. The 10 most active transfers are highlighted in dark grey. These transfers have been mapped into the IE language tree (Fig. 5a). If the geographical proximity can explain most of the frequent exchanges (e.g., between the West and North Germanic groups), some of them occur between the groups located far away from each other in the tree (e.g., between the Celtic and Indic, or Iranian and Celtic groups).

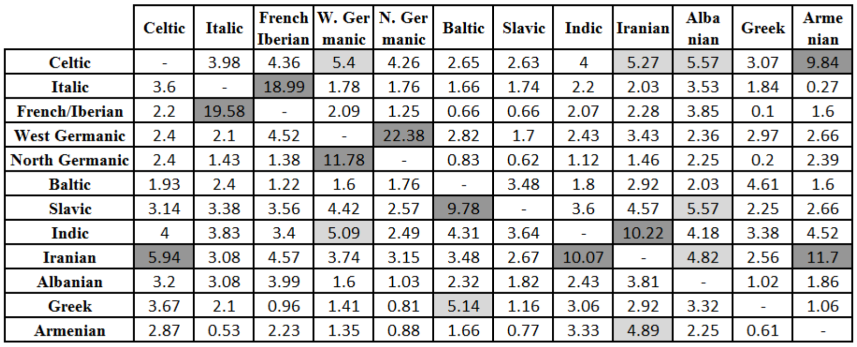
We can also observe a number of very active exchanges between the cluster combining the Indic and Iranian groups, and that combining the Celtic, Italic, French/Iberian, West/North Germanic and Slavic groups. These results suggest that despite the fact that the Iranian and Celtic groups are located far away from each other in the phylogenetic tree (Fig. 1), there is a strong relationship between them.

Fig. 4 reports the percentages of words of a given group affected by transfers originating from other groups. Similarly to the results reported in Fig. 3, the highest values were found for the neighbor groups. One can also notice that the cluster combining the Indic and Iranian groups has a sustained influence on the other groups. In the same way, we mapped the 10 most intensive transfers into the IE evolutionary tree (Fig. 5b). Some other high percentages (in light grey) can be explained either by well-known historical migration events (e.g., between the Armenian and Iranian groups) or should be investigated in more detail (e.g., between the Slavic and the Albanian groups). For instance,



	Celtic	Italic	French Iberian	W. Ger manic	N. Ger manic	Baltic	Slavic	Indic	Iranian	Alba nian	Greek	Arme nian
Celtic	-	53	82	88	58	16	68	89	83	52	30	37
Italic	54	-	357	29	24	10	45	49	32	33	18	1
French/Iberian	33	261	-	34	17	4	17	46	36	36	1	6
West Germanic	36	28	85	-	305	17	44	54	54	22	29	10
North Germanic	36	19	26	192	-	5	16	25	23	21	2	9
Baltic	29	32	23	26	24	-	90	40	46	19	45	6
Slavic	47	45	67	72	35	59	-	80	72	52	22	10
Indic	60	51	64	83	34	26	94	-	161	39	33	17
Iranian	89	41	86	61	43	21	69	224	-	45	25	44
Albanian	48	41	75	26	14	14	47	54	60	-	10	7
Greek	55	28	18	23	11	31	30	68	46	31	-	4
Armenian	43	7	42	22	12	10	20	74	77	21	6	-

Fig. 3. Total numbers of word borrowing events between each pair of language groups. For instance, 53 words of the Italic group were borrowed from the languages of the Celtic group; 10 highest values are highlighted in dark grey and 12 following highest values in light grey.



	Celtic	Italic	French Iberian	W. Ger manic	N. Ger manic	Baltic	Slavic	Indic	Iranian	Alba nian	Greek	Arme nian
Celtic	-	3.98	4.36	5.4	4.26	2.65	2.63	4	5.27	5.57	3.07	9.84
Italic	3.6	-	18.99	1.78	1.76	1.66	1.74	2.2	2.03	3.53	1.84	0.27
French/Iberian	2.2	19.58	-	2.09	1.25	0.66	0.66	2.07	2.28	3.85	0.1	1.6
West Germanic	2.4	2.1	4.52	-	22.38	2.82	1.7	2.43	3.43	2.36	2.97	2.66
North Germanic	2.4	1.43	1.38	11.78	-	0.83	0.62	1.12	1.46	2.25	0.2	2.39
Baltic	1.93	2.4	1.22	1.6	1.76	-	3.48	1.8	2.92	2.03	4.61	1.6
Slavic	3.14	3.38	3.56	4.42	2.57	9.78	-	3.6	4.57	5.57	2.25	2.66
Indic	4	3.83	3.4	5.09	2.49	4.31	3.64	-	10.22	4.18	3.38	4.52
Iranian	5.94	3.08	4.57	3.74	3.15	3.48	2.67	10.07	-	4.82	2.56	11.7
Albanian	3.2	3.08	3.99	1.6	1.03	2.32	1.82	2.43	3.81	-	1.02	1.86
Greek	3.67	2.1	0.96	1.41	0.81	5.14	1.16	3.06	2.92	3.32	-	1.06
Armenian	2.87	0.53	2.23	1.35	0.88	1.66	0.77	3.33	4.89	2.25	0.61	-

Fig. 4. Percentages of words affected by borrowing from other groups. For instance, 3.98% of the words of the Italic group have the Celtic origin. The same color notations as in Fig. 3, were adopted here.

Armenian borrowed so many words from the Iranian languages that it was at first considered a part of the Indo-Iranian languages, and was not recognized as an independent group of the Indo-European languages for many decades [15] (see the value of 11.7% for the transfers form Iranian to Armenian in Fig. 4). On the other hand, Baltic languages are extremely well preserved, retaining archaic features similar to ancient Latin and Greek. Similarities of the Baltic languages to ancient Greek (see the value of 5.14% for Greek to Baltic in Fig. 4) and Sanskrit (see value of 4.31% for Indic to Baltic in Fig. 4) were noted long ago by Franz Bopp, the founder of comparative linguistic [3]. Overally, 37% of the considered words were affected by borrowing from other language groups. The analogous results were obtained for the words of the lexical category (36.9%) and functional category (37.1%).

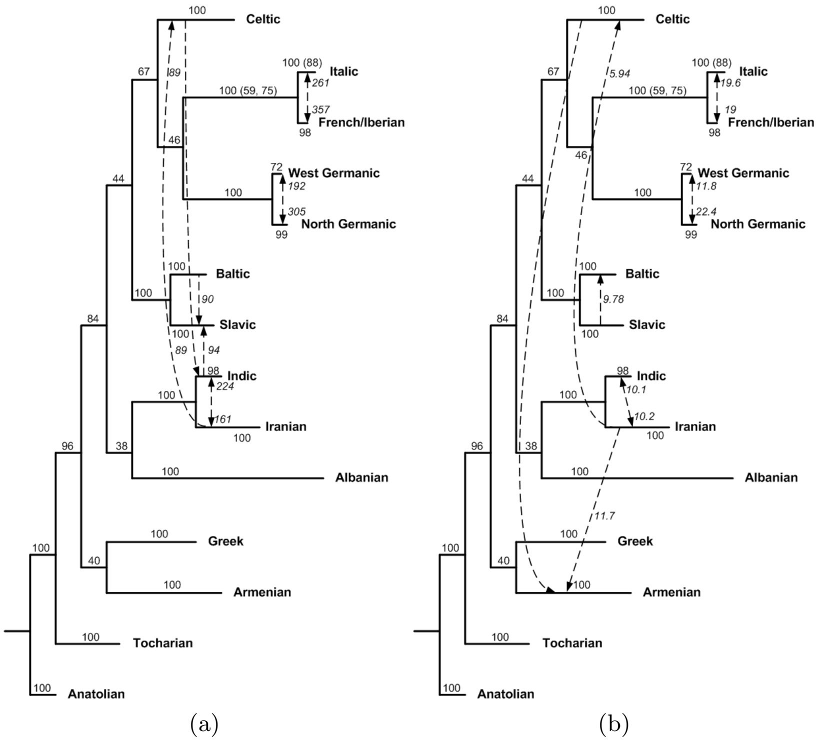


Fig. 5. Ten most frequent word exchanges between the IE language groups in terms of (a) total numbers of transferred words, and (b) percentages of affected words by group.

5 Conclusion

In this paper, we reconstructed a phylogenetic network of the Indo-European language family. The obtained network allowed us to represent the word borrowing events that have an important influence on the evolution of the IE languages. We found that 37% of the IE words are affected by borrowing from other IE groups. Very similar results were obtained for the lexical and functional categories. This means that the word borrowing process does not depend on the broad lexical/functional category. However, the obtained result should be interpreted with caution because some of the word similarities, even for words belonging to different language groups, can be due to accidental resemblance. In the future, we plan to conduct a refined study where the cognate content information [5] will be taken into account. This should eliminate the impact of the accidental word similarities. We also found that the clusters combining the Indic and Iranian groups, and the Celtic, Italic, French/Iberian, West/North Germanic groups have much closer relationships than it is represented in the traditional IE tree [6]. This may be the evidence of a much closer common ancestry between these two clusters or of an intensive migration of the ancestors of the involved nations. In the future, it would be important to carry out a more comprehensive words borrowing analysis based on the 850 words of the Basic English [9]. Basic English is an English-based controlled language created by Charles Kay Ogden [9] (in essence, a simplified subset of English) as an international auxiliary language. Such a new analysis could help find more recent activities of borrowing. It would be also interesting to establish a parallel between each of the determined high word borrowing activities (see Figs. 3 and 4) and the historical events, external to the internal language systems, such as wars, migrations, or important commercial trades between related nations.

References

1. Q.D. Atkinson and R.D. Gray. Curious parallels and curious connections: phylogenetic thinking in biology and historical linguistics. *Syst Biol*, 54:513-26, 2005.
2. J.-P. Barthelémy and A. Guénoche. *Trees and Proximity Representations, Wiley, New York*, 1991.
3. F. Bopp. A Comparative Grammar of the Sanskrit, Zend, Greek, Latin, Lithuanian, Gothic, German, and Slavonic Languages. Translated principally by Lieutenant Eastwick. *London: Madden and Malcolm*, 1845-1856, 1867.
4. J. Diamond and P. Bellwood. Farmers and their languages: the first expansions. *Science*, 300:597-603, 2003.
5. I. Dyen, J.B. Kruskal, and P. Black. Comparative IE Database Collected by Isidore Dyen, http://www.ntu.edu.au/education/langs/ielex/IE-RATE1. 1997.
6. R.D. Gray and Q.D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426:435-439, 2003.
7. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707-710, 1966.
8. V. Makarenkov, A. Boc, C.F. Delwiche, A.B. Diallo, and H. Philippe. New efficient algorithm for modeling partial and complete gene transfer scenarios. In V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna, editors, *IFCS 2006, Series: Studies in Classification, Data Analysis, and Knowledge Organization, Springer Verlag*, 341-349, 2006.
9. C. K. Ogden. Basic English: A General Introduction with Rules and Grammar. Publisher: Paul Treber & Co., Ltd. London, 1930.
10. M. Pagel. In Time Depth in Historical Linguistics. In C. Renfrew, A. McMahon, and L. Trask, editors, 189-207, 2000.
11. K. Rexová, D. Frynta, and J. Zrzavý. Cladistic analysis of languages: indo-european classification based on lexicostatistical data. *Cladistics*, 19:120-27, 2003.
12. D.R. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Math Biosc*, 53:131-147, 1981.
13. N. Saitou and M. Nei. The neighbor-joining method:a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4:406-425, 1987.
14. M. Swadesh. Lexico-statistical dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. In Proceedings of the *American Philosophical Society*, 96:452-463, 1952.
15. J. Waterman. A history of the German language. *U of Washington Press*, 1976.