Lexical evolution rates derived from automated stability measures

J. Stat. Mech. (2010) P03015

(http://iopscience.iop.org/1742-5468/2010/03/P03015)

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

Lexical evolution rates derived from automated stability measures

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

[1] DIMADEFA, Facoltà di Economia, Università di Roma 'La Sapienza', I-00161 Roma, Italy
[2] Dipartimento di Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy
E-mail: fpetroni@gmail.com and serva@univaq.it

Received 7 December 2009
Accepted 6 February 2010
Published 18 March 2010

Online at stacks.iop.org/JSTAT/2010/P03015
doi:10.1088/1742-5468/2010/03/P03015

Abstract. Phylogenetic trees can be reconstructed from the matrix which contains the distances between all pairs of languages in a family. Recently, we proposed a new method which uses normalized Levenshtein distances among words with the same meaning and averages over all the items of a given list. Decisions about the number of items in the input lists for language comparison have been debated since the beginning of glottochronology. The point is that words associated with some of the meanings have a rapid lexical evolution. Therefore, a large vocabulary comparison is only apparently more accurate than a smaller one, since many of the words do not carry any useful information. In principle, one should find the optimal length of the input lists, studying the stability of the different items. In this paper we tackle the problem with an automated methodology based only on our normalized Levenshtein distance. With this approach, the program of an automated reconstruction of language relationships is completed.

Keywords: nonlinear dynamics

ArXiv ePrint: 0912.0821

Contents

1. Introduction

Glottochronology tries to estimate the time at which languages diverged, with the implicit assumption that vocabularies change at a constant average rate. The concept seems to have its roots in the work of the French explorer Dumont D'Urville. He collected comparative word lists for various languages during his voyages in the Astrolabe from 1826 to 1829 and, in his work concerning the geographical division of the Pacific [3], he introduced the concept of lexical cognates and proposed a method for measuring the degree of relation among languages. He used a core vocabulary of 115 basic terms which, impressively, contains all but three of the terms of the Swadesh 100-item list. Then, he assigned a distance from 0 to 1 to any pair of words with the same meaning and finally he was able to obtain the relationship for any pair of languages. His conclusion is famous: *La langue est partout la même.*

The method used by modern glottochronology, developed by Morris Swadesh [18] in the 1950s, measures distances from the percentage of shared cognates. Recent examples are the studies of Gray and Atkinson [5] and Gray and Jordan [6]. Cognates are words inferred to have a common historical origin, and cognacy decisions are made by trained and experienced linguists. Nevertheless, the task of counting the number of cognate words in a list is far from being trivial and results may vary for different studies. Furthermore, these decisions may imply an enormous working time.

Recently, we proposed a new automated method [14, 16] which has some advantages. The first is that it avoids subjectivity, the second is that results can be replicated by other scholars assuming that the database is the same, the third is that no specific linguistic knowledge is requested, and the last, but surely not the least, is that it allows for rapid comparison of a very large number of languages. We applied our method to the Indo-European and the Austronesian groups considering, in both cases, fifty different languages.

In our work, we defined the distance between two languages by considering a normalized Levenshtein distance between words with the same meaning and we averaged over the two hundred words contained in a 200-word list [20]. The normalization, which takes into account word length, plays a crucial role, and no sensible results would have been found without.

doi:10.1088/1742-5468/2010/03/P03015

J. Stat. Mech. (2010) P03015

At almost the same time, the above described automated method was used and developed by another large group of scholars [1, 9]. In their work, they used lists of 40 words, while we used lists of 200. Their choice was made according to a careful study of the stability of different words.

Decisions about the number of words in the input lists for language comparison has been debated since the beginning of glottochronology, Swadesh himself switched from 200-word lists to 100-word ones. The point is that a large vocabulary comparison is only apparently more accurate; in fact, many of the words do not carry any information on language similarity, and their inclusion in the lists only has the effect of increasing the error noise that may hide the wanted results. In fact, words evolve because of lexical changes, borrowings and replacement at a rate which is not the same for all of them. The speed of lexical evolution is different for different meanings and it is probably related to the frequency of use of the associated words [13]. Those meanings with a high rate of change can be useless for establishing relationships among languages. This may not hold true for young families. This point has also been tested in this work.

The idea of inferring the stability of an item from its similarity in related languages goes back a long way in the lexicostatistical literature [8, 12, 19]. In this paper we tackle this problem with an automated methodology based on the normalized Levenshtein distance. For any meaning, and any linguistic group, we are able to find a number which measures its stability (or degree of evolution speed) in a completely objective and reproducible manner. With this approach, the program of an automated reconstruction of language relationships is completed. This is different from the approach in [1, 9], since they have a combined approach, their lists are chosen according to a stability study which makes use of cognates, and then they reconstruct the language phylogeny by using Levenshtein distance.

In section 2 we define the lexical distance between words and we also sketch our method for computing the time divergence between languages. Section 3 is the core of the paper; there we define the automated stability measures of the meanings and we make some preliminary studies concerning the distribution and ranking of stability for Indo-European languages. In section 4 we study correlations and Fouldy–Robinson differences associated with lists of different lengths. We here take the decision about the meanings that should be included in the lists. Conclusions and an outlook appear in section 5.

2. Definition of distance

We define here the lexical distance between two words, which is a variant of the Levenshtein (or edit) distance. The Levenshtein distance is simply the minimum number of insertions, deletions, or substitutions of a single character needed to transform one word into the other. Our definition is taken as the edit distance divided by the number of characters in the longer of the two compared words.

More precisely, given two words α_i and β_j their distance $D(\alpha_i, \beta_j)$ is given by

$$D(\alpha_i, \beta_j) = \frac{D_l(\alpha_i, \beta_j)}{L(\alpha_i, \beta_j)} \tag{1}$$

where $D_l(\alpha_i, \beta_j)$ is the Levenshtein distance between the two words and $L(\alpha_i, \beta_j)$ is the number of characters of the longer of the two words α_i and β_j. Therefore, the distance can take any value between 0 and 1. Obviously $D(\alpha_i, \alpha_i) = 0$.

J. Stat. Mech. (2010) P03015

The normalization is an important novelty and it plays a crucial role; no sensible results can been found without it [14, 16].

We use distance between pairs of words, as defined above, to construct the lexical distances between languages. For any pair of languages, the first step is to compute the distance between words corresponding to the same meaning in the Swadesh list. Then, the lexical distance between each language pair is defined as the average of the distances between all words [14, 16]. As a result we have a number between 0 and 1 which we claim to be the lexical distance between two languages.

Assume that the number of languages is N and the list of words for any language contains M items. Any language in the group is labeled with a Greek letter (say α) and any word of that language by α_i with $1 \leq i \leq M$. Then, two words α_i and β_j in the languages α and β have the same meaning (they correspond to the same meaning) if $i = j$.

Then the distance between two languages is

$$D(\alpha, \beta) = \frac{1}{M} \sum_i D(\alpha_i, \beta_i) \tag{2}$$

where the sum goes from 1 to M. Notice that only pairs of words with the same meaning are used in this definition. This number is in the interval [0, 1]; obviously $D(\alpha, \alpha) = 0$.

The results of the analysis is an $N \times N$ upper triangular matrix whose entries are the $N(N-1)$ nontrivial lexical distances $D(\alpha, \beta)$ between 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 of computing time.

A phylogenetic tree can be constructed from the matrix of lexical distances $D(\alpha, \beta)$, but this gives only the topology of the tree whereas the absolute timescale is missing. Therefore, we perform [14, 16] a logarithmic transformation of lexical distances which is analogous to using the adjusted fundamental formula of glottochronology [17]. In this way we obtain a new $N \times N$ upper triangular matrix whose entries are the times of divergence between all pairs of languages. This matrix preserves the topology of the lexical distance matrix but it also contains the information concerning absolute timescales. Then, the phylogenetic tree can be straightforwardly constructed.

In [14, 16] we tested our method by constructing the phylogenetic trees of the Indo-European group and of the Austronesian group. In both cases we considered $N = 50$ languages. The database [20] that we used in [14, 16] is composed of $M = 200$ words for any language. The main source for the database for the Indo-European group is the file prepared by Dyen *et al* in [4]. For the Austronesian group we used as the main source the lists contained in the huge database in [7].

Criticism has been made to our proposal [11] on the basis of the fact that our reconstructed tree presents some incongruence: for example, the early separation of Armenian, which is not grouped together with Greek (which, in our tree, separates just after Armenian). However, the structure of the top of the Indo-European tree is debated and no universally accepted conclusion exists, although important results have been obtained [2, 5, 10].

In our previous work we have adopted the historically motivated choice of 200-word lists with the meanings proposed by Swadesh. Our aim, in this paper, is to establish in a objective manner the proper length and the composition of the lists. In order to reach this goal we need to separately study the stability of any meaning.

J. Stat. Mech. (2010) P03015

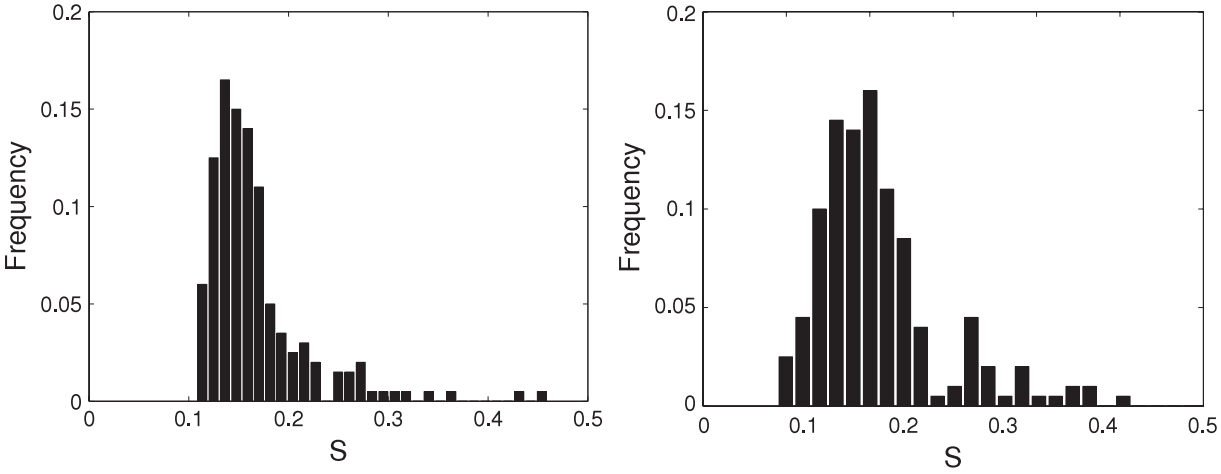

Figure 1. Stability histogram of meanings for Indo-European (left) and Austronesian (right) languages. The fat tail on the right of the histogram indicates that some items have a very large stability.

3. Stability of meanings

We take now decisions concerning stability of meanings. Our aim is to obtain an automated procedure, which avoids, also at this level, the use of cognates. For this purpose, it is necessary to obtain a measure of the typical distance of all pairs of words corresponding to a given meaning in a language family. The meaning is indicated by the label i and α_i is the corresponding word in the language α. Therefore, we define the stability as

$$S(i) = 1 - \frac{2}{N(N-1)} \sum_{\alpha>\beta} D(\alpha_i, \beta_i) \tag{3}$$

where the sum goes over all possible $N(N-1)/2$ possible language pairs α, β in the family, using the fact that $D(\alpha_i, \beta_i) = D(\beta_i, \alpha_i)$.

With this definition, $S(i)$ is inversely proportional to the average of the distances $D(\alpha_i, \beta_i)$ and takes a value between 0 and 1. The averaged distance is smaller for those words corresponding to meanings with a lower rate of lexical evolution, since they tend to remain more similar in two languages. Therefore, to a larger $S(i)$ there corresponds a greater stability.

We computed the $S(i)$ for the 200 meanings of 50 languages of the Indo-European and Austronesian families. To have a first qualitative understanding of the distribution of the $S(i)$ we plot the associated histogram in figure 1. We can see that there is a fat tail on the right of the histograms, indicating that there are some meanings with a quite large stability. This tail is very much at variance with a standard Gaussian behavior. We remark that similar plots were computed in [13] where the rates of lexical evolution are obtained by the standard glottochronology approach.

To understand better the behavior of the stability distribution, we plotted $S(i)$, in decreasing rank, for the 200 meanings in the list. In figure 2 there are reported the data concerning Indo-European and Austronesian families. At the beginning the stability drops

J. Stat. Mech. (2010) P03015

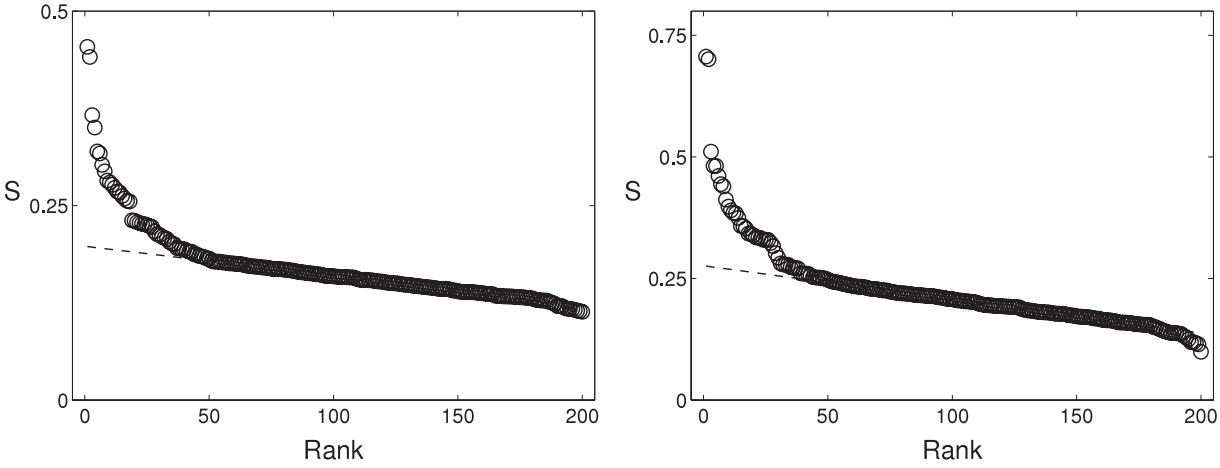

Figure 2. Stability in a decreasing ranking for the 200 meanings of the Indo-European (left) and Austronesian (right) languages. At the beginning stability has large values, but it drops rapidly, and then, between the 50th position and the 180th it decreases linearly; finally it drops again. The straight line between position 51 and position 180 underlines the initial and final deviations from the linear behavior.

rapidly; then, between the 50th position and the 180th, it decreases slowly and almost linearly with rank; finally at the end the stability drops again. We stress again that this behavior is not Gaussian—for which high and low stability parts of the curve would be symmetric. The curve is fitted by a straight line in the central part of the data, between position 51 and position 180, in order to highlight the initial and final deviations from the linear behavior.

A preliminary conclusion is that one should surely keep all the meanings with higher information, take at least some of the most stable meanings in the linear part of the curve and exclude completely those meanings with lower information. Nevertheless, at this stage it is difficult to say how many items should be maintained, since this number could be anywhere between 50 and 180.

It is necessary to carry out a deeper analysis of the stability to reach a conclusion. Indeed, we need to know the minimum number of meanings allowing for a precise computation of distances between languages and, consequently, permitting an accurate construction of the phylogenetic tree. In order to reach this goal we need a careful analysis of correlations among distances computed with the whole list and distances computed with shorter lists. It is also necessary to compare the phylogenetic trees by using a proper measure, the most natural being the Robinson–Foulds difference.

4. Correlations

As mentioned in section 3, first of all we need to evaluate the impact of shorter lists on our estimate for the distances between languages. In order to reach this goal, we compute the coefficient of correlation $c(n)$ between distances $D(\alpha, \beta)$ obtained for the whole list of 200 items and the distances $D_n(\alpha, \beta)$ obtained only for the n most stable items (obviously, $D(\alpha, \beta) = D_{200}(\alpha, \beta)$).

J. Stat. Mech. (2010) P03015

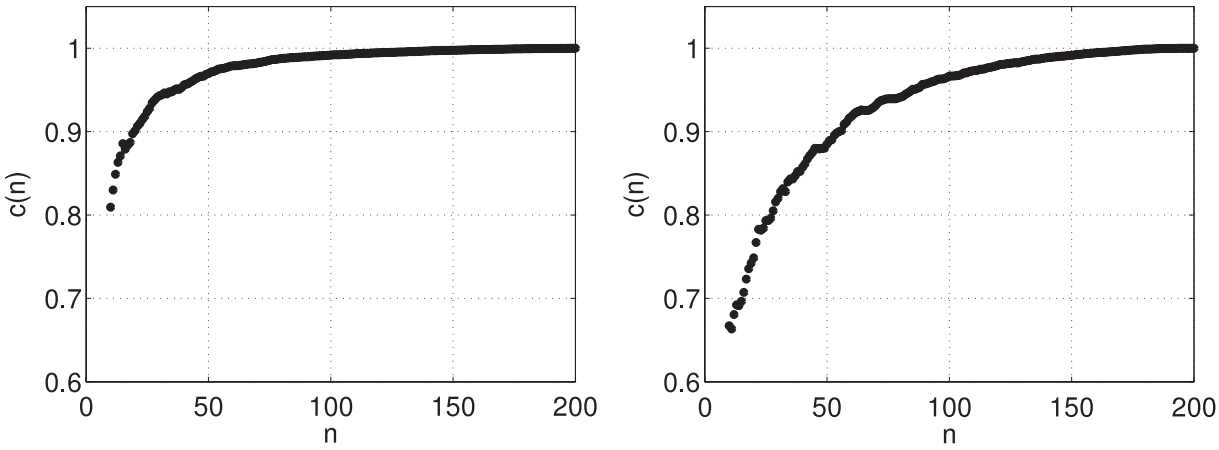

Figure 3. Correlation coefficient $c(n)$ for distances for Indo-European (left) and Austronesian (right) languages. The coefficient $c(n)$ measures the correlation between the distances estimated with n items and the distances estimated with 200 items. $c(n)$ reaches a value larger than 99% at $n = 100$.

The correlation coefficient $c(n)$ is computed in a standard way, using averages over all possible pairs of languages, and it takes the value 1 only when there is complete coincidence between $D_n(\alpha, \beta)$ and $D(\alpha, \beta)$. The correlations are plotted in figure 3 for the Indo-European and Austronesian families.

From the figure, one can observe that the correlation reaches a value larger than 99% with 100 meanings.

The problem is again that our choice for the length of the lists depends on our choice for the minimum accepted correlation coefficient. If we accept 97%, we are satisfied by lists of 50 meanings, while if we need 99%, we have to take lists of 100 meanings.

To resolve this problem we estimated the Robinson–Foulds difference [15] between the trees generated starting from $D_n(\alpha, \beta)$ and the tree generated starting from $D(\alpha, \beta)$. The RF difference, which is plotted in figure 4 for the Indo-European and Austronesian families, measures the degree of similarity between two trees. To lower values there correspond trees which are more similar.

As one can see from figure 4, the RF difference drops rapidly until $n \sim 100$. Than it remains almost constant for all values greater than $n = 100$ (the RF difference is equal to zero when $n = 200$ but this is expected since $D_{200}(\alpha, \beta) = D(\alpha, \beta)$). This result says that with 100 meanings one is able to capture all the information regarding language distance and larger lists produce the same output. In other words, the 100 meanings which have been eliminated carry small, if not vanishing, information.

The complete list of the 100 most stable terms for the Indo-European group can be found in table 1. The list is ordered by ranking, and the stability value is written correspondingly for any item.

To test whether the list of most stable words has any use when young families are taken into account we build the phylogenetic tree for the western Romance family obtained with both lists (the 200-word list and the most stable 100-word list), finding no differences in the trees. This suggest that for young families the procedure of finding the most stable words is not particularly useful.

J. Stat. Mech. (2010) P03015

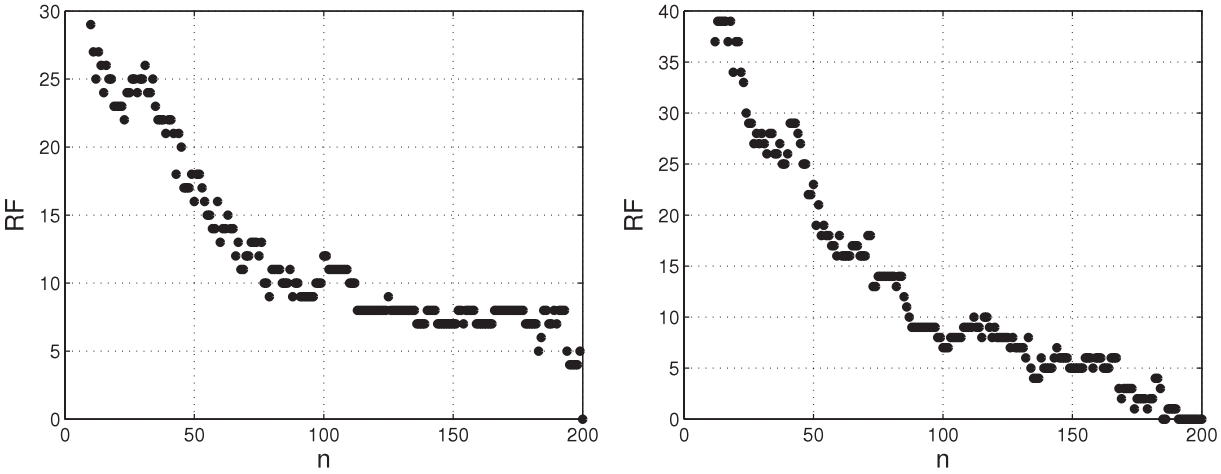

Figure 4. Robinson–Foulds difference between trees of Indo-European (left) and Austronesian (right) languages computed with lists of 200 items and lists of n items. The RF difference measures the degree of similarity between trees. More similar trees have a smaller difference. The RF difference drops rapidly until $n \sim 100$; then it remains almost constant for all greater values of n.

In conclusion, one has to consider lists with the 100 meanings with highest stability, compute the matrix of lexical distances, transform it into the matrix of divergence times and, finally, construct the tree. The elimination of the 100 items with lowest stability has the positive effect of reducing the working time necessary for an accurate check of all items and, therefore, reducing errors due to misspelling or inaccurate transliterations. Furthermore, shorter lists allow for comparison of languages whose available vocabulary is small.

5. Discussion and conclusions

In previous works [14, 16] we proposed an automated method for evaluating the distance between languages. Here we propose a method that is also automatic and gives lists of the most stable meanings. The novelty is that combining [14, 16] with the results presented here, everything can be done automatically. Stable meanings, distances, divergence times and phylogenetic trees can all be obtained by using simple objective arguments based on the normalized Levenshtein distance.

We do not claim that our combined method produces better results than the standard glottochronology approach, but they are surely comparable. The advantages of this approach can be summarized here as follows: it avoids subjectivity since all results can be replicated by other scholars assuming that the database is the same; it allows for rapid comparison of a very large number of languages; it can be used also for those language groups for which the use of cognates is very complicated or even impossible. In fact, the only work is to prepare the lists, while all the remaining work is carried out by a computer program.

We want to stress that we do not make use of any historical information, such as 'cognacy', as regards the languages that we compare. Cognates are words inferred to have

J. Stat. Mech. (2010) P03015

Table 1. List of the 100 most stable meanings according to the $S(i)$ measure described in the text.

Word	$S(i)$	Word	$S(i)$	Word	$S(i)$	Word	$S(i)$
You	0.453 95	Three	0.441 02	Mother	0.366 27	Not	0.350 33
New	0.319 61	Nose	0.316 9	Four	0.302 26	Night	0.294 03
Two	0.282 14	Name	0.279 62	Tooth	0.276 77	Star	0.272 69
Salt	0.267 92	Day	0.266 95	Grass	0.262 31	Sea	0.259 06
Die	0.256 02	Sun	0.255 35	One	0.230 93	Feather	0.230 55
Give	0.228 64	Sit	0.227 57	Stand	0.226 44	Meat	0.226 1
Long	0.224 91	Five	0.223 53	Hand	0.222 61	Short	0.216 76
Father	0.213 19	Smoke	0.212 13	Far	0.209 98	Worm	0.208 46
Dry	0.207	Scratch	0.203 43	Person	0.201 29	When	0.200 11
Wind	0.195 35	Snake	0.194 85	Sing	0.194 34	Stone	0.193 69
Suck	0.191 96	Mouth	0.190 67	Dig	0.190 52	Live	0.187 16
Root	0.187 15	Hair	0.185 22	Smooth	0.184 57	Water	0.183 78
Tongue	0.181 94	Animal	0.181 9	Year	0.178 92	Red	0.178 15
Man	0.178 01	Tie	0.177 89	Snow	0.176 97	Sew	0.176 86
There	0.176 57	Breathe	0.175 78	Flower	0.175 66	Mountain	0.175 45
Fruit	0.175 08	Bark	0.175 02	Sand	0.174 43	Leaf	0.173 9
Warm	0.172 83	Green	0.172 69	Liver	0.172 05	Hunt	0.171 68
Sky	0.171 56	Know	0.171 17	Bone	0.170 56	Spit	0.170 36
Heart	0.170 23	Pull	0.169 84	Right	0.168 9	We	0.168 58
Husband	0.168 53	Foot	0.168 3	Drink	0.168 28	See	0.167 64
Lie	0.167 63	Fish	0.166 93	Woman	0.166 56	Louse	0.166 24
Straight	0.165 34	Yellow	0.164 87	Sleep	0.164 3	Black	0.164 08
Who	0.163 51	Seed	0.162 99	Wing	0.162 88	Cut	0.162 45
Count	0.161 73	Thin	0.161 56	Sharp	0.161 1	Float	0.160 28
Fall	0.159 68	Earth	0.159 65	Kill	0.159 26	Burn	0.159 18

a common historical origin. For instance, the Spanish word *leche* and the Greek word *gala* are cognates. Also the English *wheel* and the Hindi word *cakra* are cognates. These two identifications are possible because of historical records. In our work we only compare words with the same meaning quantifying their difference, without the need for historical input. Also a preliminary separation of languages into families, like Indo-European and Austronesian, is not strictly necessary.

We would like to mention that recently, together with other scholars [2], we have applied the method described here as a starting point for a deeper analysis of relationships among languages. The point is that a tree is only an approximation, which, obviously, skips more complex phenomena such as horizontal transfer. These phenomena are reflected in the matrix of distances as deviations from the ultrametric structure. It seems that the approach in [2] allows for some more accurate understanding of some important topics, such as migration patterns and homeland locations of families of languages.

Acknowledgments

We warmly thank Søren Wichmann for helpful discussion. We also thank Philippe Blanchard, Armando Neves, Luce Prignano and Dimitri Volchenkov for critical comments on many aspects of the paper. We are indebted to S J Greenhill, R Blust and

R D Gray for the authorization to use *The Austronesian Basic Vocabulary Database*, http://language.psy.auckland.ac.nz/austronesian which we consulted in January 2008.

References

[1] Bakker D, Brown C H, Brown P, Egorov D, Grant A, Holman E W, Mailhammer R, Muller A, Velupillai V and Wichmann S, *Adding typology to lexicostatistics: a combined approach to language classification*, 2010 *Linguist. Typol.* **13** 167–79
[2] Blanchard Ph, Petroni F, Serva M and Volchenkov D, *Geometric representations of language taxonomies*, 2010 *J. Comput. Speech Lang.* accepted for publication
[3] D'Urville D, *Sur les îles du Grand Océan*, 1832 *Bull. Soc. Géogr.* **17** 1
[4] Dyen I, Kruskal J B and Black P, *FILE IE-DATA*, 1997 available at http://www.wordgumbo.com/ie/cmp/iedata.txt
[5] Gray R D and Atkinson Q D, *Language-tree divergence times support the Anatolian theory of Indo-European origin*, 2003 *Nature* **426** 435
[6] Gray R D and Jordan F M, *Language trees support the express-train sequence of Austronesian expansion*, 2000 *Nature* **405** 1052
[7] Greenhill S J, Blust R and Gray R D, *The Austronesian basic vocabulary database*, 2003–2008 http://language.psy.auckland.ac.nz/austronesian
[8] Kroeber A, *Yokuts dialect survey*, 1963 *Anthropol. Rec.* **11** 177
[9] Holman E W, Wichmann S, Brown C H, Velupillai V, Muller A and Bakker D, *Explorations in automated lexicostatistics*, 2008 *Folia Linguist.* **42.2** 331
[10] Nakhleh L, Ringe D and Warnow T, *Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages*, 2005 *Language* **81** 81
[11] Nichols J and Warnow T, *Tutorial on computational linguistic phylogeny*, 2008 *Lang. Linguist. Compass* **2** 760
[12] Oswalt R, *Towards the construction of a standard lexicostatistic list*, 1971 *Anthropol. Linguist.* **13** 421
[13] Pagel M, Atkinson Q D and Meadel A, *Frequency of word-use predicts rates of lexical evolution throughout Indo-European history*, 2007 *Nature* **449** 717
[14] Petroni F and Serva M, *Languages distance and tree reconstruction*, 2008 *J. Stat. Mech.* P08012
[15] Robinson D F and Foulds L R, *Comparison of phylogenetic trees*, 1981 *Math. Biosci.* **53** 131
[16] Serva M and Petroni F, *Indo-European languages tree by Levenshtein distance*, 2008 *Europhys. Lett.* **81** 68005
[17] Starostin S, *Comparative-historical linguistics and Lexicostatistics*, 1999 *Historical Linguistics and Lexicostatistics* (Melbourne: Association for the History of Language) pp 3–50
[18] Swadesh M, *Lexicostatistic dating of prehistoric ethnic contacts*, 1952 *Proc. Am. Phil. Soc.* **96** 452
[19] Thomas D D, *Basic vocabulary in some Mon-Khmer languages*, 1960 *Anthropol. Linguist.* **2.3** 7
[20] The database, modified by the authors, is available at the following web address: http://univaq.it/∼serva/languages/languages.html

J. Stat. Mech. (2010) P03015