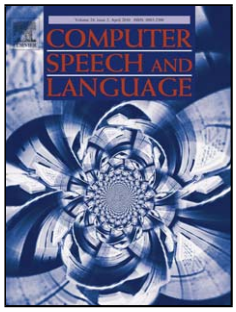
Geometric Representations of Language Taxonomies

Ph. Blanchard, F. Petroni, M. Serva, D. Volchenkov



Title: Geometric Representations of Language Taxonomies

Authors: Ph. Blanchard, F. Petroni, M. Serva, D. Volchenkov

PII: S0885-2308(10)00040-9
DOI: doi:10.1016/j.csl.2010.05.003
Reference: YCSLA 456

To appear in:

Received date: 9-10-2009
Revised date: 11-3-2010
Accepted date: 7-5-2010

Please cite this article as: Blanchard, Ph., Petroni, F., Serva, M., Volchenkov, D., Geometric Representations of Language Taxonomies, *Computer Speech & Language* (2008), doi:10.1016/j.csl.2010.05.003

Geometric Representations of Language Taxonomies

Ph. Blanchard[1]*, F. Petroni[2]†, M. Serva[3]‡, D. Volchenkov[4]§

March 11, 2010

[1] *Bielefeld University, Postfach 100131, D-33501, Bielefeld, Germany*

[2] *DIMADEFAS, Facoltà di Economia, Università di Roma "La Sapienza", Via del Castro Laurenziano 9, 00161 Roma, Italy*

[3] *Dipartimento di Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy*

[4] *Center of Excellence Cognitive Interaction Technology, Universität Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany*

Keywords: Language taxonomy, lexicostatistic data analysis, Indo-European and Polynesian origins.

Abstract

A Markov chain analysis of a network generated by the matrix of lexical distances allows for representing complex relationships between different languages in a language family geometrically, in terms of distances and angles. The fully automated method for construction of

*E-Mail: *blanchard@physik.uni-bielefeld.de*
†E-Mail: *fpetroni@gmail.com*
‡E-Mail: *serva@univaq.it*
§E-Mail:*volchenk@physik.uni-bielefeld.de*

language taxonomy is tested on a sample of fifty languages of the Indo-European language group and applied to a sample of fifty languages of the Austronesian language group. The Anatolian and Kurgan hypotheses of the Indo-European origin and the 'express train' model of the Polynesian origin are thoroughly discussed.

1 Introduction

Changes in languages go on constantly affecting words through various innovations and borrowings [1]. Although tree diagrams have become ubiquitous in representations of language taxonomies, they obviously fail to reveal full complexity of language affinity characterized by many phonetic, morphophonemic, lexical, and grammatical isoglosses; not least because of the fact that the simple relation of *ancestry* basic for a branching family tree structure cannot grasp complex social, cultural and political factors molding the extreme historical language contacts [2]. As a result, many evolutionary trees conflict with each other and with the traditionally accepted family arborescence [1]; the languages known as isolates cannot be reliably classified into any branch with other living languages [3]; the tree-reconstruction phylogenetic methods applied to the language families that do not develop by binary splitting lead to deceptive conclusions [4].

Virtually all authors using the phylogenetic analysis on language data agree upon that a *network*, or a *web* rather than trees can provide a more appropriate representation for an essentially multidimensional phylogenetic signal [5]. Networks have already appeared in phylogenetic analysis [6]-[12] either as a number of additional edges in the usual phylogenetic trees representing contacts and combined interactions between the individual languages and language groups, or as the considerable reticulation in a central part of

the tree-like graphs representing a conflict between the different splits that are produced in the data analysis. However, the more comprehensive the graphical model is, the less clear are its visual apprehension and interpretation [1].

In the present paper, we show how the relationships between different languages in the language family can be represented geometrically, in terms of distances and angles, as in Euclidean geometry of everyday intuition. Our method is fully automated and based on the statistical analysis of orthographic realizations of the meanings of Swadesh vocabulary containing 200 words essentially resistant to changes. First, we have tested our method for the Indo-European language family by construction of language taxonomy for the fifty major languages spoken in Europe, on the Iranian plateau, and on the Indian subcontinent selected among about 450 languages and dialects of the whole family [13]. Second, we have investigated the Austronesian phylogeny considered again over 50 languages chosen among those 1,200 spoken by people in Indonesia, the Philippines, Madagascar, the central and southern Pacific island groups (except most of New Guinea), and parts of mainland Southeast Asia and the island of Taiwan.

2 Applying phylogenetic methods to language taxonomies

Applying phylogenetic methods to language taxonomies is a process containing a series of discrete stages [1, 2], each one requires the application of techniques developed in different disciplines. In the first, *encoding* stage, the relations between languages is expressed in a numerical form suitable for further analysis. Various lexicostatistic techniques have been used in this stage so far (see [1], for a review). As a result each language is character-

ized by a vector (string), with components indicating the presence/absence of some features, traits, and other linguistic variables, readily converted into a matrix of lexical distances quantifying the perceptible affinity of languages in the group.

The numerical data containing the phylogenetic signal obtained in the first stage of the process lack a standard metric that makes a direct comparative analysis of the linguistic data impossible. Therefore, in the second stage, the various agglomerative clustering techniques are implemented in order to get the simplified *representations* of the data set. For example, the unweighted pair group method with arithmetic mean (UPGMA) is used in glottochronology to produce a tree from the distance matrix [14]. The neighbor joining (NJ) [15] and their variations are widely used for tree-like representations of language phylogeny. Since the phylogenetic signal is virtually multidimensional, trees and networks come at the cost of losing information. Eventually, in the *interpretation* stage, the linguistic meanings of the identified components have to be assessed. The last step is by no means trivial since, in the course of analysis, the initial linguistic features encoded in the data set appear to be strongly entangled due to the multiple transformations of coordinate systems and the phylogenetic signal may become unclear due to dramatic dimensionality reduction in the data set.

The phylogenetic methodology described above has been thoroughly criticized by linguists in each of its stages [2]. It is obvious that the direct use of techniques initially developed in genetics and palaeontology to language taxonomies is inappropriate, as the nature of interactions between the different languages in a language family and, say, between the genes in a genome is strikingly different.

Below, we present a fully automated method for building genetic language taxonomies that in many respects seems to be more relevant for the analysis of language data sets. Novelty of our approach is both in the *encoding* stage

and in the *representation* one implying some novelty in the *interpretation* stage.

3 The data set we have used

The data set [16] we have used in order to construct the language taxonomy is composed by 50 languages of the Indo-European group (IE) and 50 languages of the Austronesian group (AU). To minimize the effect of bias between orthographic and phonetic realizations of meanings, a short list of 200 words which are known to change at a slow rate are used, rather than a complete dictionary. The main source for the database for the IE group was the file prepared by Dyen *et al* [17]. This database contains the Swadesh list of vocabulary with basic 200 meanings which seem maximally resistant to change, including borrowing [18], for 96 languages. The words are given there without diacritics and adopted for using classic linguistic comparative methods to extract sets of 'cognates' – words that can be related by consistent *sound* changes. Some words are missing in [17] but for our choice of 50 languages we have filled most of the gaps and corrected some errors by finding the words from Swadesh lists and from dictionaries freely available on the web.

For the AU group, the huge database [19] has been used under the authors' permission that we acknowledge. The AU database is adopted to reconstruct systematic *sound correspondences* between the languages in order to uncover historically related 'cognate' forms and is under the permanent cleaning and development, with the assistance of linguistic experts correcting mistakes and improving the cognacy judgments. The lists in [19] contain more than 200 meanings which do not completely coincide with those in the original Swadesh list. For our choice of fifty AU languages we have retained only those words which are included in the both data sets of [17] and of

the original vocabulary [17, 22]. The resulting list has still many gaps due to missing words in the data set [19] and because of the incomplete overlap between the list of [19] and the original Swadesh list [17, 22]. We have filled some of the gaps by finding the words from Swadesh's lists available on the web and by direct knowledge of the Malagasy language (by *M.S.*).

We used the English alphabet (26 characters plus *space*) in our work to make the language data suitable for numerical processing. Those languages written in the different alphabets (i.e. Greek, etc.) were already transliterated into English in [17]. In [19], many letter-diacritic combinations are used which we have replaced by the underlying letters reducing again the set of characters to the standard English alphabet. Interestingly, the abolition of all diacritics favoring a "simple" alphabet allowed us to obtain a reasonable result. The database modified by the Authors is available at [16]. Readers are welcome to modify, correct and add words to the database.

4 The relations among languages encoded in the matrix of lexical distances

Complex relations between languages may be expressed in a numerical form with respect to many different features [1]. In traditional glottochronology [20], the percentage of significant words replaced while languages diverged from a common ancestor is counted. The concept of *cognates*, the words inherited from the ancestor language, as proved by regular sound correspondences, was introduced in the early work [21] of D. d'Urville about the geographical division of the Pacific. The method used by modern glottochronology, developed by M. Swadesh in the 1950s, measures distances from the percentage of shared cognates [22]. Constructing ancestral forms of words requires trained and experienced linguists; it is very time consuming

and cannot be automated. Statistical models used in language phylogeny (see for example [3, 8, 23, 24, 25]) describe how a set of characters may randomly evolve within a family of languages provided the relevant substitution, replacement, or confusion probabilities are taken on. Usually, statistical models have been exploited within the tree-paradigm of the language data representations. Linguists have objected to a tacit assumption that real language data can be amenable to representation as opposition between two or more 'discrete states', all equally different from each other, related by means of some 'transition probabilities' [2].

The standard Levenshtein (edit) distance accounting for the minimal number of insertions, deletions, or substitutions of single letters needed to transform one word into the other used previously in information theory [26] has also been implemented for the purpose of automatic clustering of languages [27, 8, 28] to compare the phonetic or phonological realizations of a particular vocabulary across the range of languages. The standard edit distance also gives deceptive results [29] if applied to the orthographic realizations of meanings in the different languages, since lengthy words provide more room for editing being therefore responsible for a decisive statistical impact distorting the results on language classification essentially. In order to compare two words having the same meaning albeit different lengths, the actual edit distance have to be normalized by the number of characters in these words. In [25], the original edit distance has been rescaled by the *average* length of the two words being compared. In our work, being guided by [30, 31], while comparing two words, w_1 and w_2, we use the edit distance divided by the number of characters of the *longer of the two*,

$$D\left(w_1, w_2\right) \;=\; \frac{\left\|w_1, w_2\right\|_L}{\max\left(\left|w_1\right|, \left|w_2\right|\right)} \tag{1}$$

where $\left\|w_1, w_2\right\|_L$ is the standard Levenshtein distance between the words w_1 and w_2, and $|w|$ is the number of characters in the word w. For instance, ac-

cording to (1) the normalized Levenshtein distance between the orthographic realizations of the meaning *milk* in English and in German (*Milch*) equals 2/5. Such a normalization seems natural since the deleted symbols from the longer word and the empty spaces added to the shorter word then stand on an equal footing: the shorter word is supplied by a number of spaces to match the length of the longer one. The obvious advantage of (1) against the normalization used in [25] is that $D(w_1, w_2)$ takes values between 0 and 1 for any two words, w_1 and w_2, so that $D(w, w) = 0$, and $D(w_1, w_2) = 1$ when all characters in these words are different. Moreover, it is clear that the normalized edit distance defined in (1) is symmetric, i.e. $D(w_1, w_2) = D(w_2, w_1)$. The normalized edit distance between the orthographic realizations of two words (1) can be interpreted as the probability of mismatch between two characters picked from the words at random.

In order to obtain the lexical distances between the two languages, l_1 and l_2, we compute the average of the normalized Levenshtein distances (1) over Swadesh's vocabulary [22] of 200 meanings - the smaller the result is, the more affine are the languages,

$$d(l_1, l_2) = \frac{1}{200} \cdot \sum_{\alpha \in \text{Swadesh list}} D\left(w_\alpha^{(l_1)}, w_\alpha^{(l_2)}\right), \qquad (2)$$

where α is a meaning from Swadesh's vocabulary, and $w_\alpha^{(l)}$ is its orthographic realization in the language l. It is obvious that $d(l, l) = 0$, and $d(l_1, l_2) = 1$ if none of Swadesh's words belonging to the language l_1 has any common character with those words of the same meanings in the language l_2 that is already improbable even over the short list of 200 meanings. The lexical distance (2) between two languages, l_1 and l_2, can be interpreted as the average probability to distinguish them by a mismatch between two characters randomly chosen from the orthographic realizations of Swadesh's meanings. It is worth a mention that although the lexical distance defined by (2) can be calculated formally for any pair of languages, we have used it only for

the evaluation of distances between the languages belonging to the same language family because of we like to construct the geometric representation of relations within the particular language families and not of relations between the different families, which is also possible in the framework of our method. As a result, for the two samples of 50 languages selected from the IE and AU language families, we obtained the two symmetric $50 \times 50-$matrices, $d(l_1, l_2) = d(l_2, l_1)$, with vanishing diagonal elements, $d(l, l) = 0$; each matrix therefore contains $1,225$ independent entries. The encoding by lexical distances (2) is fully automated and therefore not time consuming at variance with the cognacy approach used in glottochronology. Comparing the edit distances between languages based on orthographic realizations might reflect different kinds of distances between languages (social, cultural, political) and not only genetic [32]. The phylogenetic trees from the lexical distance matrices (2) were constructed in [31, 30].

5 The structural component analysis on language data

Component analysis is a standard tool in diverse fields from neuroscience to computer graphics. It helps to reduce a complex data set to a lower dimension suitable for visual apprehension and to reveal its simplified structures. Independent component analysis (ICA) [33] and Principal component analysis (PCA) [34] are widely used for separating a multivariate signal into additive subcomponents. The mutual statistical *independence* of the non-Gaussian source signals are supposed for the data subjected for the ICA analysis. The method finds the independent components by maximizing the statistical independence of the data instances being an efficient tool for separating independent signals mixed together like in the classical example of

the "cocktail party problem", where a number of people are talking simultaneously in a room, and one is trying to follow one of the discussions. It is obviously inapplicable for reconstructing language phylogenies.

In the standard PCA analysis, the source signals are considered as simply *linearly* correlated, while all possible high-order dependencies are removed from the data set. In the course of the PCA method, the data instances are ordered according to their variance with respect to the mean by moving as much of the variance as possible into the first few dimensions. However, there is no reason to suggest that the directions of maximum variance recovered by the standard PCA method are good enough for identification of principal components in the linguistic data. Although both the statistical methods [33, 34] are applied on a multitude of real world problems, their predictions largely fail not only on the essentially non-random, strongly correlated data sets, but even on multi-modal Gaussian data. It is clear that the standard techniques of component analysis have to be dramatically improved for any meaningful application on language data. Since all languages within a language family interact with each other and with the languages of other families in 'real time', it is obvious that any historical development in language cannot be described only in terms of 'pair-wise' interactions, but it reflects a genuine higher order influence among the different language groups. Generally speaking, the number of parameters describing all possible parallels we may observe between the linguistic data from the different languages would increase exponentially with the data sample size. The only hope to perform any useful data analysis in such a case relies upon a proper choice of features that re-expresses the data set to make all contributions from an asymptotically infinite number of parameters *convergent* to some non-parametric *kernel*.

It is important to mention that any symmetric matrix of lexical distances (2) uniquely determines a weighted undirected fully connected graph, in

which vertices represent languages, and edges connecting them have weights equal to the relevant lexical distances between languages (2). Since the graph encoded by the matrix (2) is relatively small (of 50 vertices) and essentially not random, it is obviously out of the usual context of complex network theory [35]. A suitable method for the structural analysis of networks (weighted graphs) by means of *random walks* (or Markov chains, in a more general context) has been formulated in [36, 37, 38]. Being a version of the *kernel PCA* method [42], it generalizes PCA to the case where we are interested in principal components obtained by taking all higher-order correlations between data instances.

Before we explain how the most meaningful features of the lexical data encoded in the matrix (2) can be detected, let us note that there are infinitely many matrices that match all the structure of $d\left(l_i, l_j\right)$ and contain all the information about the relationships between languages estimated by means of the lexical distances (2). It is remarkable that all these matrices are related to each other by means of a linear transformation, which can be interpreted as a random walk [36, 37] defined on the weighted undirected graph determined by the matrix of lexical distances $d\left(l_i, l_j\right)$. We have to emphasize that random walks appear in our approach in concern to neither any particular assumption regarding to evolutionary processes in language (as we do not concern ourselves with the problems of modeling contagion or the spread of information through a society), nor the Bayesian analysis used previously [43, 3, 44] to construct the self-consistent tree-like representations in linguistic phylogenies, but as the *unique* linear transformation (in the class of stochastic matrices) consistent with all of the structure of the matrix of lexical distances calculated with respect to Swadesh's list of meanings.

A random walk associated to the matrix of lexical distances $d\left(l_i, l_j\right)$ calculated over the Swadesh vocabulary for a sample of N different languages (in our case, $N = 50$ for both language families) is defined by the transition

probabilities

$$T\left(l_i, l_j\right) \;=\; \Delta^{-1}\, d\left(l_i, l_j\right) \tag{3}$$

where the diagonal matrix $\Delta = \mathrm{diag}\left(\delta_{l_1}, \delta_{d_2}, \ldots \delta_{l_N}\right)$ contains the *cumulative* lexical distances $\delta_{l_i} = \sum_{j=1}^{N} d\left(l_i, l_j\right)$, for each language l_i. Diagonal elements of the matrix T are equal to zero, since $d\left(l_i, l_i\right) = 0$, for any language l_i. The matrix (3) is a stochastic matrix, $\sum_{j=1}^{N} T\left(l_i, l_j\right) = 1$, being nothing else, but the normalized matrix of lexical distances (2), in which a vector of probabilities $\mathbf{f}\left(l_i\right) \in [0,1]^N$, a row of the matrix $T\left(l_i, l_j\right)$, is attributed to each language l_i, with respect to all other languages in the language family,

$$\mathbf{f}\left(l_i\right) \;=\; \left(\frac{d\left(l_i, l_1\right)}{\delta_{l_i}}, \frac{d\left(l_i, l_2\right)}{\delta_{l_i}}, \ldots \frac{d\left(l_i, l_N\right)}{\delta_{l_i}}\right). \tag{4}$$

Each element of the vector (4) is a conditional probability describing the level of confidence that the language l_i can be identified successfully by comparing the orthographic representation of a randomly chosen Swadesh's meaning with that of the other language l_j, given the both languages belong to the same language family. It is worth a mention that since the sum of all elements in the probability vector, $\sum_{j=1}^{N} \left(f\left(l_i\right)\right)_j = 1$, for any language l_i, it is assumed that we can always confidently identify (with probability 1) a language by comparing its orthographic realizations with those from all other languages in the group.

Consequently, random walks defined by the transition matrix (3) describe the statistics of a sequential process of language classification. Namely, while the elements of the matrix $T\left(l_i, l_j\right)$ evaluate the successful identification of the language l_i provided the language l_j has been identified certainly, the elements of the squared matrix, $T^2\left(l_i, l_j\right) = \sum_{k=1}^{N} T\left(l_i, l_k\right) \cdot T\left(l_k, l_j\right)$, ascertain the successful identification of the language l_i from l_j through an intermediate language, the elements of the matrix T^3 give the probabilities to identify the language through two intermediate steps, and so on. By the way, the

whole host of complex and indirect relationships between orthographic representations of Swadesh's meanings encoded in our approach in the matrix of lexical distances (2) is uncovered by the powers T^n, $n \geq 1$, [36, 37].

Under the successive actions of T, any probability distribution vector $\mathbf{f}$ converges to a *stationary* distribution,

$$\pi \;=\; \lim_{n\to\infty} \mathbf{f}\, T^n \;=\; \mathbf{f}\, T^{\infty}, \tag{5}$$

where

$$\pi \;=\; \left(\frac{\delta_{l_1}}{\delta}, \frac{\delta_{l_2}}{\delta}, \ldots \frac{\delta_{l_N}}{\delta} \right), \quad \delta \equiv \sum_{i=1}^{N} \delta_{l_i} \tag{6}$$

is the 'center of mass', which *does not* coincide with the simple centroid vectors (means) calculated with respect to either columns or rows of a data matrix, in the course of the standard PCA analysis.

Random walks ascribe the total probability of successful classification for any two languages in the language family,

$$P\left(l_i, l_j\right) \;=\; \lim_{n\to\infty} \sum_{k=0}^{n} T^k\left(l_i, l_j\right) \;=\; \frac{1}{1-T}. \tag{7}$$

The operator $(1-T)^{-1}$ in the r.h.s. of the above equation diverges along the direction corresponding to the stationary distribution π (6) which belongs to the maximal eigenvalue 1 of the transition matrix (3), so that the last expression in (7) is formal. Nevertheless, we can use the Moore-Penrose generalized inverse matrix [40] instead of $(1-T)^{-1}$. The use of generalized inverses is common in the study of finite Markov chains [41]. Such a generalized inverse provides the unique best fit solution (with respect to least squares) to the system of linear equations described by the matrix $(1-T)^{-1}$ that lacks a unique solution. Under the Moore-Penrose inverse, any probability distribution vector $\mathbf{f}\left(l_i\right)$ is naturally translated into a perspective projection

$$\phi_i = \mathcal{P}_{\pi}\left(\mathbf{f}\left(l_i\right)\right),$$

with the vector of stationary distribution π as the center of projection (see the Appendix for details). We can use these projections in order to classify languages with respect to the center of mass π of the entire language family.

The kernel function required for the kernel PCA component analysis is expressed as the dot product (see [42] and references therein)

$$J \;=\; (\phi_i, \phi_j) \tag{8}$$

and constitute a square symmetric Gram $N \times N$-matrix. Each diagonal element $\|\phi_i\|^2 \equiv J_{ii}$ is the *first-passage time* [45] of random walks to $\mathbf{f}\,(l_i)$ defined on the weighted undirected graph determined by the matrix of lexical distances (2). The off-diagonal entries J_{ij} quantify the interference of two random walks concluding at $\mathbf{f}\,(l_i)$ and $\mathbf{f}\,(l_j)$ respectively [36, 37].

It is remarkable that the matrix J plays the essentially same role for the structural component analysis, as the covariance matrix does for the usual PCA analysis. Like the covariance values reflect the structure and redundancy in the linearly correlated data, the large diagonal values of J correspond to the notable heterogeneity of the data instances, while the large magnitudes of the off-diagonal terms correspond to high redundancy in the data sample. However, in contrast to the covariance matrix which best explains the variance in the data with respect to the mean, the matrix J traces out all higher order dependencies among data entities.

6 Principal structural components of the lexical distance data

High-dimensional data, which require more than two or three dimensions to represent a complex nexus of relationships, are difficult to interpret. The standard goal of the component analysis is to minimize the redundancy in the data sample quantified by the off-diagonal elements J_{ij}. It is readily

achieved by solving an eigenvalue problem for the real positive symmetric kernel matrix (8). Namely, there is a real orthogonal matrix Q, $Q^{\top}Q = 1$, (where $Q^{\top}$ stands for the transposed matrix Q) such that

$$\Lambda \;=\; Q^{\top}JQ \tag{9}$$

is a diagonal matrix. Each column vector q_k of the matrix Q is an eigenvector of the linear transformation that determines a direction where J acts as a simple rescaling, $Jq_k = \lambda_k q_k$, with some real eigenvalue $\lambda_k \geq 0$ indicating the characteristic first-passage time associated to the virtually independent component q_k; each one represents an independent trait detected in the matrix of lexical distances $d\,(l_i, l_j)$ calculated over the Swadesh list of meanings.

The independent components $\{q_k\}$, $k = 1, \ldots, N$, define an orthonormal basis in $\mathbb{R}^N$ which specifies each language l_i by N numerical coordinates, $l_i \rightarrow (q_{1,i}, q_{2,i}, \ldots q_{N,i})$, which are the signed distances from the point representing the language l_i to the axes associated to the virtually independent components. Languages that cast in the same mould in accordance with the N individual data features are revealed by geometric proximity in Euclidean space spanned by the eigenvectors $\{q_k\}$ that might be either exploited visually, or accounted analytically. The rank-ordering of data traits $\{q_k\}$, in accordance to their eigenvalues $0 = \lambda_1 < \lambda_2 \leq \ldots \leq \lambda_N$, provides us with the natural geometric framework for dimensionality reduction. The minimal eigenvalue $\lambda_1 = 0$ corresponds to the vector of stationary distribution $\pi = q_1^2$ containing no information about components.

At variance with the standard PCA analysis [34], where the *largest* eigenvalues of the covariance matrix are used in order to identify the principal components, as being characterized by the largest variance with respect to the mean, while building language taxonomy, we are interested in detecting the groups of the *most similar* languages, with respect to the selected group of features. The components of maximal similarity are identified with the

eigenvectors belonging to the *smallest* non-trivial eigenvalues. In particular, we use the three consecutive components $(q_{2,i}, q_{3,i}, q_{4,i})$ as the three Cartesian coordinates of a language point $l_i(x, y, z)$ in order to build a three-dimensional geometric representation of language taxonomy. Points symbolizing different languages in space of the three major data traits are contiguous if the orthographic representations of Swadesh's meanings in these languages are similar. Although, we are doubtful of that such a statistical similarity detected automatically on a finite sample of lexicostatistical data can be directly related to the traditional isoglosses discussed by linguists, they would definitely help to formulate the plausible isogloss hypothesis for future testing (see Sec. 7).

7 Geometric representation of the Indo-European family

Many language groups in the IE family had originated after the decline and fragmentation of territorially-extreme polities and in the course of migrations when dialects diverged within each local area and eventually evolved into individual languages. In Fig. 1, we have shown the three-dimensional geometric representation of 50 languages of the IE language family in space of its three major data traits detected in the matrix of lexical distances calculated over the Swadesh list of meanings. Due to the striking central symmetry of the representation, it is natural to describe the positions of language points l_i with the use of spherical coordinates,

$$r_i \;=\; \sqrt{q_{2,i}^2 + q_{3,i}^2 + q_{4,i}^2}\;, \quad \theta_i \;=\; \arccos\left(\frac{q_{4,i}}{r_i}\right), \quad \phi_i \;=\; \arctan\left(\frac{q_{3,i}}{q_{2,i}}\right), \tag{10}$$

rather than the Cartesian system.

The principal components of the IE family reveal themselves in Fig. 1 by four well-separated spines representing the four biggest traditional IE lan-

guage groups: Romance & Celtic, Germanic, Balto-Slavic, and Indo-Iranian. These groups are monophyletic and supported by the sharply localized distributions of the azimuth (φ) and inclination (zenith) angles (θ) over the languages shown in Fig. 2.A and Fig. 2.B respectively.

The Greek, Romance, Celtic, and Germanic languages form a class characterized by approximately the same azimuthal angle (Fig. 2.A), thus belonging to one plane in the three-dimensional geometric representation shown in Fig. 1, while the Indo-Iranian, Balto-Slavic, Armenian, and Albanian languages form another class, with respect to the inclination (zenith) angle (Fig. 2.B).

It is remarkable that the division of IE languages with respect to the azimuthal and zenith angles evident from the geometric representation in Fig. 1 perfectly coincides with the well-known *centum-satem* isogloss of the IE language family (the terms are the reflexes of the IE numeral '100'), related to the evolution in the phonetically unstable palatovelar order [46]. The palatovelars merge with the velars in centum languages sharing the azimuth angle, while in satem languages observed at the same zenith angle the palatovelars shift to affricates and spirants. Although the satem-centum distinction was historically the first original dialect division of the Indo-European languages [47], it is not accorded much significance by modern linguists as being just one of many other isoglosses crisscrossing all IE languages [48]. The basic phonetic distinction of the two language classes does not justify in itself the areal groupings of historical dialects, each characterized by some phonetic peculiarities indicating their independent developments. The appearance of the division similar to the centum-satem isogloss (based on phonetic changes only) may happen because of the systematic sound correspondences between the Swadesh words across the different languages of the same language family.

The projections of Albanian, Greek, and Armenian languages onto the axes of the principal components of the IE family are rather small, as they

occupy the center of the diagram in Fig. 1. Being eloquently different from others, these languages can be resolved with the use of some minor components q_k, $k > 3$. Remarkably, the Greek and Armenian languages always remain proximate confirming Greeks' belief that their ancestors had come from Western Asia [49].

8 In search of lost time

Geometric representations of language families can be conceived within the framework of various physical models that infer on the evolution of linguistic data traits. In traditional glottochronology [22], the time at which languages diverged is estimated on the assumption that the core lexicon of a language changes at a constant average rate. This assumption based on an analogy with the use of carbon dating for measuring the age of organic materials was rejected by mainstream linguists considering a language as a social phenomenon driven by unforeseeable socio-historical events not stable over time [2]. Indeed, mechanisms underlying evolution of dialects of a proto-language evolving into individual languages are very complex and hardly formalizable.

In our method based on the statistical evaluation of differences in the orthographic realizations of Swadesh's vocabulary, a complex nexus of processes behind the emergence and differentiation of dialects within each language group is described by the single degree of freedom, along the radial direction (see (10)) from the origin of the graph shown in Fig. 1, while the azimuthal (φ) and zenith (θ) angles are specified by a language group.

It is worth a mention that the distributions of languages along the radial direction are remarkably heterogeneous indicating that the rate of changes in the orthographic realizations of Swadesh's vocabulary was anything but stable over time. Being ranked within the own language group and then

plotted against their expected values under the normal probability distribution, the radial coordinates of languages in the geometrical representation Fig. 1 show very good agreement with univariate normality, as seen from the normal probability plots in Fig. 3(A-D).

The hypothesis of normality of these distributions can be justified by taking on that for a long time the divergence of orthographic representations of the core vocabulary was a *gradual* change accumulation process into which many small, independent innovations had emerged and contributed additively to the outgrowth of new languages. Perhaps, the orthographic changes arose due to the fixation of phonetic innovations developed in the course of long-lasting interactions with non-IE languages in areas of their intensive historical contacts.

In physics, the univariate normal distribution is closely related to the time evolution of a mass-density function $\rho(r,t)$ under homogeneous diffusion in one dimension,

$$\rho(r,t) = \frac{1}{2\pi\sigma^2} \cdot \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right),$$

in which the mean value μ is interpreted as the coordinate of a point where all mass was initially concentrated, and variance $\sigma^2 \propto t$ grows linearly with time. If the distributions of languages along the radial coordinate of the geometric representation do fit to univariate normality for all language groups, then in the long run the value of variance in these distributions grew with time at some approximately constant rate. We have to emphasize that the locations of languages might not be distributed normally if it were not true; we did not do any assumption above. Again, the constant increment rates of variance of radial positions of languages in the geometrical representation Fig. 1 has nothing to do with the traditional glottochronological assumption about the steady borrowing rates of cognates [50]. For clarity, we have used a simple code to produce a sequence of normally distributed integer numbers,

with linearly growing variance: $[6, 7, 4, 3, 6, 4, 11, 7, 9, 4, 5, 1, 7, 2, 16]$; they obviously do not grow linearly. It is also important to mention that the values of variance σ^2 calculated for the languages over the individual language groups (see Fig. 3(A-D)) do not correspond to physical time but rather give a statistically consistent estimate of age for each language group. In order to assess the pace of variance changes with physical time and calibrate our dating method, we have to use the historically attested events.

Although historical compendiums report us on grace, growth, and glory succeeded by the decline and disintegration of polities in days of old, they do not tell us much about the simultaneous evolution in language. It is beyond doubt that massive population migrations and disintegrations of organized societies both destabilizing the social norms governing behavior, thoughts, and social relationships can be taken on as the chronological anchors for the onset of language differentiation. However, the idealized assumption of a punctual *split* of a proto-language into a number of successor languages shared implicitly by virtually all phylogenetic models is problematic for a linguist well aware of the long-lasting and devious process by which a real language diverges [2]. We do not aspire to put dates on such a fuzzy process but rather consider language as a natural appliance for dating of those migrations and fragmentation happened during poorly documented periods in history.

While calibrating the dating mechanism in our model, we have used the four anchor events [51]:

1. the last Celtic migration (to the Balkans and Asia Minor) (by 300 BC),
2. the division of the Roman Empire (by 500 AD),
3. the migration of German tribes to the Danube River (by 100 AD),
4. the establishment of the Avars Khaganate (by 590 AD) overspreading

Slavic people who did the bulk of the fighting across Europe.

It is remarkable that a very slow variance pace of a millionth per year

$$\frac{t}{\sigma^2} \;=\; (1.367 \pm 0.002) \cdot 10^6 \tag{11}$$

is evaluated uniformly, with respect to all of the anchoring historical events mentioned above.

The time–variance ratio (11) deduced from the well attested events allows us to retrieve the probable dates for

1. the break-up of the Proto-Indo-Iranian continuum preceding 2,400 BC, in a good agreement with the migration dates from the early Andronovo archaeological horizon [52];

2. the end of common Balto-Slavic history as early as by 1,400 BC, in support of the recent glottochronological estimates [53] well agreed with the archaeological dating of Trziniec-Komarov culture, localized from Silesia to Central Ukraine;

3. the separation of Indo-Arians from Indo-Iranians by 400 BC, probably as a result of Aryan migration across India to Ceylon, as early as in 483 BC [54];

4. the division of Persian polity into a number of Iranian tribes migrated and settled in vast areas of south-eastern Europe, the Iranian plateau, and Central Asia by 400 BC, shortly after the end of Greco-Persian wars [55].

9 Evidence for Proto-Indo-Europeans

The basic information about the Proto-Indo-Europeans arises out of the comparative linguistics of the IE languages. There were a number of proposals about early Indo-European origins in so far. For instance, the *Kurgan*

scenario postulating that the people of an archaeological "Kurgan culture" (early 4^{th} millennium BC) in the Pontic steppe were the most likely speakers of the proto- IE language is widely accepted [56]. The *Anatolian* hypothesis suggests a significantly older age of the IE proto-language as spoken in Neolithic Anatolia and associates the distribution of historical IE languages with the expansion of agriculture during the Neolithic revolution in the 8^{th} and 6^{th} millennia BC [47].

It is a subtle problem to trace back the diverging pathways of language evolution to a convergence in the IE proto-language since symmetry of the modern languages assessed by the statistical analysis of orthographic realizations of the core vocabulary mismatches that in ancient time. The major IE language groups have to be reexamined in order to ascertain the locations of the individual proto-languages as if they were extant. In our approach, we associate the mean μ of the normal distribution of languages belonging to the same language group along the radial coordinate r with the expected location of the group proto-language. Although we do not know what the exact values of means were, the sample means calculated over the several extant languages from each language group give us the appropriate estimators. There is a whole interval around each observed sample mean within which, the true mean of the whole group actually can take the value.

In order to target the locations of the five proto-languages (the Proto-Germanic, Latin, Proto-Celtic, Proto-Slavic, and Proto-Indo-Iranian) with the 95% confidence level, we have supposed that variances of the radial coordinate calculated over the studied samples of languages are the appropriate estimators for the true variance values of the entire groups. The expected locations of the proto-languages, together with the end points of the 95% confidence intervals, are displayed on the normal plots, in Fig. 3(A-D). Let us note that we did not include the Baltic languages into the Slavic group when computing the Proto-Slavic center point because these two groups exhibit

different statistics, so that such an inclusion would dramatically reduce the confidence level for the expected locations of the proto-languages. Although the statistical behavior of the proto-languages in the geometric representation of the IE family is not known, we assume that it can be formally described by the 'diffusion scenario', as for the historical IE languages. Namely, we assume that the locations of the five proto-languages from a statistically determined central point fit to multivariate normality. Such a null hypothesis is subjected to further statistical testing, in which the chi-square distribution is used to test for goodness of fit of the observed distribution of the locations of the proto-languages to a theoretical one. The chi-square distribution with k degrees of freedom describes the distribution of a random variable $Q = \sum_{i=1}^{k} X_i^2$ where X_i are k independent, normally distributed random variables with mean 0 and variance 1.

In Fig. 4, we have used a simple graphical test to check three-variate normality by extending the notion of the normal probability plot. The locations of proto-languages have been tested by comparing the goodness of fit of the scaled distances from the proto-languages to the central point (the mean over the sample of the five proto-languages) to their expected values under the chi-square distribution with three degrees of freedom. In the graphical test shown in Fig. 4, departures from three-variant normality are indicated by departures from linearity. Supposing that the underlying population of parent languages fits to multivariate normality, we conclude that the determinant of the sample variance-covariance matrix has to grow linearly with time. The use of the previously determined time–variance ratio (11) then dates the initial break-up of the Proto-Indo-Europeans back to 7,000 BC pointing at the early Neolithic date, to say nothing about geography, in agreement with the Anatolian hypothesis of the early Indo-European origin [47, 49, 46, 3, 57, 31].

The linguistic community estimates of dating for the proto IE language lie between 4,500 and 2,500 BC, a later date than the Anatolian theory predicts.

These estimations are primarily based on the reconstructed vocabulary (see [58] and references therein) suggesting a culture spanning the Early Bronze Age, with knowledge of the wheel, metalworking and the domestication of the horse and thus favoring the Kurgan hypothesis. It is worth a mention that none of these words are found in the Swadesh list encompassing the basic vocabulary related to agriculture that emerged perhaps with the spread of farming, during the Neolithic era. Furthermore, the detailed analysis of the terms uncovered a great incongruity between the terms found in the reconstructed proto-IE language and the cultural level met with in the Kurgans lack of agriculture [59]. Let us note that our dating (2,400 BC) for the migration from the Andronovo archaeological horizon (see Sec. 8) and the early break-up of the proto-Indo-Iranian continuum estimated by means of the variance (see Fig. 3.C) is compatible with the Kurgan time frame. However, despite the Indo-Iranian group of languages being apparently the oldest among all other groups of the IE family, we cannot support the general claim of the Kurgan hypothesis, at least on the base of Swadesh's lexicon.

10 In Search of Polynesian origins

The colonization of the Pacific Islands is still the recalcitrant problem in the history of human migrations, despite many explanatory models based on linguistic, genetic, and archaeological evidences have been proposed in so far. The origins, relationships, and migration chronology of Austronesian settlers have constituted the sustainable interest and continuing controversy for decades. The components probe for a sample of 50 AU languages immediately uncovers the both Formosan (F) and Malayo-Polynesian (MP) branches of the entire language family (see Fig. 5).

The distribution of azimuth angles shown in Fig. 6.A identifies them as two monophyletic jets of languages that cast along either axis spanning the

entire family plane. The clear geographic patterning is perhaps the most remarkable aspect of the geometric representation. It is also worth mentioning that the language groupings as recovered by the component analysis of lexical data reflect profound historical relationships between the different groups of AU population. For instance, the Malagasy language spoken in Madagascar casts in the same mould as the Maanyan language spoken by the Dayak tribe dwelling in forests of Southern Borneo and the Batak Toba language of North Sumatra spoken mostly west of Lake Toba.

Despite Malagasy sharing much of its basic vocabulary with the Maanyan language [60], many manifestations of Malagasy culture cannot be linked up with the culture of Dayak people: the Malagasy migration to East Africa presupposes highly developed construction and navigation skills with the use of out-rigger canoes typical of many Indonesian tribes which the Dayak people however do not have, also some of the Malagasy cultivations and crop species (such as wet rice) cannot be found among forest inhabitants. In contrast, some funeral rites (such as the second burial, *famadihana*) typical of the leading entities of the Madagascar highlands are essentially similar to those of Dayak people. A possible explanation is that population of the Dayak origin was brought to Madagascar as slaves by Malay seafarers [30]. As the Dayak speakers formed the majority in the initial settler group, in agreement with the genetic parental lineages found in Madagascar [61], their language could have constituted the core element of what later became Malagasy, while the language of the Malay dominators was almost suppressed, albeit its contribution is still recovered by the exploration of the leading traits on language data.

The AU language family forks at the northernmost tip of the Philippines, the Batanes Islands located about 190 km south of Taiwan (see Fig. 6.B). On the distribution of azimuth angles shown in Fig. 6.A, the Itbayaten language representing them in the studied sample is pretty close to the azimuth, $\varphi = 0$,

bridging over the separating language family branches (Fig. 6.B). By the way, the MP-offset descends from the northern Philippines (the northern Luzon Island) and springs forth eastward through the Malay Archipelago across Melanesia culminating in Polynesia (Fig. 7); in accordance with the famous 'express train' model of migrations peopled the Pacific [62]. In its turn, the F-branch embarks on the southwest coast of Taiwan and finds its way to the northern Syueshan Mountains inhabited by Atayal people that compose many ethnic groups with different languages, diverse customs, and multiple identities. Evidently, both the offshoots derived their ancestry in Southeast Asia as strengthened by multiple archaeological records [62], but then evolved mostly independently from each other, on evidence of the Y-chromosome haplotype spread over Taiwanese and Polynesian populations [63]. The Bayesian methods for the language phylogeny trees [43] also evinced the earliest separation of these two branches of the AU language family. However, in the recent pulse-pause scenario [44], the Taiwanese origin of the entire AU family was suggested because of the "considerable diversity of Formosan languages". It is important to note that diversity itself is by no means a reliable estimate provided symmetry is downplayed (e.g., in spite of the greatest diversity, the Indo-Iranian language group is not an origin of the entire IE language family).

The distribution of languages spoken within Maritime Southeast Asia, Melanesia, Western Polynesia and of the Paiwan language group in Taiwan over the distances from the center of the diagram representing the AU language family in Fig. 5 conforms to univariate normality (see Fig. 8) suggesting that an interaction sphere had existed encompassing the whole region, from the Philippines and Southern Indonesia through the Solomon Islands to Western Polynesia, where ideas and cultural traits were shared and spread as attested by trade [64, 65] and translocation of farm animals [66, 67] among shoreline communities.

Although the lack of documented historical events makes the use of the developed dating method difficult, we may suggest that variance evaluated over Swadesh's vocabulary forges ahead approximately at the same pace uniformly for all human societies involved in trading and exchange forming a singular cultural continuum. Then, the time–age ratio (11) deduced from the previous chronological estimates for the IE family returns 550 AD if applied to the Austronesians as the likely break-up date of their cultural continuum, pretty well before 600-1,200 AD while descendants from Melanesia settled in the distant apices of the Polynesian triangle as evidenced by archaeological records [68, 69, 70].

11 Austronesian languages riding an express train

The distributions of languages spoken in the islands of East Polynesia and of the Atayal language groups in Taiwan over the radial coordinate from the center of the geometric representation shown in Fig. 5 break from normality, so that the general 'diffusive scenario' of language evolution used previously for either of the chronological estimates is obviously inapplicable to them. For all purposes, the evolution of these extreme language subgroups cannot be viewed as driven by independent, petty events. Although the languages spoken in Remote Oceania clearly fit the general trait of the entire MP-branch, they seem to evolve without extensive contacts with Melanesian populations, perhaps because of a rapid movement of the ancestors of the Polynesians from South-East Asia as suggested by the 'express train' model [62] consistent with the multiple evidences on comparatively reduced genetic variations among human groups in Remote Oceania [71, 72, 73].

In order to obtain reasonable chronological estimates, an alternative mech-

anism on evolutionary dynamics of the extreme language subgroups in space of traits of the AU language family should be reckoned with. The simplest 'adiabatic' model entails that no words had been transferred to or from the languages riding the express train to Polynesia, so that the lexical distance among words of the most distanced languages tends to increase primarily due to random permutations, deletions or substitutions of phonemes in the words of their ancestor language. Under such circumstances the radial coordinate of a remote language riding an 'express train' in the geometric representation (see Fig. 5) effectively quantifies the duration of its relative isolation from the Austronesian cultural continuum. Both of the early colonization of a secluded island by Melanesian seafarers and of the ahead of time migration of the indigenous people of Taiwan to highlands can be discerned by the excessively large values of the radial coordinates r of their languages. In Fig. 9, we have presented the log-linear plot, in which the radial coordinates of remote languages were ranked and then plotted against their expected values under the exponential distribution (shown by the dash-dotted line in Fig. 9).

The radial coordinates of the languages at the distant margins of the AU family diagram shown in Fig. 9 may be deduced as evolving in accordance with the simple differential equation

$$\dot{r} \;=\; ar \tag{12}$$

where $\dot{r}$ means the derivative of r with respect to isolation time, and $a > 0$ is some constant quantifying the rate of radial motion of a language riding the express train in space of the major traits of the AU family. The suggested model of language taxonomy evolution is conceived by that while the contact borrowings are improbable the orthographic realizations of Swadesh's meanings would accumulate emergent variations in spellings, so that the radial coordinate of a remote language can formally grow unboundedly with isolation time.

A simple equation mathematically similar to (12) has been proposed by M. Swadesh [22] in order to describe the change of cognates in time, in the framework of the glottochronological approach. In our previous work [30], another similar equation has been suggested for the purpose of modeling the time evolution of normalized edit distances between languages. We have to emphasize that the statistical model (12) can be hardly related to them both, as the radial coordinate r in the geometrical representations of language families described above does not have a direct relation to neither the percentage of cognates, nor the edit distance.

Then the relative dates estimating the duration of relative isolation of the distant languages from the extensive contacts with other Austronesian languages can be derived basing on the assumption (12) as

$$t_1 - t_2 \;=\; \frac{1}{a} \cdot \ln \frac{r_1}{r_2} \tag{13}$$

where $r_2 > r_1$ are the radial coordinates of the languages from the center of the sample diagram shown in Fig. 5.

Tahiti located in the archipelago of Society Islands is the farmost point in the geometric representation of the Austronesian family and the foremost Austronesian settlement in the Remote Oceania attested as early as 300 BC [68], the date we placed the incipience of the Tahitian society. According to many archaeological reconstructions [68, 69, 70], descendants from West Polynesia had spread through East Polynesian archipelagos and settled in Hawaii by 600 AD and in New Zealand by 1,000 AD testifying the earliest outset dates for the related languages. It is worth mentioning that all stride times between the offsets of these three Polynesian languages hold consistently the same rate

$$a \;=\; (4.27 \pm 0.01) \cdot 10^{-4} \tag{14}$$

affirming the validity of the 'adiabatic' conjecture described above and allowing us to assign the estimated dates to the marks of the horizontal axis

of the timing diagram presented in Fig. 9. The language divergence among Atayal people distributed throughout an area of rich topographical complexity is neatly organized by the myths of origin place, consanguine clans, and geographical barriers that have lead to the formation of a unique concept of ethnicity remarkable for such a geographically small region as Taiwan. The complexity of the Atayal ethnic system and the difficulty of defining the ethnic borders hindered the classification of the Atayal regional groups and their dialects which has been continuously modified throughout the last century.

In our work, we follow the traditional classification [74] of the Atayal group into three branches based on their places of origin: Sediq (Sedek), Ciuli (Tseole) Atayal, and Squiliq (Sekilek) Atayal. In account with the standard lexicostatistic arguments [75], the Sediq dialect subgroup could have split off from the rest of the Atayal groups about 1,600 years ago, as both the branches share up to a half of the cognates in the 200 words of basic vocabulary. This estimated date is very tentative in nature and calls for a thorough crosschecking. The Atayal people had been recognized as they had started to disperse to the northern part of Taiwan around 1,750 AD [76]. Being formed as the isolated dialect subgroups in island interiors, they showed the greatest diversity in race, culture, and social relations and sometimes considered each other as enemies and prime head hunting targets.

Given the same rate of random phonetic changes as derived for the Polynesian languages, the 'adiabatic' model of language evolution returns the stride times of 1,000 years between the Sediq dialect subgroup and Squiliq Atayal and of 860 years between the Ciuli and Squiliq Atayal languages. Consistently, Sediq is estimated to have branched off from the other Atayal languages 140 years before the main Atayal group split into two. The Squiliq subgroup had been attested during the latest migration of Atayal people, as late as 1,820 AD [76]. Perhaps, a comprehensive study of the Atayal dialects by their symmetry can shed light on the origins of the Atayal ethnic system

and its history.

12 Conclusion

We have presented the new paradigm for the language phylogeny based on the analysis of geometric representations of the major traits on language data. The proposed method is fully automated.

On the encoding stage, we evaluated the lexical distances between languages by means of the mean normalized edit distances between the orthographic realizations of Swadesh's meanings. Then, we considered an infinite sequential process of language classification described by random walks on the matrix of lexical distances. As a result, the relationships between languages belonging to one and the same language family are translated into distances and angles, in multidimensional Euclidean space. The derived geometric representations of language taxonomy are used in order to test the various statistical hypotheses about the evolution of languages.

Our method allows for making accurate inferences on the most significant events of human history by tracking changes in language families through time. Computational simplicity of the proposed method based primarily on linear algebra is its crucial advantage over previous approaches to the computational linguistic phylogeny that makes it an invaluable tool for the automatic analysis of both the languages and the large document data sets that helps to infer on relations between them in the context of human history.

13 Acknowledgments

We profoundly thank R.D. Gray for the permission to use the Austronesian Basic Vocabulary Database [77] containing lexical items from languages spoken throughout the Pacific region.

We are deeply grateful to J. Nichols, T. Warnow, S. Wichmann, R. Gray, and J. Salmons for their kind advises and multiple consultations during the preparation of the present work.

Appendix

The stationary distribution of random walks (6) defines a unique measure on the set of languages with respect to which the transition operator T (3) is self-adjoint,

$$\hat{T} \;=\; \frac{1}{2}\left(\pi^{1/2}\,T\,\pi^{-1/2} + \pi^{-1/2}\,T^{\top}\,\pi^{1/2}\right), \tag{15}$$

where $T^{\top}$ is the adjoint operator, and π is the diagonal matrix

$$\pi \;=\; \mathrm{diag}\left(\pi_1, \ldots, \pi_N\right), \quad \pi_i \;=\; \frac{\delta_{l_i}}{\delta}, \quad i = 1, \ldots, N.$$

The spectral properties of the self-adjoint operators related to random walks and diffusions are widely used in the analysis of complex networks [37] and in spectral graph theory [39]. All eigenvalues of (15) are real $1 = \nu_1 > \nu_2 \geq \ldots \geq \nu_N \geq -1$, with orthonormal eigenvectors $\{\psi_k\}_{k=1}^{N}$ mapping the nodes V of the graph uniquely determined by the matrix of lexical distances onto the $(N-1)$-dimensional unit hyper-sphere, $\psi_k : V \to S_1^{(N-1)}$.

The inverse Laplace operator $L^{-1} = (1-T)^{-1}$ quantifying the probability of successful classification of languages is not invertible over S_1^{N-1}, but over $S_1^{N-1} - \{\psi_1\}$, the orthogonal complement of the first eigenvector ψ_1 (belonging to the largest eigenvalue of (15) $\nu_1 = 1$) that corresponds to the transient process of random walks toward the stationary distribution $\pi = \psi_1^2$. This orthogonal complement is homeomorphic to the projective hyper-plane $P\mathbb{R}_{\pi}^{(N-1)}$ constructed by linearly mapping points of the unit hyper-sphere S_1^{N-1} from ψ_1 as the center of projection.

Each language l_i has an image in $P\mathbb{R}^{N-1}$ determined by the vector

$$\phi_i = \left(\frac{\psi_{2,i}}{\psi_{1,i}\sqrt{(1-\nu_2)}}, \ldots, \frac{\psi_{N,i}}{\psi_{1,i}\sqrt{(1-\nu_N)}} \right) \tag{16}$$

where all $\psi_{1,i} = \sqrt{\pi_i} > 0$. The kernel function required for the component analysis is expressed as the dot product,

$$J = (\phi_i, \phi_j) .$$

References

[1] Nichols, J., Warnow, T. (2008) "Tutorial on computational linguistic phylogeny." *Language and Linguistics Compass* **2** (5), 760.

[2] Heggarty, P. (2006) "Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully be Applied to Language Data and to Dating Language?" In P. Forster& C. Renfrew (Eds.) *Phylogenetic Methods and the Prehistory of Languages*, p.183, McDonald Institute for Archaeological Research, Cambridge.

[3] Gray, R.D., Atkinson, Q.D. (2003). "Language-tree divergence times support the Anatolian theory of Indo-European origin." *Nature* **426**, 435.

[4] Ben Hamed, M., Wang, F. (2006). "Stuck in the forest: Trees, networks and Chinese dialects". *Diachronica* **23**(1): iv 230, pp. 29.

[5] Heggarty, P. (2008). "Splits or Waves? Trees or Webs? Network Analysis of Language Divergence", *AHRC Conference on Cultural and Linguistic Diversity*, Great Missenden, 9th-13th December.

[6] Forster,P., Toth, A (2003). "Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European." In *Proceedings of the National Academy of Sciences USA*, **100**(15), p. 9079.

[7] Nakhleh, L., Ringe, D., Warnow, T. (2005) "Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages." *Language* **81**(2), 382-420.

[8] McMahon, A., McMahon, R. (2005) *Language classification by numbers.* Oxford, UK: Oxford University Press.

[9] Bryant, D., Filimon, F., Gray, R.D. (2005). "Untangling our past: Languages, Trees, Splits and Networks." In *The Evolution of Cultural Diversity: Phylogenetic Approaches* by Mace, R., Holden, C., Shennan, S. (Eds.), UCL Press, London, p. 69.

[10] Barbançon, F., Warnow, T., Evans, S.N., Ringe, D.A. Jr, Nakhleh, L. (2006) "An experimental study comparing linguistic phylogenetic reconstruction methods". In *Proceedings of a Conference on Language and Genes*, University of California, Santa Barbara.

[11] Holden,C.J., Gray, R.D.(2006) "Exploring Bantu linguistic relationships using trees and networks." In P. Forster& C. Renfrew (Eds.) *Phylogenetic Methods and the Prehistory of Languages*, p. 19, McDonald Institute for Archaeological Research, Cambridge.

[12] Gray, R.D., Greenhill, S.J, Ross, R.M. (2007). "The Pleasures and Perils of Darwinizing Culture (with phylogenies)." *Biological Theory* **2**(4), p. 360.

[13] Gordon, R.G. Jr. (ed.) (2005) *Ethnologue: Languages of the World*, the 15th edition. Dallas, TX.: SIL International. Online version: *http://www.ethnologue.com/*.

[14] Michener, C.D., Sokal, R.R. (1957) "A quantitative approach to a problem in classification. " *Evolution* **11**, 130.

[15] Saitou, N., Masatoshi, N. (1987) "The neighborhood joining method: a new method of constructing phylogenetic trees." *Molecular Biology and Evolution* **4**, 406.

[16] The database is available at *http://univaq.it/∼serva/languages/languages.html*

[17] Dyen, I., Kruskal, J., Black, P. (1997) *Comparative Indo-European Database* collected by Isidore Dyen. (Available at *http://www.wordgumbo.com/ie/cmp/iedata.txt*) Copyright (C) 1997 by Isidore Dyen, Joseph Kruskal, and Paul Black. The file was last modified on Feb 5, 1997. Redistributable for academic, non-commercial purposes.

[18] A. McMahon, P. Heggarty, R. McMahon, N. Slaska,(2005) "Swadesh sublists and the benefits of borrowing: an andean case study. " *Transactions of the Philological Society* **103**(2), 147.

[19] Greenhill, S.J., Blust, R., Gray, R.D. (2008) "The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics." *Evolutionary Bioinformatics* **4**, 271. The Austronesian Basic Vocabulary Database is available at *http://language.psy.auckland.ac.nz/austronesian.*

[20] Dyen, I., Kruskal, J.B., Black, P. (1992) "An Indo-european classification: A lexicostatistical experiment." *Transactions of American Philosophical Society* **82**(5) 1-132.

[21] D. d'Urville, (1832) "Sur les îles du Grand Océan". *Bulletin de la Société de Goégraphie* **17**, 1.

[22] Swadesh, M. (1952) "Lexico-statistic dating of prehistoric ethnic contacts." *Proceedings of the American Philosophical Society* **96**, 452.

[23] Wang, W.S.-Y., Minett, J.W.(2005) "Vertical and Horizontal Transmission in Language Evolution" *Transactions of the Philological Society* **103**(2), 121.

[24] Warnow, T. , Evans, S.N., Ringe, D.A. Jr., Nakhleh, L. (2006) "A stochastic model of language evolution that incorporates homoplasy and borrowing." In P. Forster& C. Renfrew (Eds.) *Phylogenetic Methods and the Prehistory of Languages*, p.75, McDonald Institute for Archaeological Research, Cambridge.

[25] Ellison, T.M., Kirby, S. (2006) "Measuring Language Divergence by Intra-Lexical Comparison". In Proceedings of the *21st International Conference on Computational Linguistics & 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17-21 July 2006.

[26] Levenshtein, V.I. (1966) "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady* **10**, 707.

[27] Nerbonne, J., Heeringa, W., Kleiweg, P. (1999) "Edit Distance and Dialect Proximity"; p.5-15, in Sankoff, D., Kruskal J. (Eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison* , Stanford: CSLI Press.

[28] Kessler, B. (2005) "Phonetic comparison algorithms." *Transactions of the Philological Society* **103**(2), 243 .

[29] Batagelj, V., Pisanski, T., Keržic, D. (1992) "Automatic clustering of languages". *Computational Linguistics* **18**(3), 339.

[30] Petroni, F., Serva, M. (2008) "Language distance and tree reconstruction." *Journal of Statistical Mechanics: Theory and Experiment*, P08012.

[31] Serva, M., Petroni, F. (2008) "Indo-European languages tree by Levenshtein distance". *Europhysics Letters* **81**, 68005.

[32] We thank our reviewer for this profound comment.

[33] Hyvärinen, A., Karhunen, J., Oja, E. (2001) *Independent Component Analysis*, New York, Wiley.

[34] Jolliffe, I.T. (2002) *Principal Component Analysis*, Springer Series in Statistics **XXIX**, 2nd ed., Springer, NY.

[35] Dorogovtsev, S.N. (2010) *Lectures on Complex Networks,* Oxford University Press, Oxford.

[36] Blanchard, Ph., Volchenkov, D. (2008) "Intelligibility and first passage times in complex urban networks." *Proceedings of the Royal Society A* **464**, 2153.

[37] Blanchard, Ph., Volchenkov, D.(2009) *Mathematical Analysis of Urban Spatial Networks*. In Springer series: Understanding Complex Systems **XIV**.

[38] Volchenkov, D. (2010) "Random Walks and Flights over Connected Graphs and Complex Networks". In *Communications in Nonlinear Science and Numerical Simulation*, *http://dx.doi.org/10.1016/j.cnsns.2010.02.016.*

[39] Chung, F. (1997) *Lecture notes on spectral graph theory*, AMS Publications, Providence.

[40] Penrose, R. (1955) "A generalized inverse for matrices". *Mathematical Proceedings of the Cambridge Philosophical Society* **51**, 406.

[41] Meyer, C.D. (1975) "The role of the group generalized inverse in the theory of finite Markov chains", *The Review of Society for Industrial and Applied Mathematics* (SIAM Review) **17**, 443.

[42] Schölkopf, B., Smola, A.J., Müller, K.-R. (1998) "Nonlinear component analysis as a kernel eigenvalue problem." *Neural Computation* **10**, 1299.

[43] Gray, R.D., Jordan, F.M. (2000) "Language trees support the express-train sequence of Austronesian expansion." *Nature* **405**, 1052.

[44] Gray, R.D., Drummond, A.J., Greenhill, S.J. (2009) "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement." *Science* **323**, 479.

[45] Lovász, L. (1993) "Random walks on graphs: Survey." *Bolyai Society Mathematical Studies* **2**, 1, Keszthely, Hungary.

[46] Gamkrelidze, Th.V., Ivanov, V.V. (1995) *Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and a Proto-Culture.* Mouton de Gruyter Series *Trends in Linguistics: Studies and Monographs* **80**, Berlin, New York.

[47] Renfrew, C.(1987) *Archaeology and Language: The Puzzle of Indo-European Origins*. New York, Cambridge University Press.

[48] Baldi, Ph. (2002) *The Foundations of Latin*. Mouton de Gruyter Series *Trends in Linguistics: Studies and Monographs* **117**, Berlin, New York.

[49] Gamkrelidze, Th.V., Ivanov, V.V. (1990) "The early history of Indo-European languages. " *Scientific American* **262**(3), 110.

[50] Embelton, S.M. (1986)*Statistics in Historical Linguistics*, Bochum, Brockmeyer.

[51] Fouracre, P. (1995-2007) *The New Cambridge Medieval History*, Cambridge University Press.

[52] Bryant, E. (2001) *The Quest for the Origins of Vedic Culture: The Indo-Aryan Migration Debate*. Oxford University Press.

[53] Novotná, P. , Blažek, V. (2007) "Glottochronolgy and its application to the Balto-Slavic languages". *Baltistica* **XLII** (2), 185.

[54] Mcleod, J. (2002) *The History of India*. Greenwood Pub. Group.

[55] Green, P. (1996) *The Greco-Persian Wars*. Berkeley, Los Angeles, London, University of California Press.

[56] Gimbutas, M. (1982) "Old Europe in the Fifth Millenium B.C.: The European Situation on the Arrival of Indo-Europeans", in E.C. Polomé (eds.) *The Indo-Europeans in the Fourth and Third Millennia*, Ann Arbor: Karoma Publishers.

[57] Renfrew, C. (2003) "Time Depth, Convergence Theory, and Innovation in Proto-Indo-European". In Proceedings of the Conference *Languages in Prehistoric Europe*, p. 227, Eichstätt University, 4-6 October 1999, Heidelberg, published in (2003).

[58] Mallory,J.P. (1991) *In Search of the Indo-Europeans: Language, Archaeology, and Myth*. London, Thames & Hudson.

[59] Krell, K.S. (1998) "Gimbutas Kurgan-PIE Homeland Hypothesis: A Linguistic Critique. In Blench R., Spriggs, M. (Eds.) *Archaeology and Language*, **II** p. 267, London, Routledge.

[60] Dahl, O.C. (1951) *Avhandlinger utgitt av Egede-Instituttet* **3**, 408, Arne Gimnes Forlag.

[61] Hurles, M.E., Sykes, B.C., Jobling, M.A., Forster, P.(2005) "The dual origins of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and paternal lineages." *American Journal of Human Genetics* **76**, 894.

[62] Diamond, J.M. (1988) "Express train to Polynesia." *Nature* **336**, 307.

[63] Su, B., *et al* (2000) "Polynesian origins: Insights from the Y chromosome". *Proceedings of the National Academy of Sciences* **97** (15), 8225.

[64] Bellwood, P., Koon, P. (1989) "Lapita colonists leave boats unburned!" *Antiquity* **63** (240), 613.

[65] Kirch, P.V. (1997) *The Lapita Peoples: Ancestors of the Oceanic World*. Cambridge, Mass., Blackwell.

[66] Matisoo-Smith, E.,Robins, J.H. (2004) "Origins and dispersals of Pacific peoples: Evidence from mtDNA phylogenies of the Pacific rat." *Proceedings of the National Academy of Sciences* **101** (24), 9167.

[67] Larson, G., *et al*, (2007) "Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in Island Southeast Asia and Oceania. " *Proceedings of the National Academy of Sciences* **104** (12), 4834.

[68] Kirch, P.V. (2000) *On the road of the winds: an archaeological history of the Pacific Islands before European contact*. University of California Press, Berkley, CA.

[69] Anderson, A., Sinoto,Y. (2002) "New radiocarbon ages for colonization sites in East Polynesia." *Asian Perspectives* **41**, 242.

[70] Hurles, M.E., *et al*, (2003) "Untangling Pacific settlement: the edge of the knowable", *Trends in Ecology & Evolution* **18**, 531.

[71] Lum, J.K., Jorde, L.B., Schiefenhovel, W. (2002) "Affinities among Melanesians, Micronesians, and Polynesians: A Neutral, Biparental Genetic Perspective." *Human Biology* **74**, 413.

[72] M. Kayser *et al*, (2006) "Melanesian and Asian Origins of Polynesians: mtDNA and Y Chromosome Gradients Across the Pacific. " *Molecular Biology and Evolution* **23**, 2234.

[73] Friedländer, J.S., *et al* (2008) "Genetic Structure of Pacific Islanders." *PLoS Genetics* (Public Library of Science) **4**(1), e19.

[74] Utsurikawa, N. (1935) *A Genealogical and Classificatory Study of the Formosan Native Tribes*. Tokyo: Toko shoin.

[75] Li, P.J., (1983) "Types of lexical derivation of men's speech in Mayrinax." *Bulletin of the Institute of History and Philology*, Academia Sinica **54** (3) 1.

[76] Li, P.J., (2001) "The Dispersal of The Formosan Aborigines in Taiwan." *Language and Linguistics* **2**(1), 271.

[77] Available at *http://language.psy.auckland.ac.nz/austronesian/*.

Vitae

1. **Philippe Blanchard** (Fig. 10) obtained his Ph.D. at the ETH-Zürich in mathematical physics. His main research interests lie in the use of functional analysis and probability theory, in quantum and statistical physics, in epidemiology and in sociology. He has authored more than

240 scientific papers and many books. He is director of the Research Centre BiBoS (Bielefeld-Bonn Stochastics) and editor of "Progress in Mathematical Physics" and "Mathematical Physics, Analysis and Geometry". He is a professor of mathematical physics at Bielefeld University, honorary professor at the East China Normal University (Shanghai) and scientific advisor of the Ecole Polytechnique Fédérale de Lausanne (EPFL).

2. **Filippo Petroni** (Fig. 11) was born near L'Aquila (Italy) on 10th december 1975. He graduated cum laude in Physics from Turin University. Later he was granted a Ph.D. in Applied Mathematics from the University of Newcastle upon Tyne (UK). Since february 2009 he is a postdoctoral fellow at the University of Rome "La Sapienza". He is interested in the study of complex systems

3. **Maurizio Serva** (Fig. 12) was born in Rome on 5th July 1959. He graduated cum laude in Physics at the University of Rome "La Sapienza". Later he obtained his Ph.D. in theoretical physics at the University of Bielefeld (Germany). Since June 1990 he is Researcher in Mathematical Physics (permanent position) at the University of L'Aquila. He temporarily worked in various Universities and institutions in Italy, France, Switzerland, Brazil, Argentina, Germany, Madagascar, Sweden, Denmark, and Switzerland. His research topics are stochastic models in biology and linguistics, stochastic finance, quantum and statistical mechanics.

4. **Dimitri Volchenkov** (Fig. 13) obtained his Ph.D. in theoretical physics at the Saint-Petersburg State University (Russia) and habilitated in CNRS Centre de Physique Theorique (Marseille, France). He worked in Texas A&M University (USA), Zentrum für Interdisziplinäre Forschung (Bielefeld, Germany), Centre de Physique Theorique (Marseille, France),

Bielefeld-Bonn Stochastic Research Center (Germany). He is the Researcher at the Center of Excellence Cognitive Interaction Technology (Bielefeld, Germany). His research interests are the non-perturbative quantum-field theory methods in stochastic dynamics and plasma turbulence, urban spatial networks and their impact on poverty and environments, stochastic analysis of complex networks, and physics of dance.

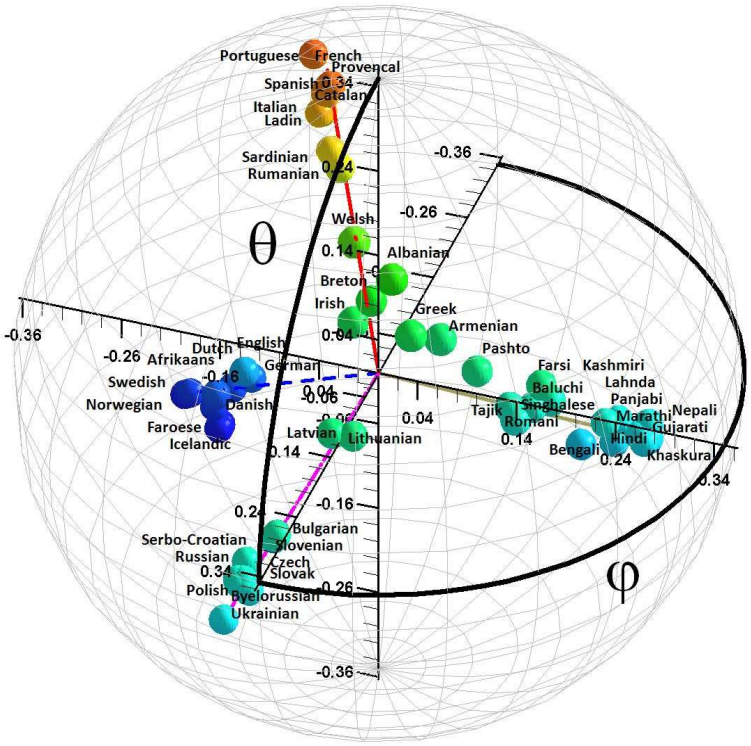
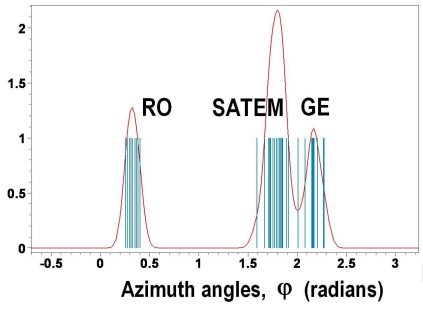


Figure 1: The three-dimensional geometric representation of the IE language family in space of the major data traits (q_2, q_3, q_4) color coded. The origin of the graph indicates the 'center of mass' $q_1 = \pi$ of the matrix of lexical distances $d(l_i, l_j)$, not the Proto-IE language. Due to the central symmetry of representation, it is convenient to use the spherical coordinates to identify the positions of languages: the radius from the center of the graph, the inclination angle θ, and the azimuth angle φ.

a.



b.

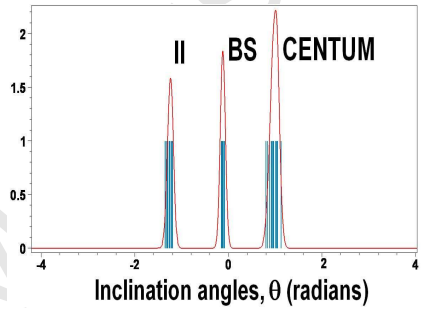


Figure 2: A). The kernel density estimates of the distributions of azimuthal angles in the three-dimensional geometric representation of 50 languages of the IE language family, together with the absolute data frequencies. Romance (RO), Germanic (GE), and the satem languages (SATEM) are easily differentiated with respect to the azimuthal angles. B). The kernel density estimates of the distributions of inclination (zenith) angles in the three-dimensional geometric representation of 50 languages of the IE language family, together with the absolute data frequencies. Indo-Iranian (II), Balto-Slavic (BS), and the centum languages (CENTUM) are attested by the inclination (zenith) angles.

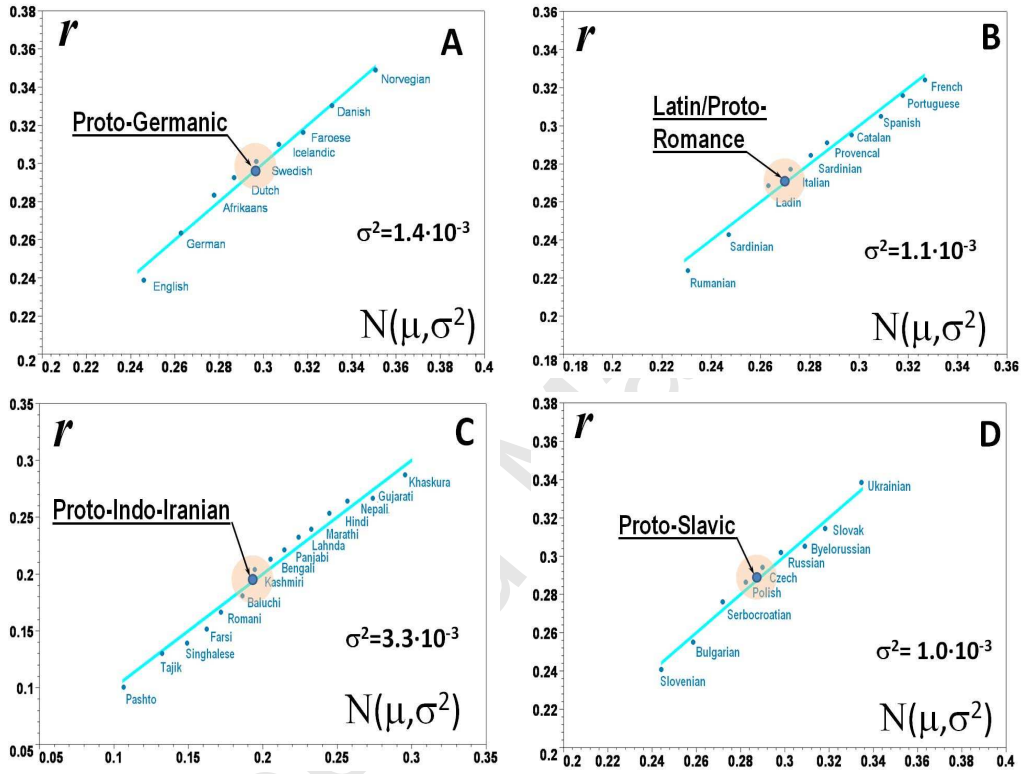


Figure 3: The panels A-D show the normal probability plots fitting the distances r of language points from the 'center of mass' to univariate normality. The data points were ranked and then plotted against their expected values under normality, so that departures from linearity signify departures from normality. The values of variance are given for each language group. The expected locations of the proto-languages, together with the end points of the 95% confidence intervals, are displayed on the normal plots by circles.

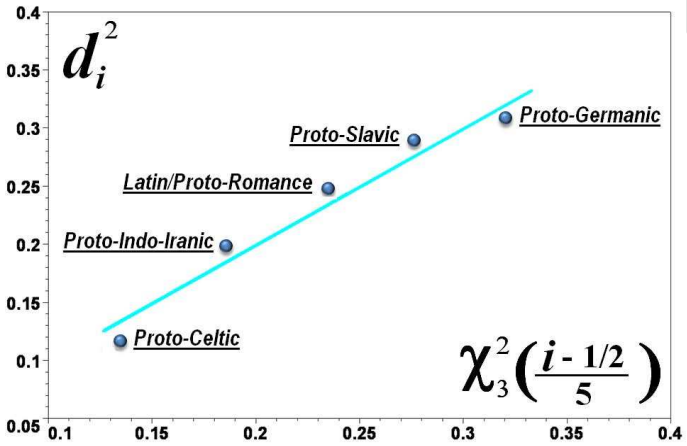


Figure 4: The graphical test to check three-variate normality of the distribution of the distances d_i of the five proto-languages from a statistically determined central point is presented by extending the notion of the normal probability plot. The chi-square distribution is used to test for goodness of fit of the observed distribution: the departures from three-variant normality are indicated by departures from linearity.

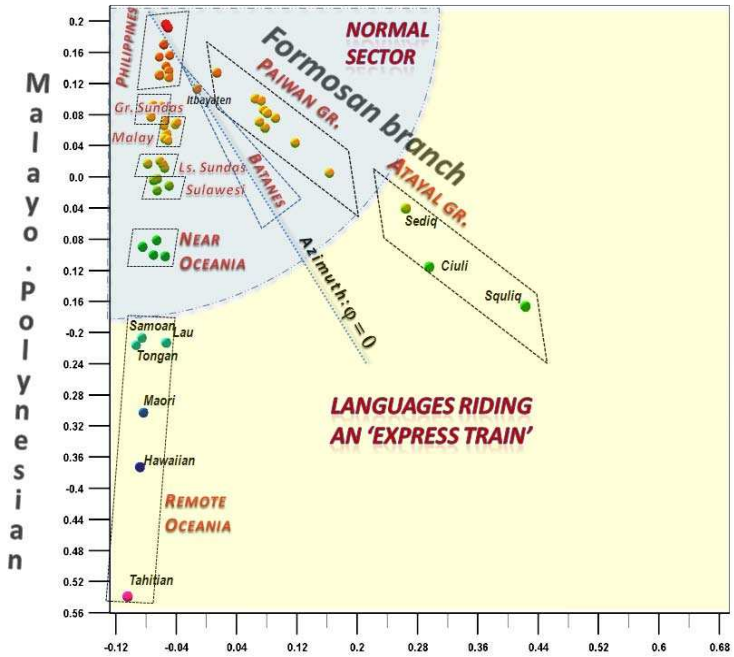


Figure 5: The geometric representation of the 50 AU languages in space of the major data traits (q_2, q_3) shows the remarkable geographic patterning. It is convenient to use the polar coordinates: the radius from the center of the graph, $r_i = \sqrt{q_{2,i}^2 + q_{3,i}^2}$, and the azimuth angle $\varphi = \arctan\left(\frac{q_{3,i}}{q_{2,i}}\right)$, to identify the positions of languages. For languages in the 'normal sector', the distribution of radial coordinates conforms to univariate normality. At variance with them, languages located at the distant margins of the AU family apparently follow the 'express train' evolution model (see Sec. 11) The 'normal sector' consists of the following languages: from Philippines, *Bontoc, Kankanay, Ilokano, Hanunoo, Cebuano, Tagalog, Pangasinan, Mansaka, Maranao*; from Great Sunda and Malay, *Malagasy, Maanyan, Ngaiu dayak, Toba batak, Bali, Malay, Iban, Sasak, Sunda, Javanese*; from Lesser Sunda and Sulawesi, *Sika, Kambera, Wolio, Baree, Buginese, Manggarai, Sangir, Makassar*; from Near Oceania, *Manam, Motu, Nggela, Mota*; of Paiwan group (Taiwan) *Pazeh, Thao, Puyuma, Paiwan, Bunun, Amis, Rukai, Siraya, Kavalan.*

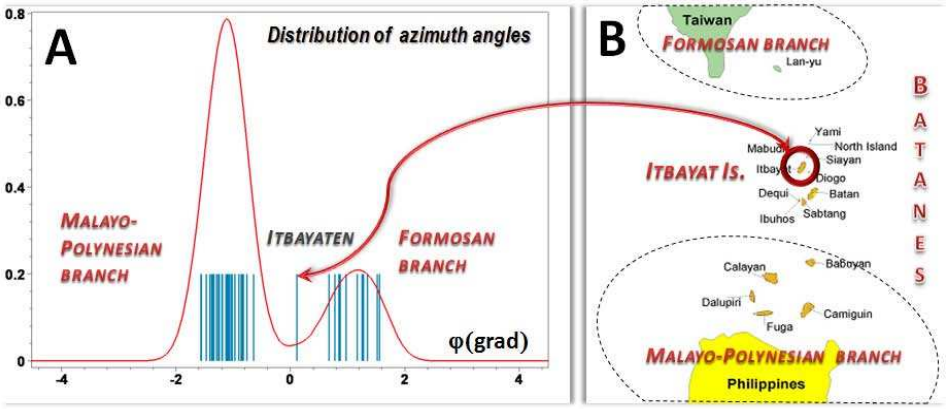


Figure 6: A.) The distribution of azimuth angles in the geometric representation of the 50 AU languages shown in Fig. 5. B.) The Itbayaten language is pretty close to the azimuth, $\varphi = 0$, bridging over the language family branches lexically and geographically.

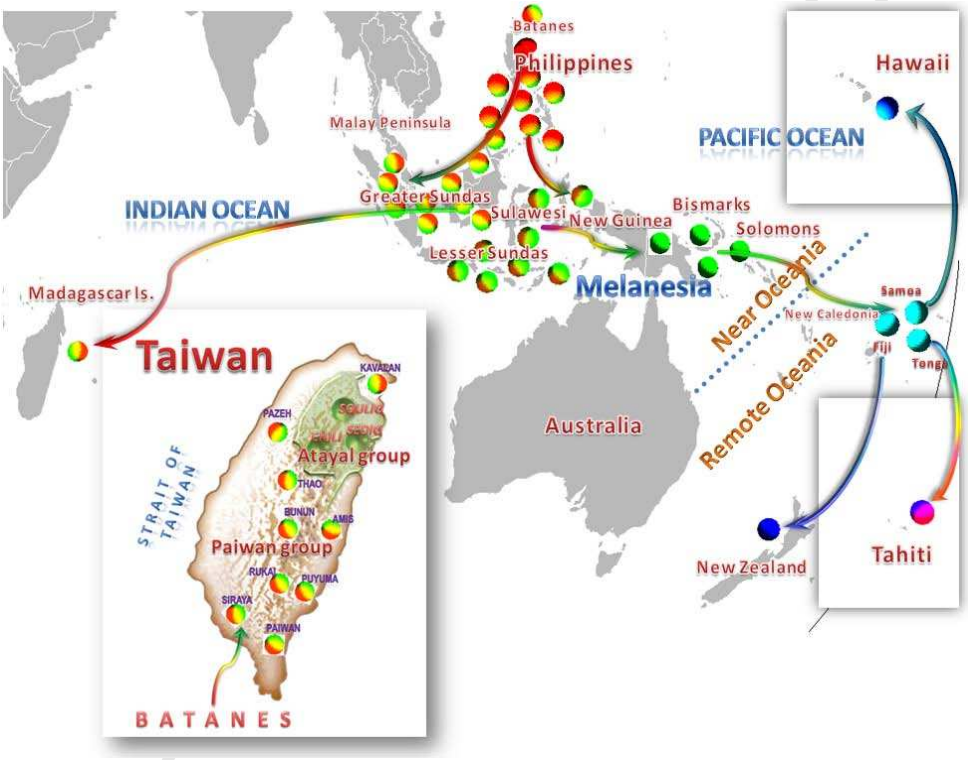


Figure 7: The geometric representation of the 50 AU languages (Fig. 5) projected onto the geographic map uncovers the possible route of Austronesian migrations.

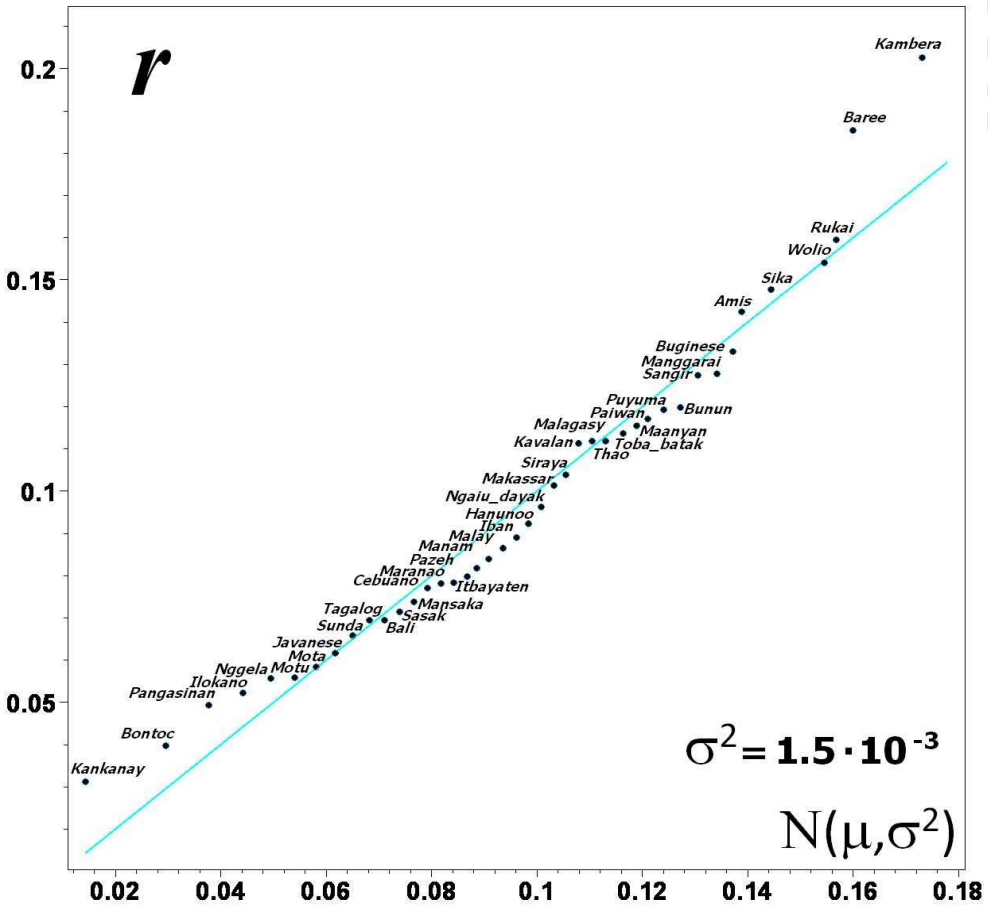


Figure 8: The normal probability plot fitting the distances r of language points from the 'center of mass'of the geometrical representation of the AU language family to univariate normality. The data points for languages belonging to the 'normal sector' shown in Fig. 5 were ranked and then plotted against their expected values under normality, so that departures from linearity signify departures from normality. The value of variance over all languages belonging to the 'normal sector' is $\sigma^2 = 1.5 \cdot 10^{-3}$.

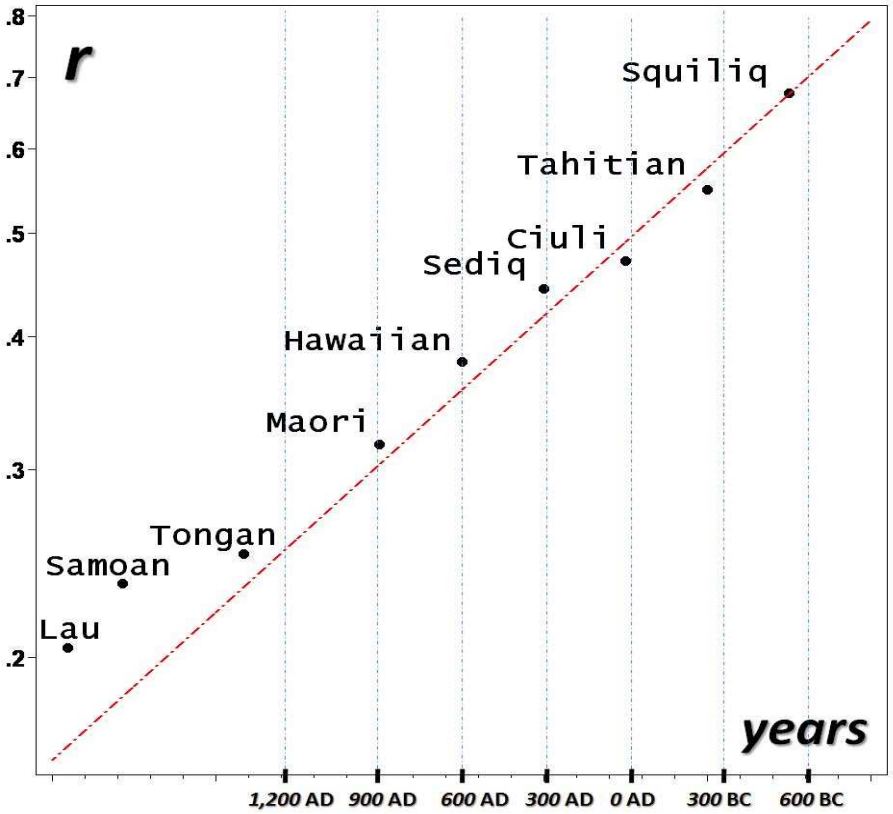


Figure 9: The log-linear plot fitting the distances r to remote languages riding an 'express train' in the geometric representation (see Fig. 5) to an exponential distribution. The radial coordinates of the languages were ranked and then plotted against their expected values under the exponential distribution. As usual, the departures from linearity signify departures from the tested distribution (given by the dash-dotted line).

Figure 10: Professor Dr. Philippe Blanchard, Bielefeld University (Germany).

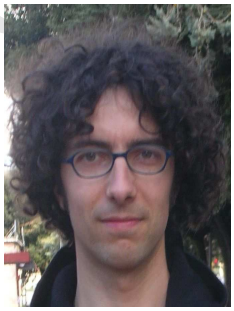


Figure 11: Dr. Filippo Petroni, the Postdoctoral fellow at the University of Rome "La Sapienza" (Italy).

Figure 12: Dr. Maurizio Serva, the Researcher in Mathematical Physics, University of L'Aquila (Italy).

Figure 13: D.Sc.(habil.) Dimitri Volchenkov, the researcher at the Center of Excellence Cognitive Interaction Technology (Bielefeld, Germany).