A LETTERS JOURNAL EXPLORING THE FRONTIERS OF PHYSICS

OFFPRINT

Indo-European languages tree by Levenshtein distance

M. SERVA and F. PETRONI

EPL, **81** (2008) 68005

TAKE A LOOK AT THE NEW EPL

Europhysics Letters (EPL) has a new online home at
www.epljournal.org

Take a look for the latest journal news and information on:

- reading the latest articles, free!
- receiving free e-mail alerts
- submitting your work to EPL

www.epljournal.org

A Letters Journal Exploring the Frontiers of Physics

March 2008

EPL, **81** (2008) 68005 www.epljournal.org
doi: 10.1209/0295-5075/81/68005

Indo-European languages tree by Levenshtein distance

M. Serva[1] and F. Petroni[2]

[1] *Dipartimento di Matematica, Università dell'Aquila - I-67010 L'Aquila, Italy*
[2] *GRAPES, B5, Sart Tilman - B-4000 Liège, Belgium*

received 17 October 2007; accepted in final form 30 January 2008
published online 29 February 2008

PACS 87.23.Ge – Dynamics of social systems
PACS 87.23.Kg – Dynamics of evolution
PACS 89.75.Hc – Networks and genealogical trees

Abstract – The evolution of languages closely resembles the evolution of haploid organisms. This similarity has been recently exploited (Gray R. D. and Atkinson Q. D., *Nature*, **426** (2003) 435; Gray R. D. and Jordan F. M., *Nature*, **405** (2000) 1052) to construct language trees. The key point is the definition of a distance among all pairs of languages which is the analogous of a genetic distance. Many methods have been proposed to define these distances; one of these, used by glottochronology, computes the distance from the percentage of shared "cognates". Cognates are words inferred to have a common historical origin, and subjective judgment plays a relevant role in the identification process. Here we push closer the analogy with evolutionary biology and we introduce a genetic distance among language pairs by considering a renormalized Levenshtein distance among words with same meaning and averaging on all words contained in a Swadesh list (Swadesh M., *Proc. Am. Philos. Soc.*, **96** (1952) 452). The subjectivity of process is consistently reduced and the reproducibility is highly facilitated. We test our method against the Indo-European group considering fifty different languages and the two hundred words of the Swadesh list for any of them. We find out a tree which closely resembles the one published in Gray and Atkinson (2003), with some significant differences.

Introduction. – Glottochronology uses the percentage of shared "cognates" between languages to calculate their distances. These "genetic" distances are logarithmically proportional to divergence times if a constant rate of lexical replacement is assumed. Cognates are words inferred to have a common historical origin, their identification is often a matter of sensibility and personal knowledge. Therefore, subjectivity plays a relevant role. Furthermore, results are often biased since it is easier for European or American scholars to find out those cognates belonging to western languages. For instance, the Spanish word *leche* and the Greek word *gala* are cognates. In fact, *leche* comes from the Latin *lac* with genitive form *lactis*, while the genitive form of *gala* is *galactos*. This identification is possible because of our historical records, hardly it would have been possible for languages, let us say, of central Africa.

Our aim is to avoid this subjectivity and construct a languages' tree which can be easily replicated by other scholars. To reach this goal, we compare words with same meaning belonging to different languages only considering orthographical differences. More precisely, we use a modification of the Levenshtein distance (or edit distance) to measure the distance between pairs of words in different languages. The edit distance is defined as the minimum number of operations needed to transform one word into another, where an operation is an insertion, a deletion, or substitution of a single character. Our definition of genetic distance between two words is taken as the edit distance divided by the number of characters of the longer of the two. With this definition, the distance can take any value between 0 and 1. To understand why we renormalize, let us consider the following case of one substitution between two words: if the compared words are long even if the difference between them is given by one substitution they remain very similar; while, if these words are short, let us say two characters, one substitution is enough to make them completely different. Without renormalization, the distance between the words compared in the two examples would be the same, no matter their length. Instead, introducing the normalization factor, in the first case the genetic distance is much smaller than in the second one.

We use a distance between words pairs, as defined above, to construct a distance between pairs of languages. The first step is to find lists of words with the same meaning for all languages for which we intend to construct a distance. Then, we compute the genetic distance for each pair of

words with the same meaning in one languages' pair. Finally, the distance between each languages' pair is defined as the average of the distance between words pair. As a result we have a number between 0 and 1 which we claim to be the genetic distance between two languages.

We stress that in our approach grammatical rules are not taken into account. This can be seen as a first approximation to compare languages.

Languages database. – The database we use for the present analysis [1] is composed by 50 languages with 200 words for each of them. Words are chosen according to Swadesh lists [2]. We stress that the Swadesh list considers the most common words which are used in all cultures and which are the most resistant to changes and borrowings. All languages considered belong to the Indo-European group. The database is a selection/modification of the one used in [3], where some errors have been corrected, and many missing words have been added. In the database only the English alphabet is used (26 characters plus space); those languages written in a different alphabet (*i.e.* Greek, etc.) were already transliterated into the English one in [3]. For some of the languages in our lists [1] there are still few missing words for a total number of 43 in a database of 9957. When a language has one or more missing words, these are simply not considered in the average that brings to the definition of distance. This implies that for some pairs of languages, the number of compared words is not 200 but a number always greater than or equal to 187. There is no bias in this procedure, the only effect is that the statistic is slightly reduced.

The result of the analysis described above is a 50×50 upper triangular matrix which expresses the 1225 distances among all languages' pairs.

Indeed, our method for computing distances is a very simple operation, that does not need any specific linguistic knowledge and requires a minimum computing time.

Time distance between languages. – A phylogenetic tree can be built already from this matrix, but this would only give the topology of the tree, whereas the absolute time scale would be missing. In order to have this quantitative information, some hypotheses on the time evolution of genetic distances are necessary. We assume that the genetic distance among words, on the one hand tends to grow due to random mutations and on the other hand it may reduce since different words may become more similar by accident or, more likely, by language borrowings.

Therefore, the distance D between two given languages can be thought to evolve according to the simple differential equation

$$\dot{D} = -\alpha\,(1-D) - \beta D, \tag{1}$$

where $\dot{D}$ is the time derivative of D. The parameter α is related to the increase of D due to random permutations, deletions or substitutions of characters (random mutations) while the parameter β considers the possibility that two words become more similar by a "lucky" random mutation or by words borrowing from one language to the other or both from a third one. Since α and β are constant, it is implicitly assumed that mutations and borrowings occur at a constant rate.

Note that with this choice, word substitution is statistically equivalent to the substitution of all characters in the word itself. The first reason for this approximation is the economy of parameters to be used in the model. The second, and more important, is that it is very hard to establish if a word has changed because many characters have been replaced or if the whole word has been replaced. This would be possible only by historical knowledge of the languages and it would imply again the use of cognates and a subjective analysis of the problem, something that we want to avoid with our model. In fact, the main point for the model is to use the fastest and simplest algorithm to compare languages. More complicated models could be introduced with more parameters and even with time-dependent parameters. But at the present stage we prefer to keep the model as simple as possible.

At time $T = 0$ two languages begin to separate and the genetic distance D is zero. With this initial condition the above equation can be solved and the solution can be inverted. The result is a relation which gives the separation time T (time distance) between two languages in terms of their genetic distance D,

$$T = -\epsilon\,\ln(1-\gamma D). \tag{2}$$

The values for the parameters $\epsilon = 1/(\alpha+\beta)$ and $\gamma = (\alpha+\beta)/\alpha$ can be fixed experimentally by considering two pairs of languages whose separation time (time distance) is known. We have chosen a distance of 1600 years between Italian and French and a distance of 1100 years between Icelandic and Norwegian. The resulting values of the parameters are $\epsilon = 1750$ and $\gamma = 1.09$, which corresponds to the following values $\alpha \cong 5 * 10^{-4}$ and $\beta \cong 6 * 10^{-5}$. This means that similar words may become more different at a rate that is about ten times the rate at which different words may become more similar. It should be noticed that (2) closely resembles the fundamental formula of glottochronology.

A time distance T is then computed for all pairs of languages in the database, obtaining a 50×50 upper triangular matrix with 1225 non-trivial entries. This matrix preserves the topology of the genetic distance matrix but it contains all the information concerning absolute time scales.

The phylogenetic tree in fig. 1 is constructed from the matrix using the Unweighted Pair Group Method Average (UPGMA). We use the UPGMA for its coherence with the trees associated with the coalescence process of Kingamnn type [4]. In fact, the process of languages' separation and extinction closely resembles the population dynamics associated with haploid reproduction which holds for simple organisms or for the mitochondrial DNA of complex ones as humans. This dynamics, introduced

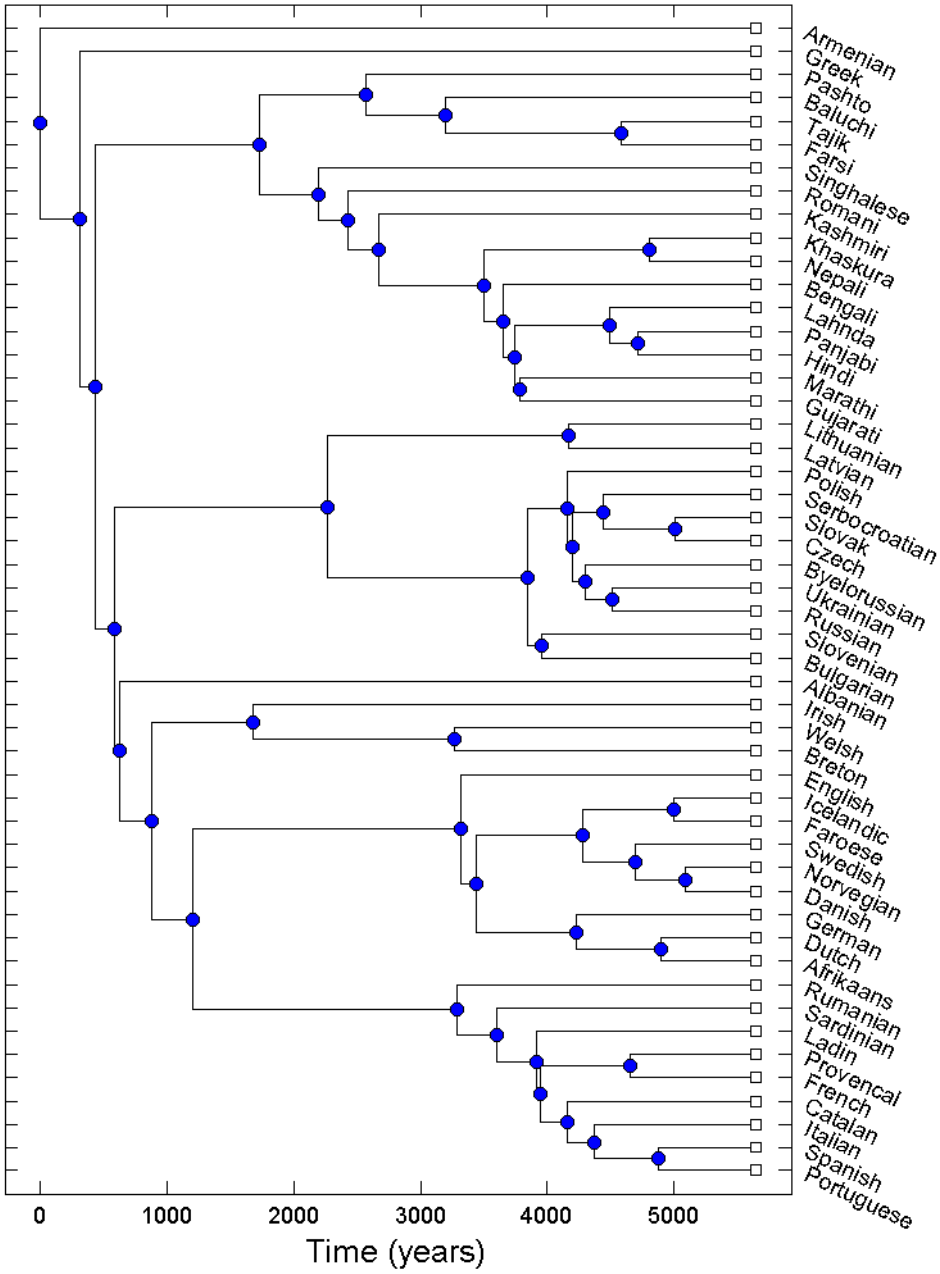

Fig. 1: Phylogenetic tree constructed from the matrix of distances using the UPGMA.

by Kingman, has been extensively studied and described, see for example [5,6]. In particular, in these two papers the distribution of distances is found and plotted and can be usefully compared with the one herein obtained and plotted in fig. 2. It should be considered that in the model of Kingman, time distances have the objective meaning of measuring time from separation while in our realistic case time distances are reconstructed from genetic distances. In this reconstruction we assume that lexical mutations and borrowings occur at a constant rate. This is true only on average, since there is an inherent randomness in this process which is not taken into account by the deterministic differential equation (1). Furthermore, the parameters α and β may vary from a pair of languages to another and also they may vary in time according to historical conditions. Therefore, the distribution in fig. 2 is not exactly the distribution in [5,6] but it could be obtained from them after a random shift of all distances.

In this paper we do not consider dead languages, we suspect, in fact, that results would be biased due to the different times in which languages existed. For example, comparison of Latin with its offspring could be a meaningless operation in the context of this research. We think that, eventually, Latin should be compared with

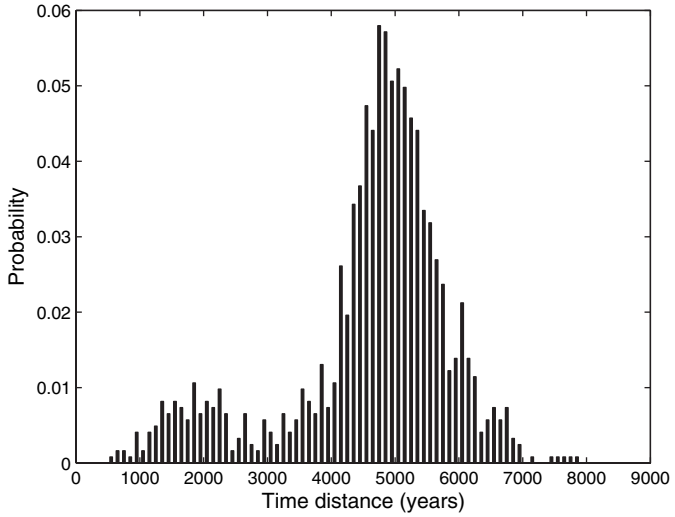

Fig. 2: Distribution of distances obtained from all the 1225 pairs of languages.

its contemporary languages and their genealogical tree constructed.

Methods. –

Database. The database used here to construct the phylogenetic tree is composed by 50 languages of the Indo-European group. The main source for the database is the file prepared by Dyen *et al.* in [3] which contains a Swadesh list of 200 words for 96 languages. This list consists of items of basic vocabulary, like body parts, pronouns, numbers, which are known to be resistant to borrowings. Many words are missing in [3] but we have filled most of the gaps by finding the words on Swadesh lists and on dictionaries freely available on the web. The file [3] also contains information on "cognates" among languages that we do not use in this work. Our selection of 50 languages is based on [3] but we avoid to consider more than one version of the same language. For example, we do not consider both Irish A and Irish B, but we choose only one of them and we do not include Brazilian but only Portuguese. Our choice among similar languages is based on keeping that language with fewer gaps. Our database is available at [1].

Tree construction. In this work the normalized Levenshtein distance is used to build up an upper triangular 50×50 matrix with 1225 entries representing the pairwise distances corresponding to 50 languages. These genetic distances are translated into time distances between languages, pairs and a new matrix of same size comes out. Then, the simple phylogenetic algorithm UPGMA [7] for tree construction is used. The reason why we choose the unweighted pair-group method, using arithmetic average (UPGMA), is that it is the most coherent with the hypothesis that the languages' tree is generated by a coalescence process of Kingman type [4]. Let us describe briefly how this algorithm works. It first identifies the two languages with shortest time distance and then it treats this pair as a new single object whose distance from the other languages is the average of the distance of its two components. Subsequently, among the new group of objects it identifies the pair with the shortest distance, and so on. At the end, one is left with only two objects (languages' clusters) which represent the two main branches at the root of the tree. We remark that, as a consequence of the construction rule, the distance between two branches is the average of the distances among all pairs of languages belonging to the two branches.

Conclusions. – We would like to compare now our results with those published in [8]. The tree in fig. 1 is similar to the one in [8] but there are some important differences. First of all, the first separation concerns Armenian, which forms a separate branch close to the root, while the other branch contains all the remaining Indo-European languages. Then, the second one is that of Greek, and only after there is a separation between the European branch and the Indo-Iranian one. This is the main difference with the tree in [8], since therein the separation at the root gives origin to two branches, one with Indo-Iranian languages plus Armenian and Greek, the other with European languages. The position of Albanian is also different: in our case it is linked to European languages while in [8] it goes with Indo-Iranian ones. Finally, the Romani language is correctly located together with Indian languages but it is not as close to Singhalese as reported in [8].

In spite of this differences, our tree seems to confirm the same conclusions reported in [8] about the Anatolian origin of the Indo-European languages, in fact, in our research, the first separation concerns the languages geographically closer to Anatolia, that is to say Armenian and Greek.

We note that while the tree reported in [8] is the result of a Bayesian posterior distribution of trees, due to how our algorithm is constructed, our tree is unique and is the outcome of a direct comparison of languages.

We want to stress that the method used here is very simple and does not require any previous knowledge of languages' origin. Also, it can be applied directly to all those language pairs for which a translation of a small group of words exists. The results could be improved if more words were added to the database and if translations and transliterations were made more accurate. Since our method is very easy to use, being the only difficulty the procedure of collecting words, we plan to extend our study to other languages' families and eventually to test competing hypotheses concerning super families, or to test controversial classifications as, for example, the case of Japanese.

* * *

We thank J. Raboanary for many discussions and examples from Malagasy dialects concerning the applicability of the Levenshtein distance to linguistics. Critical comments on many aspects of the paper by

M. Ausloos are also gratefully acknowledged. The work by FP has been supported by the European Commission Project E2C2 FP6-2003-NEST-Path-012975 Extreme Events: Causes and Consequences.

REFERENCES

[1] The database, as modified by the authors, is available at the following web address: `http://univaq.it/~serva/languages/languages.html`. Readers are welcome to modify, correct and add words to the database.

[2] Swadesh M., *Proc. Am. Philos. Soc.*, **96** (1952) 452.

[3] Dyen I., Kruskal J. B. and Black P., *FILE IE-DATA1*, available at `http://www.wordgumbo.com/ie/cmp/iedata.txt` (1997).

[4] Kingman J. F. C., *J. Appl. Probab.*, **19A** (1982) 27.

[5] Serva M., *J. Stat. Mech.: Theory Exp.* (2005) P07011.

[6] Simon D. and Derrida B., *J. Stat. Mech.: Theory Exp.* (2006) P05002.

[7] Sneath P. H. A. and Sokal R. R., *Numerical Taxonomy* (Freeman, San Francisco) 1973.

[8] Gray R. D. and Atkinson Q. D., *Nature*, **426** (2003) 435.