Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Language distance and tree reconstruction

Filippo Petroni[1] **and Maurizio Serva**[2]

[1] GRAPES, B5, Sart Tilman, B-4000 Liège, Belgium
[2] Dipartimento di Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy
E-mail: fpetroni@gmail.com and serva@univaq.it

Received 5 June 2008
Accepted 28 July 2008
Published 22 August 2008

Online at stacks.iop.org/JSTAT/2008/P08012
doi:10.1088/1742-5468/2008/08/P08012

Abstract. Languages evolve over time according to a process in which reproduction, mutation and extinction are all possible. This is very similar to haploid evolution for asexual organisms and for the mitochondrial DNA of complex ones. Exploiting this similarity, it is possible, in principle, to verify hypotheses concerning the relationship among languages and to reconstruct their family tree. The key point is the definition of the distances among pairs of languages in analogy with the genetic distances among pairs of organisms. Distances can be evaluated by comparing grammar and/or vocabulary, but while it is difficult, if not impossible, to quantify grammar distance, it is possible to measure a distance from vocabulary differences. The method used by glottochronology computes distances from the percentage of shared 'cognates', which are words with a common historical origin. The weak point of this method is that subjective judgment plays a significant role. Here we define the distance of two languages by considering a renormalized edit distance among words with the same meaning and averaging over the two hundred words contained in a Swadesh list. In our approach the vocabulary of a language is the analogue of DNA for organisms. The advantage is that we avoid subjectivity and, furthermore, reproducibility of results is guaranteed.

We apply our method to the Indo-European and the Austronesian groups, considering, in both cases, fifty different languages. The two trees obtained are, in many respects, similar to those found by glottochronologists, with some important differences as regards the positions of a few languages. In order to support these different results we separately analyze the structure of the distances of these languages with respect to all the others.

Keywords: phylogeny (theory), population dynamics (theory), analysis of algorithms, new applications of statistical mechanics

1742-5468/08/P08012+15$30.00

Contents

1. Introduction 2
2. Method and database 3
3. Time distance between languages 4
4. Tree and distance distributions 5
5. Malagasy and Romani 11
6. Discussion and conclusions 13
Acknowledgments 15
References 15

1. Introduction

The Swadesh list, in contrast to a list of arbitrary words, contains terms which are common to all cultures and which concern the basic activities of humans. The choice is motivated by the fact that the vocabulary learned during childhood changes very slowly over time. The use of Swadesh lists in glottochronology has been popular for half a century. Glottochronologists use the percentage of shared *cognates* in order to compute the distances between pairs of languages. These *lexical* distances are on average logarithmically proportional to the divergence time. In fact, changes in vocabulary accumulate year after year and two languages initially similar become more and more different. Recent examples of the use of Swadesh lists and cognates to construct language trees are the studies of Gray and Atkinson [1] and Gray and Jordan [2]. Cognates are words inferred to have a common historical origin; their identification is often a matter of sense and personal knowledge. In fact, the task of counting the number of cognate words in the list is far from trivial because cognates do not necessarily look similar. Therefore, subjectivity plays a significant role. Furthermore, results are often biased since it is easier for European or American scholars to find 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*. Also the English *wheel* and Hindi *cakra* are cognates. These two identifications are possible because of our historical records; it would hardly have been possible for languages of, let us say, Central Africa or Australia.

In this paper we use an automated method which avoids any subjectivity so that our results can be replicated by other scholars assuming that the database used is the same. For any language we write down a list of the same 200 words according to the original choice of Swadesh [3]; then we compare words with the same meaning belonging to different languages only considering orthographical differences. This may appear reductive since words may look similar by chance, while cognate words may have a completely different orthography, but we will try to convince the reader that this is indeed a simpler, more objective and more efficient choice than the traditional glottochronological approach.

J. Stat. Mech. (2008) P08012

In section 2 we describe how we measure the distance between words and then between languages. Once all the distances among pairs of languages are given it is possible to find the time separations (section 3) and finally the genealogical trees (section 4). In this work we include some results previously obtained in [4] concerning the Indo-European group of languages and we extend our results to the Austronesian group. For each of the two groups we consider 50 languages and we obtain two genealogical trees which are similar to those found by [1] and [2], with some important differences as regards the positions of a few languages and subgroups. Indeed we think that these differences carry some new information about the structure of the tree and about the positions of some languages, for which we separately analyze the structure of distances with respect to all the others (section 5).

2. Method and database

In order to measure distances between pairs of words in different languages we use a modification of the Levenshtein distance (or edit distance) which is defined as the minimum number of operations needed to transform one word into another. An operation is an insertion, deletion, or substitution of a single character. Our definition of lexical distance between two words is taken as the edit distance normalized by the number of characters of the longer of the two. The reason for the renormalization is easily understood: if two words differ by one character this is much more important for short words than it is for long words. This is taken into account by the renormalization factor.

We use distance between word 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 lexical distance for each pair of words with the same meaning in one language pair. Finally, the distance of each language pair is defined as the average of the distances between the word pairs. As a result we have a number between 0 and 1 which we claim to be the lexical distance between the two languages.

The databases used here[3] to construct the phylogenetic trees are composed of 50 languages of the Indo-European group and 50 languages of the Austronesian group. The main source for the database of the Indo-European group is the file prepared by Dyen *et al* in [5] which contains the 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 [5] but for our choice of 50 languages we have filled most of the gaps and corrected some errors by finding the words on Swadesh lists and in dictionaries freely available on the Web. For the Austronesian group we used, as the main source, the lists contained in the huge database available in [6]. The lists in [6] contain more than 200 words which do not coincide completely with the words in the original Swadesh list [3]. For our choice of 50 Austronesian languages we have retained only the words which are in [6] and also are in the original Swadesh list. The list obtained has many gaps due to missing words in [6] and because of the incomplete overlap of [6] with the original Swadesh list. Also in this case we have filled some of the gaps by finding

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

J. Stat. Mech. (2008) P08012

the words on Swadesh lists available on the Web and in one case (Malagasy) from direct knowledge of the language.

Both databases [5] and [6] contain information on 'cognates' for languages that we do not use in this work.

Our selection of the 50 Indo-European languages is based on [5] but we do not consider more than one version of the same language. For example, we do not consider both Irish A and Irish B; we choose only one of them, and we do not include Brazilian, only Portuguese. Our choice among similar languages is based on keeping that language with fewer gaps. Our choice of languages in the Austronesian group has the aim of reproducing all the main subgroups and the main branches of the tree.

In both databases 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 alphabet in [5] and [6]. Furthermore, in [6] many additional characters are used which we have transliterated into the basic English alphabet plus space.

For some of the languages in our lists (see footnote 3) there are still a few missing words: a total number of 43 in the Indo-European group and 1575 in the Austronesian group. When a language has one or more missing words, these are simply not considered in the average that contributes to the definition of distance between two languages. This implies that for some pairs of languages, the number of compared words is not 200 but a smaller number. There is no bias in this procedure; the only effect is that the statistic is slightly reduced.

The results of the analysis described above are two 50×50 upper triangular matrices with the lexical distances of the languages of the two groups. Each matrix contains the 1225 distances among all pairs in a group.

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.

3. Time distance between languages

A phylogenetic tree can be built already from one of these matrices, but this would only give the topology of the tree; the absolute timescale would be missing. In order to have this quantitative information, some hypotheses on the time evolution of lexical distances are necessary. We assume that the lexical distance among words on one hand tends to grow due to random mutations and on the other hand may reduce since different words may become more similar by accident or, more likely, through 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 increasing 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 word borrowing from one language to the other or by both from a third one. Since α and β are constant, it is implicitly assumed that mutations and borrowings occur at a constant rate.

J. Stat. Mech. (2008) P08012

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 whether a word has changed because many characters have been replaced or whether the whole word has been replaced. This would be possible only through 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 lexical 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 lexical 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 parameter are $\epsilon = 1750$ and $\gamma = 1.09$ which corresponds to the following values: $\alpha \cong 5 \times 10^{-4}$ and $\beta \cong 6 \times 10^{-5}$. 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. We use this choice of the parameters both for the Indo-European and Austronesian groups.

A time distance T is then computed for all pairs of languages in the database, obtaining two 50×50 upper triangular matrices with 1225 non-trivial entries. These matrices preserve the topology of the lexical distance matrices but contain all the information as regards absolute timescales.

In order to check the validity of the parameters used in the model we verified that the time distances between other pairs of languages estimated from the model were in the range of the time distances known from historical studies. In particular, we obtained from the model a time distance of about 2000 years for Italian and Rumanian.

4. Tree and distance distributions

The phylogenetic trees in figures 1 and 2 are constructed from the matrix using the unweighted pair group method average (UPGMA) [7]. The algorithm works as follows. It first identifies the two languages with the 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 distances 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 (language clusters) which represent the two main branches at the root of the tree.

J. Stat. Mech. (2008) P08012

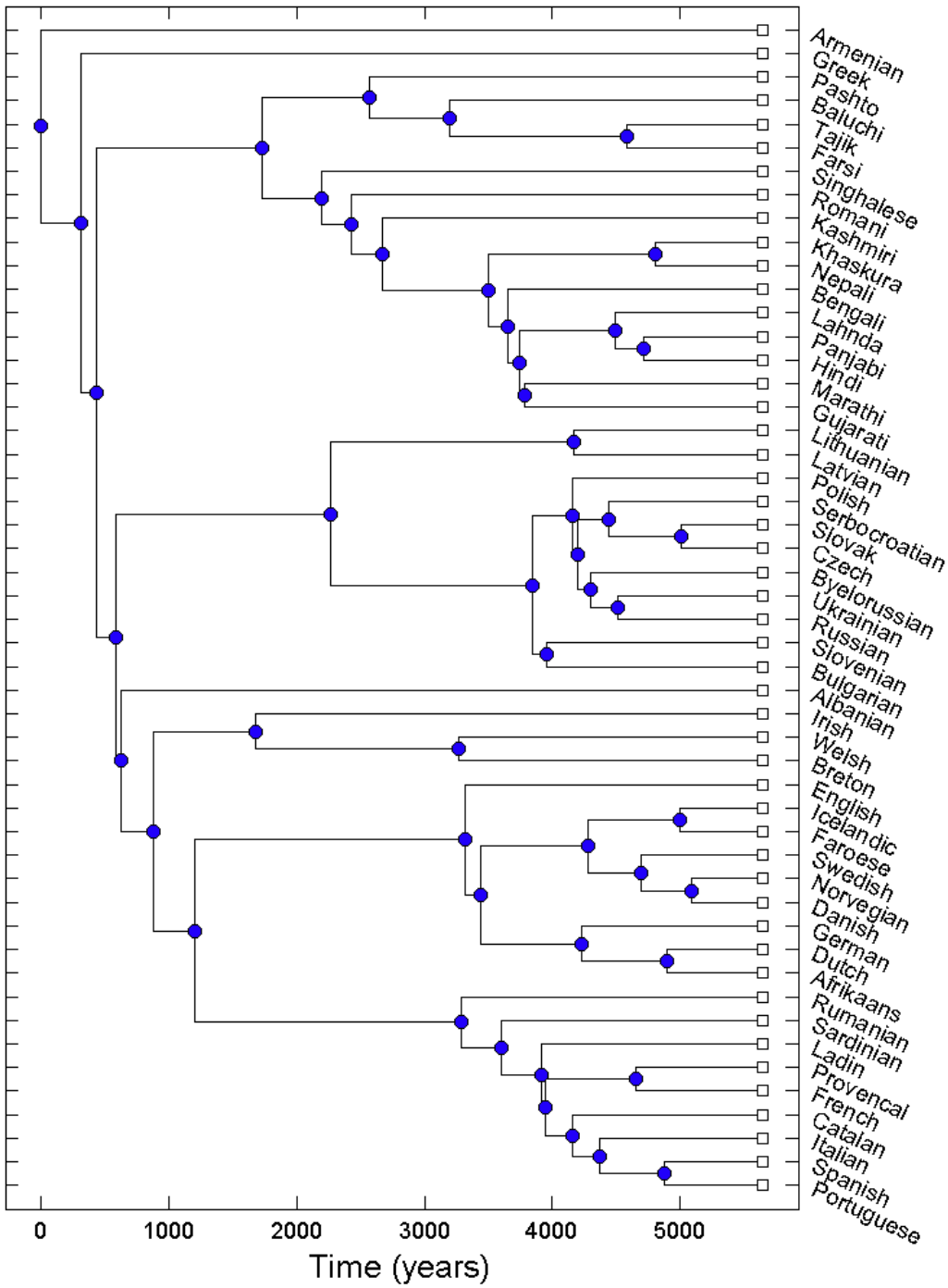

Figure 1. Indo-European phylogenetic tree constructed from the matrix of distances using the UPGMA algorithm.

To check the stability of the main results (separation into subgroups) obtained from the phylogenetic trees constructed using the UPGMA algorithm, we computed many trees where given numbers of languages were removed randomly. The computation of these trees shows a certain stability in the main features of the trees, namely all the distances between subgroups are stable if a few languages of each subgroup are removed. To show this result

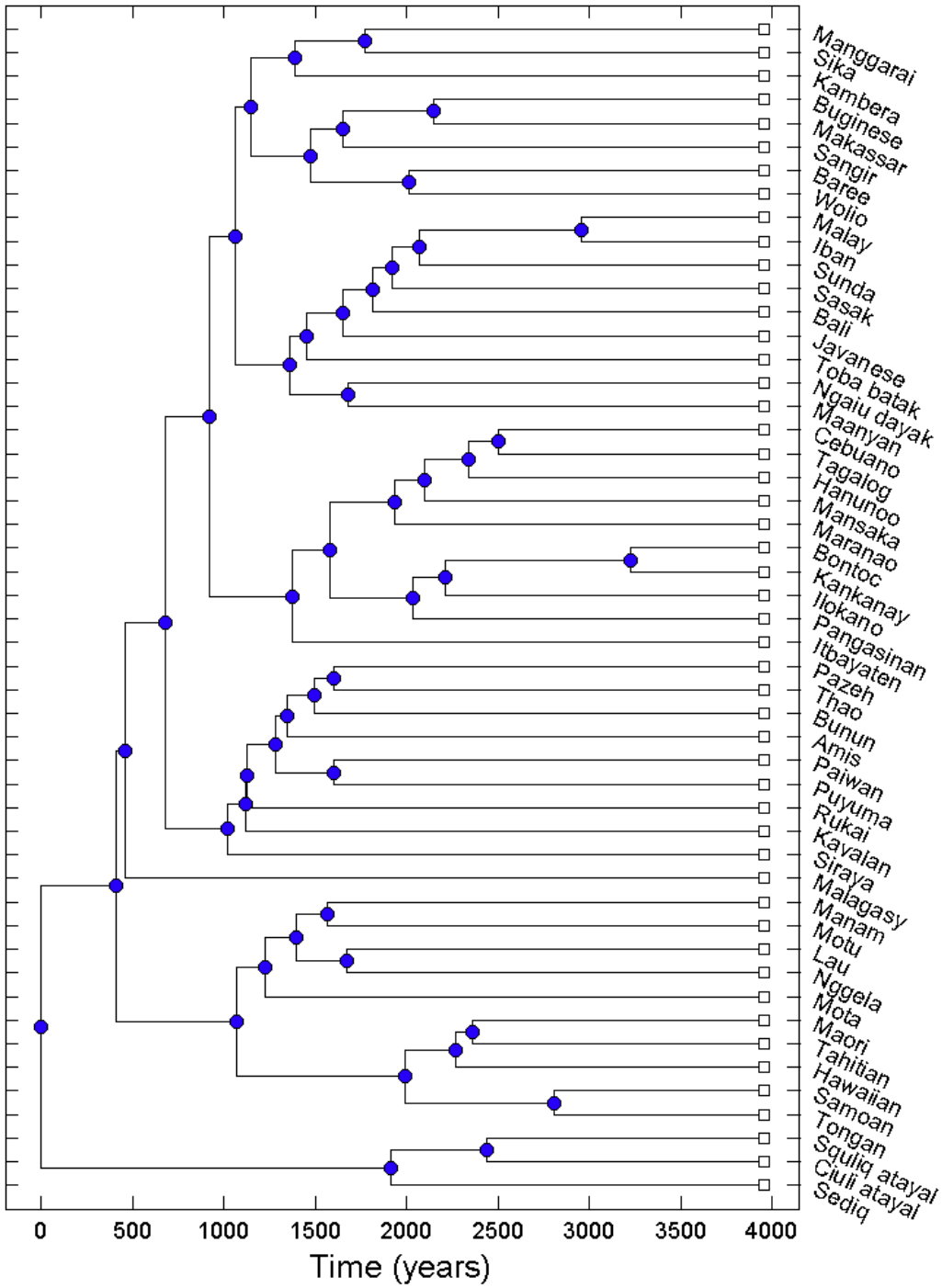

Figure 2. Austronesian phylogenetic tree constructed from the matrix of distances using the UPGMA algorithm.

we plot in figure 3 a phylogenetic tree from the Indo-European group where 25 languages from the original database were randomly removed.

We use UPGMA algorithm for its coherence with the trees associated with the coalescence process of Kingman type [8]. In fact, the process of language separation and

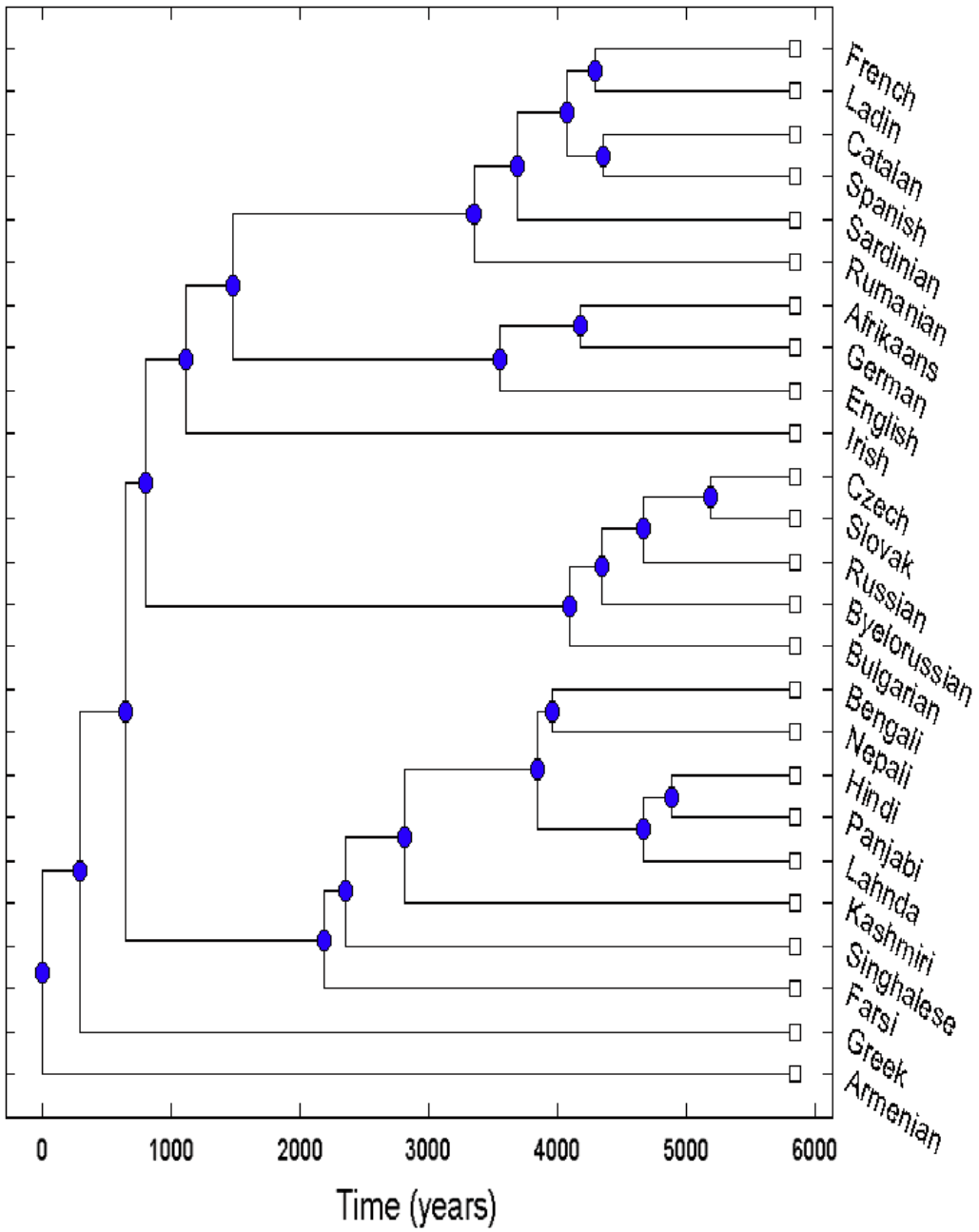

Figure 3. Indo-European phylogenetic tree constructed from a subsample of 25 languages randomly chosen from the original database.

extinction closely resembles the population dynamics associated with haploid reproduction which holds for simple organisms or for the mitochondrial DNA of complex ones such as humans. This dynamics, introduced by Kingman, has been extensively studied and described; see for example [9, 10]. In particular, in these two papers the distribution of distances is found and plotted and can be usefully compared with the one obtained herein and plotted in figures 4 and 5. It should be considered that in the model of Kingman, time distances have the objective meaning of measuring time from separation, while in our

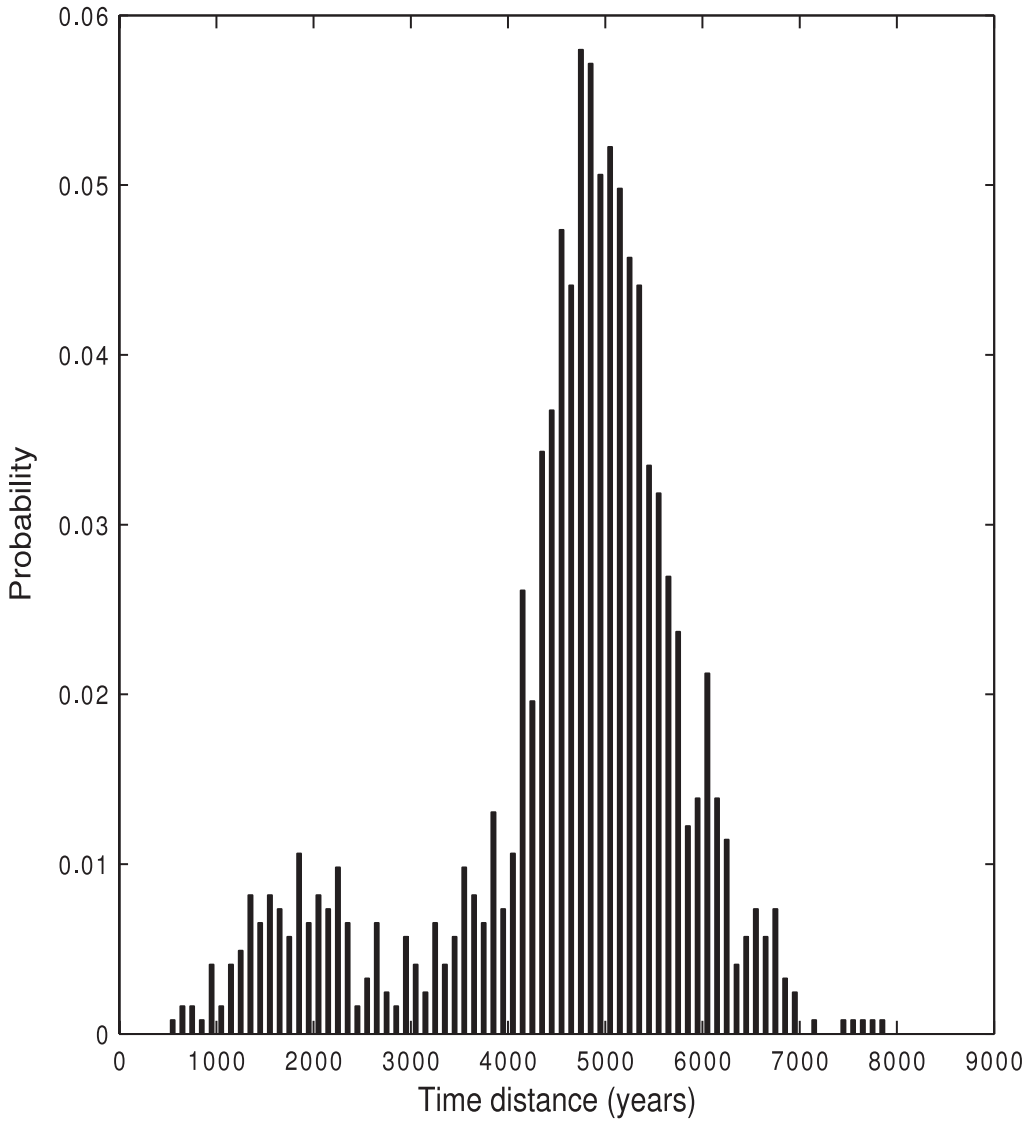

Figure 4. Distribution of distances obtained from all the 1225 pairs of languages from the Indo-European database.

realistic case time distances are reconstructed from lexical distances. In this reconstruction we assume that lexical mutations and borrowings happen 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 one pair of languages to another and also they may vary in time according to historical conditions. Therefore, the distributions in figures 4 and 5 are not exactly of the type in [9, 10] but they could be obtained from them after a random shift of all distances.

We would like to compare now our results with those published in [1] and in [2].

The tree in figure 1 is similar to the one in [1] but there are some important differences. First of all, the first separation concerns Armenian, which forms a separate branch close

J. Stat. Mech. (2008) P08012

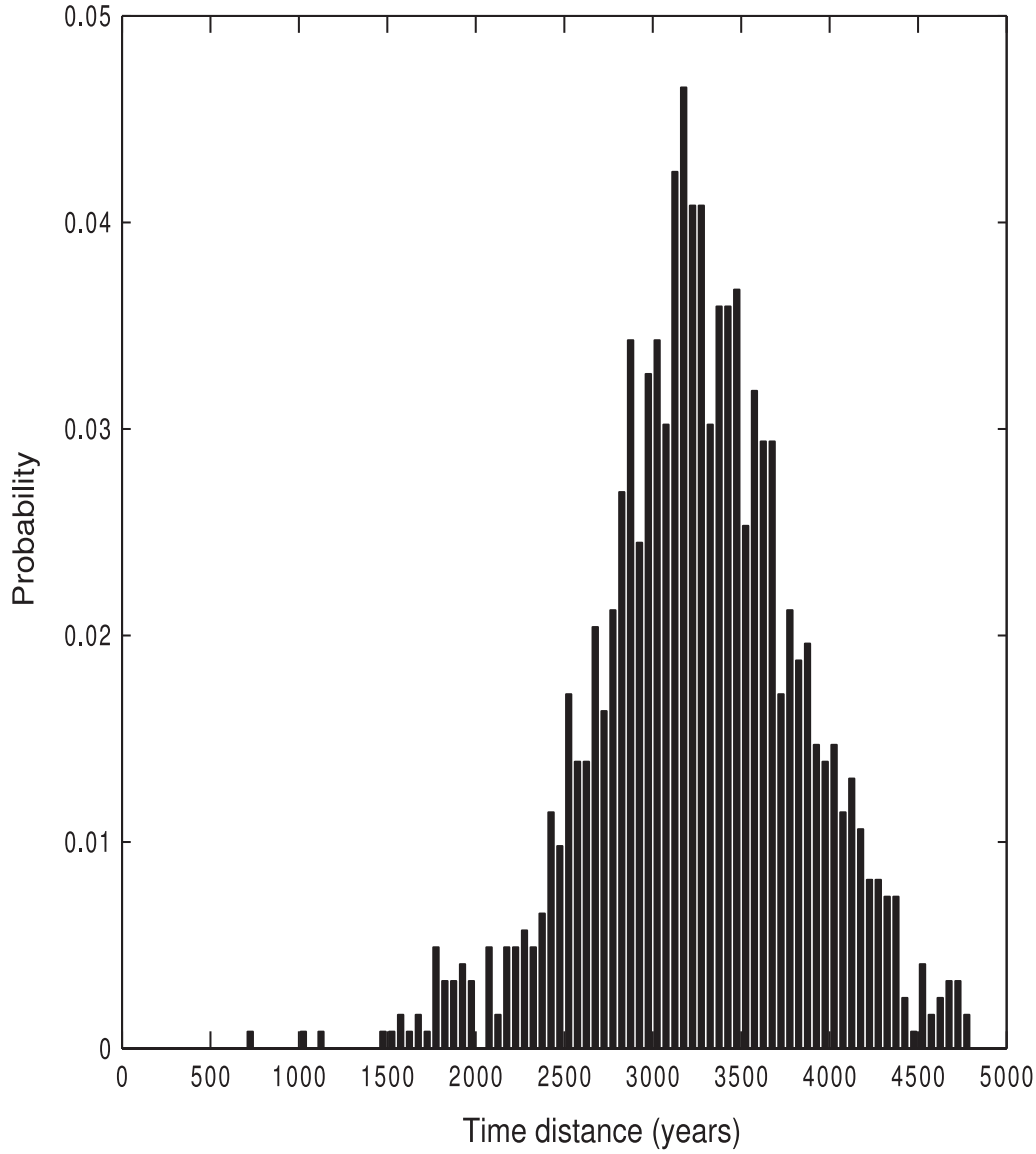

Figure 5. Distribution of distances obtained from all the 1225 pairs of languages from the Austronesian database.

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 [1], since therein the separation at the root gives rise 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 [1] 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 [1]. We will come back to this last point in section 5.

In spite of these differences, our tree seems to confirm the conclusions reported in [1] as regards 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.

Also the tree in figure 2 is similar to the one in [2] but differences here are more important. The first separation concerns Formosan (Atayal group) languages which are in the first main branch, while all the other Formosan languages (Paiwan group) are in the second main branch together with all other languages of the group. In this second main branch the first separation concerns the Oceanian languages, the second separation the Malagasy, the third all the remaining Formosan languages (Paiwan), the fourth the Philippino languages, and, finally, the fifth, the Indonesian/Sulawesi languages with two subgroups: Sulawesi and Indonesian.

The fact that the first separation concerns Formosan languages of the Atayal group seems to confirm that the Austronesian group did indeed originate in Formosa as is widely accepted by researchers. Nevertheless, the second group of Formosan languages (Paiwan) is grouped differently and closer to the Philippino and Indonesian languages. This result, if confirmed, would suggest two different waves of migration from Formosa with different origins into the Island. In this case, the early separation of the Oceanian languages would be linked with the first wave and the Philippino/Indonesian with the second. An alternative explanation would be a backward later migration toward Taiwan.

Finally, the Malagasy language is not grouped as close to Kalimantan languages as often claimed in the literature; even the closest language (Maanyan) is in that group. This fact suggests a multiple origin which we will investigate further in section 5.

5. Malagasy and Romani

In this section we discuss the position of Malagasy and Romani with respect to the other branches of their group and with respect to their closest languages. Then, we compare our results and conclusions with previous ones in the two papers [1, 2].

In our tree in figure 2 the Malagasy language is not in a cluster together with Kalimantan languages but forms an early branch by itself. Indeed, it seems from figure 2, that separation of Malagasy from Indonesian, Sulawesi, Formosan (Paiwan) and Philippino branches occurred earlier. In order to investigate this result further, we plotted, in figure 6, the average distance of Malagasy with respect to languages of various groups. It turns out that the time separation of Malagasy from Indonesian, Sulawesi, Formosan (Paiwan) and Philippino groups is almost the same and it is of about 3500 years; in contrast, the separation from Oceanic and Formosan (Atayal) groups is more ancient. This plot confirms the picture of the tree in figure 2.

Nevertheless, it turns out that the closest language to Malagasy is Maanyan which is spoken in Kalimantan. The fact that Malagasy is related to the Indonesian and Philippines languages, and more closely to the south-east Barito group of languages spoken in Kalimantan, is well known [11].

The time separation of Malagasy from Maanyan is of only 2650 years, much more recent that the 3500 years of average separation from the other languages of the four closest groups. Even more surprising, the language that is second closest is Maranao which is spoken in the Philippines, with a separation of 3000 years, while the third closest is Buginese spoken in south Sulawesi, with a separation of 3100 years. The fact that Malagasy, as expected, is very close to Maanyan but the next closest languages are not in

J. Stat. Mech. (2008) P08012

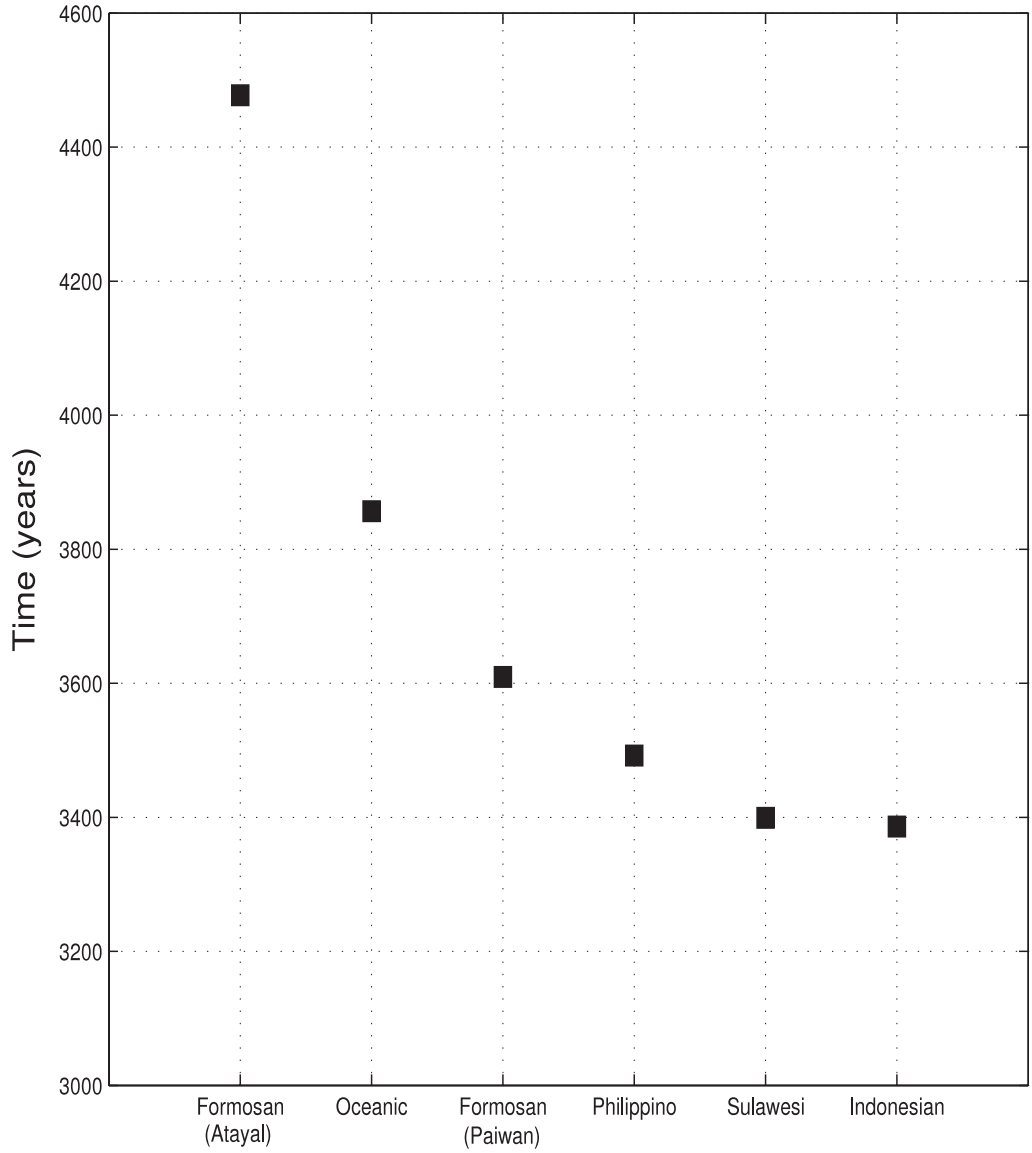

Figure 6. Time distance between Malagasy and all the other subgroups in the Austronesian group.

Kalimantan could suggest a multiple origin of the language. In fact, there are many similar loanwords in Malagasy and Malay, and there are also a number of similar loanwords in Malagasy and Javanese. The Malay and Javanese loanwords belong to all sorts of semantic domains, but Malay loanwords are particularly well represented in the domain of maritime life and navigation.

Archaeologists place the arrival of Austronesian settlers in Madagascar with their outrigger canoes in the centuries between 200 and 500 AD. Nevertheless, Malagasy mythology portrays a people, called the Vazimba, as the original inhabitants and, indeed, there are some indications that these first colonizers were part of a previous Austronesian expansion. There are also indications of successive arrivals of Austronesian immigrants during

the following centuries. This may explain why many manifestations of Malagasy culture cannot be linked up with the culture of the Dayaks of the south-east Barito area. For example, the Malagasy people use outrigger canoes, whereas south-east Barito Dayaks never do. The Malagasy migration to East Africa presupposes navigational capacities typical of many Indonesian peoples which Dayaks do not have. Some of the Malagasy musical instruments are very similar to musical instruments in Sulawesi, and some of the Malagasy cultivations (wet rice) cannot be found among Barito river inhabitants. In contrast, some funeral rites, such as the *famadihana* (second burial), are similar to those of Dayaks.

But the main problem is that it is unlikely that Maanyan speaking Dayaks realized the spectacular migrations from Kalimantan to Madagascar since they are forest dwellers with river navigation skills only. A possible explanation is that they were brought there as slaves by Malay seafarers who also took slaves from other parts of south-east Asia. If the south-east Barito speakers formed the majority in the initial group, their language could have constituted the core element of what later became Malagasy. In this way Malagasy absorbed words of the Austronesian languages of the other slaves and of the Malay seafarers. In this scenario, Malay and Javanese were spoken by the dominant class (Merina society was then divided into three classes whose individuals appear different in various respects). Later the language of the dominators was superseded but it left an important contribution to the Malagasy lexicon. This multiple origin could explain why Malagasy, which is so similar to Maanyan, is scattered in an isolated position in figure 2.

Let us discuss the case of Romani, the language of the Roma and Sinti, which is less controversial. In our tree in figure 1, it is correctly located together with Indian languages but it is not as close to Singhalese as reported in [1]. Also in this case we plotted, in figure 7, the average distance of Romani with respect to languages of various groups. It turns out that the closest group, as expected, is the Indian one, with a time separation of about 3300 years. But if we consider the closest three languages we find Nepali (2900 years), Bengali (2900 years) and Khaskura (2950 years). Nepali and Khaskura are similar and they are spoken in Nepal and northern India (the second is the language of the Gurkhas), while Bengali is spoken in north-east India and Bangladesh. We remark that these distances are not significantly different from the average distance of 3300 years for the whole Indian group and almost identical to distances for other languages spoken in northern India. This implies a geographical origin in northern India for Roma and Sinti people, according with the beliefs of the majority of researchers. In contrast, we find Romani quite far from Singhalese which in figure 1 separated 3950 years ago, a time distance which is comparable to the Romani/Italian distance (4150 years). Our results are different from those found in [1], where a close Romani/Singhalese relationship is detected.

Finally, many scholars have pointed out some similarities with Greek (and to a lesser extent with Iranian languages and Armenian) which is usually explained by a prolonged stay in Anatolia after the departure from northern India. Indeed, in figure 7 one may appreciate that, beside the Indian group, the closest language is Greek, followed by the Iranian group.

6. Discussion and conclusions

Our method correctly finds languages clusters, but while for the Indo-European group, the hierarchical organization of clusters is similar to that found by [1], for the Austronesian

J. Stat. Mech. (2008) P08012

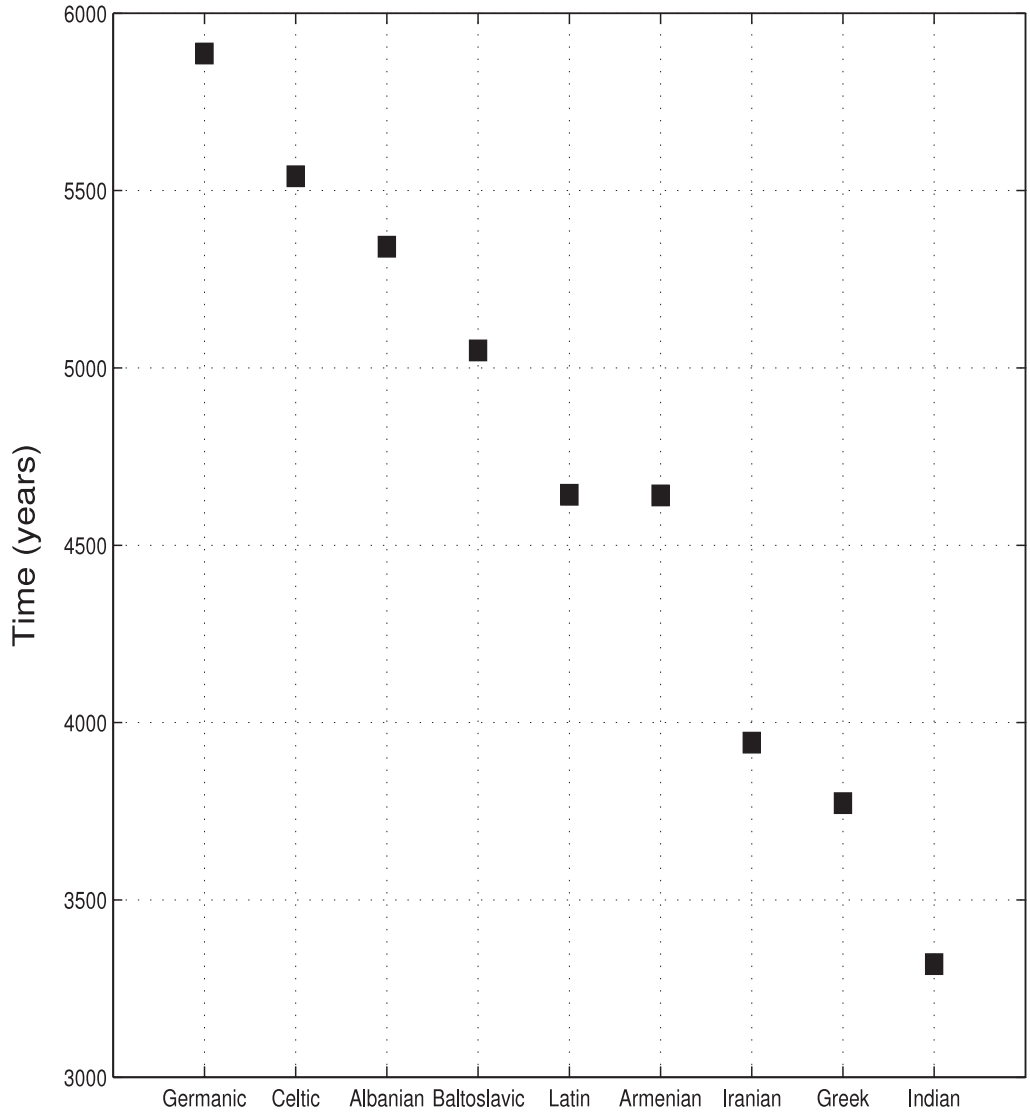

Figure 7. Time distance between Romani and all the other subgroups in the Indo-European group.

group it is quite different from that of [2] because Formosan languages split into two clusters with different positions in the tree and also because Oceanian languages separate earlier. For the Indo-European case, since the first separations concern languages geographically close to Anatolia, the Anatolian origin of the group seems to be confirmed. For the Austronesian group, in contrast, our conclusions are different, and we hypothesize two waves of migration from Formosa or a later backward migration. Furthermore, Romani seems to us a language closer to the northern India subgroup than to Singhalese, while we hypothesize a multiple origin for Malagasy.

Finally, we would like to stress again that the method used here is very simple and does not require any previous knowledge of the language origin. Also, it can be applied directly to all those language pairs for which a translation of a small group of words exists.

doi:10.1088/1742-5468/2008/08/P08012

J. Stat. Mech. (2008) P08012

Acknowledgments

We thank Lydie Irene Andriamiseza for improvements in the Malagasy database and related Austronesian Swadesh lists. Critical comments on many aspects of the paper by M Ausloos are also gratefully acknowledged. The work by FP was supported by European Commission Project E2C2 FP6-2003-NEST-Path-012975 Extreme Events: Causes and Consequences.

References

[1] Gray R D and Atkinson Q D, *Language-tree divergence times support the Anatolian theory of Indo-European origin*, 2003 *Nature* **426** 435
[2] Gray R D and Jordan F M, *Language trees support the express-train sequence of Austronesian expansion*, 2000 *Nature* **405** 1052
[3] Swadesh M, *Lexicostatistic dating of prehistoric ethnic contacts*, 1952 *Proc. Am. Phil. Soc.* **96** 452
[4] Serva M and Petroni F, *Indo-European languages tree by Levenshtein distance*, 2008 *Europhys. Lett.* **81** 68005
[5] Dyen I, Kruskal J B and Black P, 1997 *FILE IE-DATA1* (Available at http://www.ntu.edu.au/education/langs/ielex/IE-DATA1)
[6] Greenhill S J, Blust R and Gray R D, *The Austronesian basic vocabulary database*, 2003–2008 http://language.psy.auckland.ac.nz/austronesian
[7] Sneath P H A and Sokal R R, 1973 *Numerical Taxonomy* (San Francisco, CA: Freeman)
[8] Kingman J F C, *On the genealogy of large populations. Essays in statistical science*, 1982 *J. Appl. Probab.* A **19** 27
[9] Serva M, *On the genealogy of populations: trees, branches and offspring*, 2005 *J. Stat. Mech.* P07011
[10] Simon D and Derrida B, *Evolution of the most recent common ancestor of a population with no selection*, 2006 *J. Stat. Mech.* P05002
[11] Dahl O C, 1951 *Malgache et Maanyan: Une comparison linguistique* (Oslo: Egede Institett)