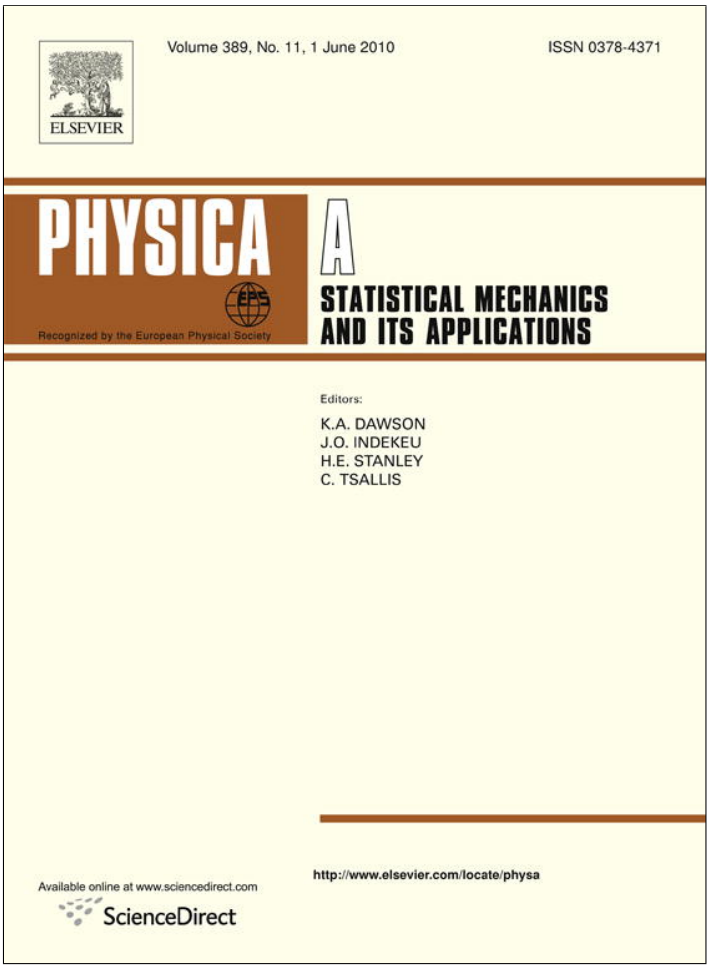
Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Physica A 389 (2010) 2280–2283

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Measures of lexical distance between languages

Filippo Petroni [a,*], Maurizio Serva [b]

[a] DIMADEFA, Facoltà di Economia, Università di Roma "La Sapienza", I-00161 Roma, Italy
[b] Dipartimento di Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy

ARTICLE INFO

Article history:
Received 9 December 2009
Received in revised form 30 January 2010
Available online 11 February 2010

Keywords:
Historical linguistics
Phylogenetics
Levenshtein distance

ABSTRACT

The idea of measuring distance between languages seems to have its roots in the work of the French explorer Dumont D'Urville (1832) [13]. He collected comparative word lists for various languages during his voyages aboard the Astrolabe from 1826 to 1829 and, in his work concerning the geographical division of the Pacific, he proposed a method for measuring the degree of relation among languages. The method used by modern glottochronology, developed by Morris Swadesh in the 1950s, measures distances from the percentage of shared cognates, which are words with a common historical origin. Recently, we proposed a new automated method which uses the normalized Levenshtein distances among words with the same meaning and averages on the words contained in a list. Recently another group of scholars, Bakker et al. (2009) [8] and Holman et al. (2008) [9], proposed a refined version of our definition including a second normalization. In this paper we compare the information content of our definition with the refined version in order to decide which of the two can be applied with greater success to resolve relationships among languages.

© 2010 Elsevier B.V. All rights reserved.

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 idea is to consider the percentage of shared cognates in order to compute the distance between pairs of languages [1]. These lexical distances are assumed to be, on average, logarithmically proportional to divergence times. 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 [2] and Gray and Jordan [3].

We recently proposed an automated method which uses Levenshtein distances among words in a list [4,5] (another automated method used to compare dialects already exists but uses a different normalization of the Levenshtein distance [6] based on the length of the alignment). To be precise, we defined the lexical distance between two languages by considering a normalized Levenshtein distance among words with the same meaning and averaging on all the words contained in a Swadesh list. The normalization is extremely important and no reasonable results can be found without it. Then, we transformed the lexical distances into separation times. This goal was reached using a logarithmic rule which is the analogue of the adjusted fundamental formula of glottochronology [7]. Finally, the phylogenetic tree could be straightforwardly constructed.

In Refs. [4,5] we tested our method by constructing the phylogenetic trees of the Indo-European and the Austronesian groups.

* Corresponding author.
E-mail addresses: fpetroni@gmail.com (F. Petroni), serva@univaq.it (M. Serva).

0378-4371/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2010.02.004

Author's personal copy

At almost the same time, the above described automated method was used and developed by another large group of scholars [8,9]. They placed the method at the core of an ambitious project, the ASJP (The Automated Similarity Judgment Program). In their work they proposed a refinement of our definition including a second normalization in the definition of lexical distance.

The goal of this paper is to compare the information contents of the two definitions in order to decide which of the two can be applied with greater success to resolve relationships among languages.

Before tackling this problem we sketch our definition of lexical distance and the modification proposed in Refs. [8,9] which is a refinement including a second normalization. Then we compare the information content of the two definitions and give our conclusion.

2. Lexical distance

Our definition of lexical distance between two words is a variant of the Levenshtein distance which 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 Levenshtein distance divided by the number of characters of the longer of the two compared words. More precisely, given two words α_i and β_j their lexical 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 $D(\alpha_i, \alpha_i) = 0$.

The normalization is an important novelty and it plays a crucial role; no sensible results can be found without it [4,5].

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 [4,5]. 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 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 result of the analysis is an $N \times N$ upper triangular matrix whose entries are the $N(N-1)$ non-trivial 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 could be constructed from the matrix of lexical distances $D(\alpha, \beta)$, but this would only give the topology of the tree and the absolute time scale would be missing. Therefore, we perform [4,5] a logarithmic transformation of lexical distances which is the analogous to using the adjusted fundamental formula of glottochronology [7]. 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 time scales. Then, the phylogenetic tree can be straightforwardly constructed.

In Refs. [4,5] 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 [10] that we used in Refs. [4,5] is composed by $M = 200$ words for any language. The main source for the database for the Indo-European group is the file prepared by Dyen et al. in Ref. [11]. For the Austronesian group we used as the main source the lists contained in the huge database in Ref. [12].

3. A second normalization

A further modification has been proposed by Refs. [8,9] in order to avoid possible similarities which could arise from accidental relative phonological similarities of languages.

Let us first define the *global distance* between languages α and β as

$$\Gamma(\alpha, \beta) = \frac{1}{M(M-1)} \sum_{i \neq j} D(\alpha_i, \beta_j) \tag{3}$$

where the sum is over all $M(M-1)$ pairs of words corresponding to different meanings in the two lists (M^2 is the total number of pairs and M is the number of pairs with the same meaning).

Author's personal copy

This quantity measures a distance of the vocabulary of the two languages, without comparing words with the same meaning. In other words, it only accounts for general similarities in the frequency and ordering of characters. The point is that, at this stage, we don't know whether $\Gamma(\alpha, \beta)$ carries information or only depends on accidental similarities.

Assuming the point of view of Refs. [8,9], it is reasonable to define a bi-normalized lexical distance as follows:

$$D_s(\alpha, \beta) = \frac{D(\alpha, \beta)}{\Gamma(\alpha, \beta)}. \tag{4}$$

Notice that while by definition $D(\alpha, \alpha) = D_s(\alpha, \alpha) = 0$, in all real cases $\Gamma(\alpha, \alpha) \neq 0$.

This second normalization should cancel the effects of accidental phonological similarities between the two languages. The idea was to avoid a situation where unrelated languages that happen to have similar sound structures (e.g., Finnish and Japanese) for that reason alone get classified together. It was assumed that eliminating Γ would have a positive effect on the classification of sets of languages that included unrelated ones and would have no appreciable effect on the classification of languages that are related.

We think that the idea of the proposed second normalization turns out to be correct only if $\Gamma(\alpha, \beta)$ is uncorrelated with the lexical distance between languages α and β. In this case, in fact, it has vanishing information concerning their relationship. In contrast, if it is positively correlated with the distance between the two languages, one can conclude that it may contain some information that could be usefully exploited. Nevertheless, the information contained in it could also be useless in establishing phylogenetic relations. This point will be further discussed in the conclusions.

4. Comparison of different definitions

In order to decide which definition is better to use, $D(\alpha, \beta)$ or $D_s(\alpha, \beta)$, we have to see whether $\Gamma(\alpha, \beta)$ is positively correlated with these distances. If it is not, we will decide to use $D_s(\alpha, \beta)$ since we eliminate errors due to accidental similarities between vocabularies. In contrast, if it is positively correlated, we would conclude that $\Gamma(\alpha, \beta)$ carries some positive information about the degree of similarity of the two languages. In this second case, two languages will be, on average, closer for smaller $\Gamma(\alpha, \beta)$, and we would decide to use $D(\alpha, \beta)$ since it incorporates the information contained in $\Gamma(\alpha, \beta)$.

In order to compute the correlation between distance and $\Gamma(\alpha, \beta)$ we proceed as follows: first we define for a generic function $f(\alpha, \beta)$ the average $\langle f \rangle$ over all possible values of α and β as follows:

$$\langle f \rangle = \frac{1}{N^2} \sum_{\alpha,\beta} f(\alpha, \beta) \tag{5}$$

which is the average value of the function $f(\alpha, \beta)$ in a linguistic group. Then, we define the correlation between $D(\alpha, \beta)$ and $\Gamma(\alpha, \beta)$ in a standard way as

$$C(\Gamma, D) = \frac{\langle (\Gamma - \langle \Gamma \rangle)(D - \langle D \rangle) \rangle}{(\langle (\Gamma - \langle \Gamma \rangle)^2 \rangle \langle (D - \langle D \rangle)^2 \rangle)^{\frac{1}{2}}}. \tag{6}$$

The result is that the correlation in the Indo-European group is 0.59173 while in the Austronesian group it is 0.46032. In both cases it is a quite high positive value (correlation may take any value between -1 and 1) and we conclude that eventual vocabulary similarities accounted for by $\Gamma(\alpha, \beta)$ carry information and are not at all accidental. The weak point is that we have checked the correlations against $D(\alpha, \beta)$ which, at least from the point of view of the proponents of the second normalization, linearly incorporates $\Gamma(\alpha, \beta)$ since $D(\alpha, \beta) = \Gamma(\alpha, \beta) D_s(\alpha, \beta)$.

From this point of view our result is not so astonishing. Nevertheless, we can also compute the correlation between the bi-normalized distance $D_s(\alpha, \beta)$ and $\Gamma(\alpha, \beta)$. The definition is the same as (6) with D_s substituting for D. We obtain that the correlation $C(\Gamma, D_s)$ in the Indo-European group is 0.54713 while in the Austronesian group it is 0.40169. These two numbers, although slightly smaller than the previous ones, are still quite high and confirm that $\Gamma(\alpha, \beta)$ contains positive information. In other words, closer languages, both in the sense of a smaller $D(\alpha, \beta)$ and a smaller $D_s(\alpha, \beta)$, will on average have smaller $\Gamma(\alpha, \beta)$.

We remark that the same correlation coefficients, both for D and for D_s, come out if the average (5) is computed neglecting the pairs where the same Greek index is repeated.

In order to complete our analysis we plot, just for the Austronesian languages group, $\Gamma(\alpha, \beta)$ as a function of $D(\alpha, \beta)$ (Fig. 1, left) and as a function of $D_s(\alpha, \beta)$ (Fig. 1, right). Any point in the figures represents a pair of languages. In both cases we perceive the positive correlation which is evidenced by the best linear fits.

We remark that the points which lie on the vertical axes at the distance 0 value correspond, in both figures, to pairs for which the same language is compared. For these points the $D(\alpha, \alpha) = D_s(\alpha, \alpha)$ are all equal to 0 while the $\Gamma(\alpha, \alpha)$ are positive. It is easy to see that the self-distances accounted by the $\Gamma(\alpha, \alpha)$, which compare words with different meaning in the same language, are, on average, smaller than the $\Gamma(\alpha, \beta)$ which compare words with different meaning in two different languages. This fact confirms that the information carried by $\Gamma(\alpha, \beta)$ is positive.

In other words, more closely related languages not only have more similar words corresponding to the same meaning, but also have more similar general occurrence and ordering of characters in words.

Author's personal copy

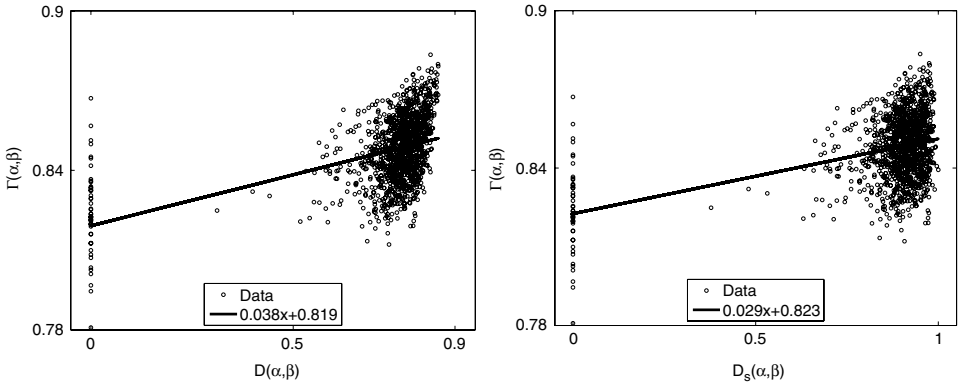


Fig. 1. Global distance $\Gamma(\alpha, \beta)$ as a function of lexical distance $D(\alpha, \beta)$ (left) and as a function of bi-normalized distance $D_s(\alpha, \beta)$ (right) for Austronesian languages. The positive correlations are evidenced by the best linear fits. The points which lie on the vertical axes at the distance 0 value correspond to pairs for which the same language is compared. For these points $D(\alpha, \alpha) = D_s(\alpha, \alpha)$ while $\Gamma(\alpha, \alpha) \neq 0$.

5. Conclusions

In this work we have analyzed two different possibilities for the definition of automated language distance. More precisely, starting from a Levenshtein distance, we have analyzed two possible normalizations. The comparison between them is only made by using statistical arguments.

Our analysis clearly shows that more closely related languages have smaller global distance. This means that not only do they have more similar words for the same meaning, but also they have more similar general occurrence and ordering of characters in words.

We would like to conclude that it is preferable to use the single-normalization definition of distance $D(\alpha, \beta)$, since otherwise a part of the information about affinities of languages is lost. Nevertheless, this conclusion should be empirically supported by comparing the distance matrices corresponding to the two definitions with matrices produced by experts.

At this stage, we have clearly shown that global distances contain some information about language relationships. Nevertheless, we are not completely sure that this information is really relevant when constructing phylogenetic trees and this remains an open question.

Acknowledgements

We warmly thank Søren Wichmann for helpful discussion. We also thank Philippe Blanchard, 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] M. Swadesh, Lexicostatistic dating of prehistoric ethnic contacts, Proceedings American Philosophical Society 96 (1952) 452–463.
[2] R.D. Gray, Q.D. Atkinson, Language-tree divergence times support the Anatolian theory of Indo-European origin, Nature 426 (2003) 435–439.
[3] R.D. Gray, F.M Jordan, Language trees support the express-train sequence of Austronesian expansion, Nature 405 (2000) 1052–1055.
[4] M. Serva, F. Petroni, Indo-European languages tree by Levenshtein distance, EuroPhysics Letters 81 (2008) 68005.
[5] F. Petroni, M. Serva, Languages distance and tree reconstruction, Journal of Statistical Mechanics: Theory and experiment (2008) P08012.
[6] Wilbert Heeringa, Measuring dialect pronunciation differences using Levenshtein distance, Ph.D. Dissertation, 2004. http://www.let.rug.nl/~heeringa/dialectology/thesis/thesis00.pdf.
[7] S. Starostin, Comparative-historical linguistics and Lexicostatistics, in: Historical Linguistics and Lexicostatistics, Association for the History of Language, Melbourne, 1999, pp. 3–50.
[8] D. Bakker, C.H. Brown, P. Brown, D. Egorov, A. Grant, E.W. Holman, R. Mailhammer, A. Müller, V. Velupillai, S. Wichmann, Adding typology to lexicostatistics: A combined approach to language classification, Linguistic Typology 13 (2009) 167–179.
[9] E.W. Holman, S. Wichmann, C.H. Brown, V. Velupillai, A. Muller, D. Bakker, Explorations in automated lexicostatistics, Folia Linguistica 42.2 (2008) 331–354.
[10] The database, 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.
[11] I. Dyen, J.B. Kruskal, P. Black, FILE IE-DATA1, 1997. Available at http://www.ntu.edu.au/education/langs/ielex/IE-DATA1.
[12] S.J. Greenhill, R. Blust, R.D. Gray, The Austronesian Basic Vocabulary Database, 2003–2008. http://language.psy.auckland.ac.nz/austronesian.
[13] D. D'Urville, Sur les îles du Grand Océan, Bulletin de la Société de Goégraphie 17 (1832) 1–21.