Nonlinear Systems and Complexity
Series Editor: Albert C. J. Luo

Hernán González-Aguilar
Edgardo Ugalde *Editors*

Nonlinear Dynamics New Directions

Models and Applications

Springer

Nonlinear Systems and Complexity

Volume 12

Nonlinear Systems and Complexity provides a place to systematically summarize recent developments, applications, and overall advance in all aspects of nonlinearity, chaos, and complexity as part of the established research literature, beyond the novel and recent findings published in primary journals. The aims of the book series are to publish theories and techniques in nonlinear systems and complexity; stimulate more research interest on nonlinearity, synchronization, and complexity in nonlinear science; and fast-scatter the new knowledge to scientists, engineers, and students in the corresponding fields. Books in this series will focus on the recent developments, findings and progress on theories, principles, methodology, computational techniques in nonlinear systems and mathematics with engineering applications. The Series establishes highly relevant monographs on wide ranging topics covering fundamental advances and new applications in the field. Topical areas include, but are not limited to: Nonlinear dynamics Complexity, nonlinearity, and chaos Computational methods for nonlinear systems Stability, bifurcation, chaos and fractals in engineering Nonlinear chemical and biological phenomena Fractional dynamics and applications Discontinuity, synchronization and control

More information about this series at http://www.springer.com/series/11433

Hernán González-Aguilar • Edgardo Ugalde
Editors

Nonlinear Dynamics New Directions

Models and Applications

Springer

Editors

Hernán González-Aguilar
Autonomous University of San Luis Potosí
San Luis Potosi
Mexico

Edgardo Ugalde
Autonomous University of San Luis Potosí
San Luis Potosi
Mexico

ISSN 2195-9994 ISSN 2196-0003 (electronic)
ISBN 978-3-319-09863-0 ISBN 978-3-319-09864-7 (eBook)
DOI 10.1007/978-3-319-09864-7
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014951913

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Levenshtein's Distance for Measuring Lexical Evolution Rates

Filippo Petroni, Maurizio Serva and Dimitri Volchenkov

Abstract The relationships between languages molded by extremely complex social, cultural and political factors are assessed by an automated method, in which the distance between languages is estimated by the average normalized Levenshtein distance between words from the list of 200 meanings maximally resistant to change. A sequential process of language classification described by random walks on the matrix of lexical distances allows to represent complex relationships between languages geometrically, in terms of distances and angles. We have tested the method on a sample of 50 Indo-European and 50 Austronesian languages. The geometric representations of language taxonomy allow for making accurate interfaces on the most significant events of human history by tracing changes in language families through time. The Anatolian and Kurgan hypothesis of the Indo-European origin and the "express train" model of the Polynesian origin are thoroughly discussed.

1 Introduction

The evolution of languages goes on like to haploid evolution for asexual organisms, as evolving reproduction, mutation and extinction. Hypotheses concerning their relationships can be verified provided a distance between languages is evaluated from the lexical differences, in analogy with the genetic distance between species.

D. Volchenkov (✉)
Cognitive Interaction Technology—Center of Excellence, Universität Bielefeld,
Postfach 10 01 31, 33501 Bielefeld, Germany
e-mail: volchenk@physik.uni-bielefeld.de

M. Serva
Dipartimento di Matematica, Università dell'Aquila,
67010 L'Aquila, Italy
e-mail: serva@univaq.it

F. Petroni
Dipartimento di Scienze Economiche ed Aziendali Università di Cagliari V.le S. Ignazio,
17 09123 Cagliari Italy
e-mail: fpetroni@unica.it

H. González-Aguilar, E. Ugalde (eds.), *Nonlinear Dynamics New Directions,*
Nonlinear Systems and Complexity 12, DOI 10.1007/978-3-319-09864-7_10

The idea to assess the dissimilarity between languages using vocabulary, has its roots in the work of the French explorer Dumont D'Urville, who collected comparative lists of 115 basic terms from various languages during his voyages aboard the Astrolabe from 1826 to 1829 and introduced the idea of measuring the similarity between words with the same meaning in his work about the geographical division of the Pacific [1]. The method used by modern glottochronology developed by Swadesh [2] estimates the distance between languages from the percentage of shared *cognates* (words inferred to have a common historical origin) assuming that vocabularies change at a constant average rate. However, the identification of cognates is often a matter of sensibility and personal knowledge, as they do not necessarily look similar, so that the task of counting the number of cognate words shared by the two languages is difficult. For instance, the Spanish word *leche* and the Greek word *gala* are cognates. In fact, *leche* comes from the Latin *lac* with genitive form *lactis*, while the genitive form of *gala* is *galactos*. This identification became possible because of our historical records that are hardly available for languages of Central Africa, Australia or Polynesia. Moreover, the comparison of languages over a large vocabulary is only apparently more accurate, as many similar words rather carry information about complex social, cultural and political factors molding the extreme historical contacts, than about the actual language similarity. It is also worth a mention that languages belonging to the same family may not share many words in common, while languages of two distinct families may share many. For instance, Brahui spoken in Pakistan, Afghanistan and Iran is a Dravidian language accordingly to its syntactic structure, despite of 85 % of its vocabulary being Indo-European (IE). Eventually, the rates of lexical changes in words are all different, being probably related to the frequency of use of the associated meanings [3]; those words with a high rate of changes might be worthless for inferring the language relatedness.

Summarizing, the successful application of phylogenetic methods to language evolution requires:

1. A distance accumulating the differences in systematic sound correspondences between the realizations of individual meanings;
2. A well-adjusted input vocabulary exhibiting uniformly high stability of items, with respect to the defined distance;
3. A suitable agglomerative clustering technique that maps the matrix of lexical distances calculated over the optimized vocabulary into low-dimensional space of language groups;
4. A plausible hypothesis on the dynamical process of language evolution that evolves the obtained geometric representation of language taxonomy in time.

In our work, we consequently fulfill the outlined program and apply it to the study of the language evolution in the IE and Austronesian (AU) language families that allows us for making accurate inferences on the most significant events of human history by tracking changes in language families through time.

2 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 [4]. 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 has been introduced in information theory [5]. In our work, being guided by [6, 7], 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*,

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

where $\|w_1, w_2\|_L$ 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, according 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 distance (1) is symmetric, $D\left(w_1, w_2\right) = D\left(w_2, w_1\right)$, and takes values between 0 and 1 for any two words, w_1 and w_2, so that $D\left(w, w\right) = 0$, and $D\left(w_1, w_2\right) = 1$ when all characters in these words are different. The normalized edit distance between the orthographic realizations of two words can be interpreted as the probability of mismatch between two characters picked from the words at random.

Given the short list $\mathcal{L}$ containing $|\mathcal{L}| = M$ meanings, we define the lexical distance between the two languages, l_1 and l_2, as the average of the normalized Levenshtein distance (1)—the smaller the result is, the more affine are the languages,

$$d\left(l_1, l_2\right) \;=\; \frac{1}{M} \cdot \sum_{\alpha \in \mathcal{L}} D\left(w_\alpha^{(l_1)}, w_\alpha^{(l_2)}\right), \tag{2}$$

where α is a meaning from the list $\mathcal{L}$, and $w_\alpha^{(l)}$ is its orthographic realization in the language l. The distance (2) is symmetric, $d(l_1, l_2) = d(l_2, l_1)$, $d(l, l) = 0$, and $d(l_1, l_2) = 1$ if and only if none of words of the $\mathcal{L}$ meanings in the language l_1 has any common character with those words 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 $\mathcal{L}$. As a result, for the two samples of 50 languages selected from the IE and AU language families, we obtained the two symmetric 50×50 matrices; each matrix therefore contains 1225 independent entries. The phylogenetic trees from the lexical distance matrices (2) were constructed and discussed in [6, 7].

3 The Short List of Meanings and Its Stability

Following [8], we define the stability of the meaning α over a sample of N languages by

$$S(\alpha) \;=\; 1 - \frac{2}{N(N-1)} \sum_{l_i > l_j} D\left(w_\alpha^{(l_i)}, w_\alpha^{(l_j)}\right) \tag{3}$$

where the sum goes over all ordered pairs (l_i, l_j) of languages in the sample. With this definition, $S(\alpha)$ takes a value between 0 and 1. The sum in the RHS. of (3) is smaller for those words corresponding to meanings with a lower rate of lexical evolution, since they tend to remain more similar in two languages. Therefore, to a larger $S(\alpha)$ there corresponds a greater stability.

We computed the stability values for the 200 meanings according to the original choice of Swadesh [2] for the 50 language samples of both language families. The main source for the database for the IE group was the file prepared by Dyen et al. [9]. This database contains Swadesh's vocabulary with basic 200 meanings which seem maximally resistant to change, including borrowing [10], for 96 languages. The words are given there without diacritical symbols 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 [9] 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 [11] has been used under the author's 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 [11] contain more than 200 meanings that do not completely coincide with those in the original Swadesh list. For our choice of 50 AU languages, we have retained only those words which are included in the both data sets of [9] and of the original vocabulary [2, 9]. The resulting list has still many gaps due to missing words in the data set [11] and incomplete overlap between the list of [11] and the original Swadesh list [2, 9]. 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 [9]. In [11], 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 diacritical symbols favouring a "simple" alphabet allowed us to obtain a reasonable result. The database modified by the authors is available [12]. Readers are welcome to modify, correct and add words to the database.

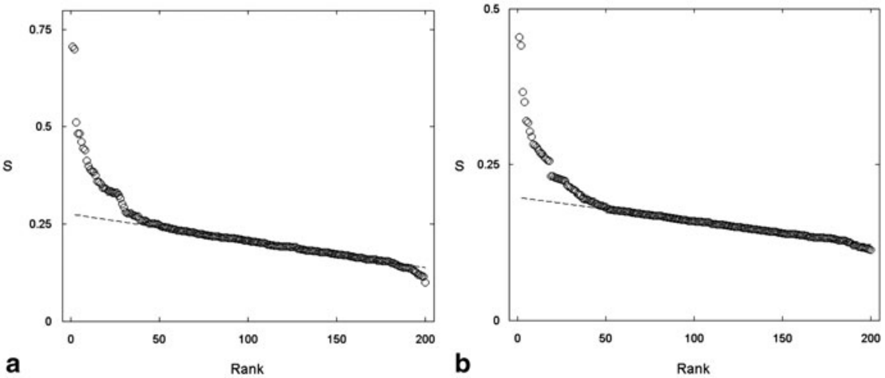

Fig. 1 Stability in a decreasing ranking for the 200 meanings over the 50 languages samples from the IE (**a**) and AU (**b**) language languages. The s*traight line* between position *51* and position *180* underlines the initial and final deviations from the linear behaviour

In Fig. 1, we have shown the plot of ranked stability values $S(\alpha)$ calculated for the 200 meanings in the short lists, for the IE and AU language groups in [13]. At the beginning the stability values drop rapidly; then, between the 50th position and the 180th, it decreases slowly and almost linearly with rank; finally at the end the stability drops again. This behaviour is not Gaussian, since high and low stability parts of the curve are not symmetric. The curve is fitted by a straight line to highlight the initial and final deviations from linearity. Clearly, one should keep all the meanings with higher information, take at least some of the most stable meanings in the linear part of the curve and exclude completely those meanings with lower information. The correlation coefficient between the stability index computed for the two groups is roughly 0.21 [13] suggesting that the stability of items in the short list depends strongly on the studied family. In order to understand whether the most stable terms in the two short lists show a large coincidence, we considered the first n items in the ranking list for both families, and we computed the number $m(n)$ of common items in the two lists. To underline the non-causal behaviour, $m(n)$ has to be compared with n^2/N, which is the average number of common items if one randomly chooses n items from any of the two lists. Then, it is natural to define $p(n)$ as $m(n)$ divided by n^2/N. If there is no relation between stability in the two families, $p(n)$ must be close to 1 for every n. The behaviour of $p(n)$ as a function of n can be seen in Fig. 2 indicating that indeed there is a non-trivial overlapping of the two lists of most stable n items since $p(n)$ is always larger than 1. This fact confirms the correlation between the two rankings, and also shows that this effect is strong only for small n ($n < 50$). For larger n, the overlapping is much closer to 1 and random coincidences prevail. This means that the most stable terms in the two lists are those that show a larger coincidence.

To give an example of the lists found with our approach, we show here a table of the 20 most stable items for the IE and AU language groups. Together with any of the items we report its stability record within the family.

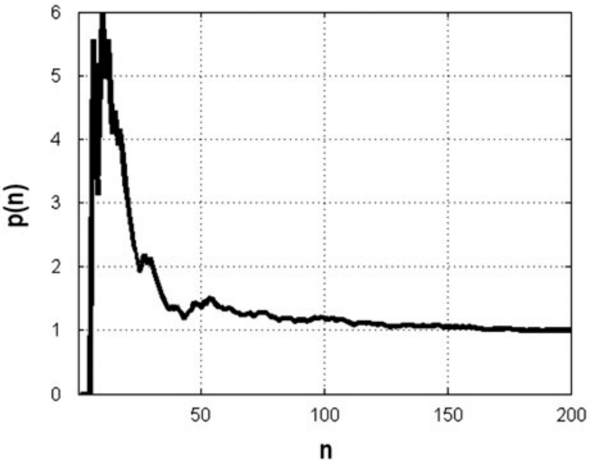

Fig. 2 The number of common items in the two lists of most stable n items obtained for the Austronesian and Indo-European families. The number is normalized by the random coincidence $n^2/200$

Table 1 The 20 most stable words for the Indo-European and Austronesian language families, with their stability values within the family

Indo-European	Stability	Austronesian	Stability
YOU	0.45395	EYE	0.70646
THREE	0.44102	FIVE	0.70089
MOTHER	0.36627	FATHER	0.51095
NOT	0.35033	DIE	0.48157
NEW	0.31961	STONE	0.48157
NOSE	0.3169	THREE	0.46087
FOUR	0.30226	TWO	0.44411
NIGHT	0.29403	LOUSE	0.43958
TWO	0.28214	ROAD	0.41217
NAME	0.27962	FOUR	0.39798
TOOTH	0.27677	HAND	0.38997
STAR	0.27269	NAME	0.38493
SALT	0.26792	LIVER	0.38375
DAY	0.26695	PUSH	0.37444
GRASS	0.26231	MOTHER	0.35821
SEA	0.25906	WE	0.35749
DIE	0.25602	EAT	0.3529
SUN	0.25535	STICK	0.34242
ONE	0.23093	I	0.34208
FEATHER	0.23055	VOMIT	0.33861

4 The Structural Component Analysis of Linguistic 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) [14] and principal component analysis (PCA) [15] are widely used for separating a multivariate signal into additive subcomponents. However, it is clear that these standard techniques of component analysis have to be dramatically improved for any meaningful application on language data, as there is no reason to suggest neither that the directions of maximum variance recovered by the standard PCA method are good enough for identification of principal components in the linguistic data, nor that the language traits are statistically independent. 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 pairwise 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. A suitable method for the structural component analysis (SCA) of networks (weighted graphs) by means of *random walks* (or Markov chains, in a more general context) has been formulated in [16–18]. Being a version of the *kernel PCA* method [19], it generalizes PCA to the case, where we are interested in principal components obtained by taking all higher-order correlations between data instances. The SCA method has been successfully applied to the analysis of language taxonomies in [20].

Let us note that there are infinitely many matrices that match all the structure of $d(l_i, l_j)$ 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,

$$T\left(l_i, l_j\right) \;=\; \Delta^{-1}\, d\left(l_i, l_j\right), \tag{4}$$

defined on the weighted undirected graph determined by the matrix of lexical distances $d(l_i, l_j)$, The diagonal matrix in (4) $\Delta = \mathrm{diag}(\delta_{l_1}, \delta_{l_2}, \ldots \delta_{l_N})$ contains the *cumulative* lexical distances $\delta_{l_i} = \sum_{j=1}^{N} d(l_i, l_j)$, 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 (4) is a stochastic matrix, $\sum_{j=1}^{N} T(l_i,l_j) = 1$, being nothing else, but the normalized matrix of lexical distances (2). Random walks defined by the transition matrix (4) describe the statistics of a sequential process of language classification. Namely, while the elements of the matrix $T(l_i,l_j)$ evaluate the probability of successful differentiation of the language l_i provided the language l_j has been identified certainly, the elements of the squared matrix T^2, ascertain the successful differentiation of the language l_i from l_j through an intermediate language, the elements of the matrix T^3 give the probabilities to differentiate the language through two intermediate steps, and so on. The whole host of complex and indirect relationships between orthographic representations of the vocabulary meanings encoded in the matrix of lexical distances (2) is uncovered by the von Neumann series estimating the characteristic time of successful classification for any two languages in the database over a language family,

$$J\left(l_i,l_j\right) \;=\; \lim_{n\to\infty}\sum_{k=0}^{n} T^n\left(l_i,l_j\right) \;=\; \frac{1}{1-\mathbf{T}}. \tag{5}$$

The last equality in (5) is understood as the group generalized inverse [20], being a symmetric, positive semi-definite matrix which plays essentially the same role for the SCA, as the covariance matrix does for the usual PCA analysis. The standard goal of a component analysis (minimization of the data redundancy quantified by the off-diagonal elements of the kernel matrix) is readily achieved by solving an eigenvalue problem for the matrix $J(l_i,l_j)$. Each column vector q_k, which determines a direction where $\mathbf{J}$ acts as a simple rescaling, $\mathbf{J}q_k = \lambda_k q_k$, with some real eigenvalue $\lambda_k = 0$, is associated to the virtually independent trait in the matrix of lexical distances $d(l_i,l_j)$. Independent components $\{q_k\}$, $k = 1,\dots N$, define an orthonormal basis in $\mathbb{R}^N$ which specifies each language l_i by N numerical coordinates, $l_i \to (q_{1,i}, q_{2,i},\dots q_{N,i})$. 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, $\lambda_0 = \lambda_1 < \lambda_2 = \dots = \lambda_N$, provides us with the natural geometric framework for dimensionality reduction. At variance with the standard PCA analysis [15], where the largest eigenvalues of the covariance matrix are used in order to identify the principal components, 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. Since the minimal eigenvalue $\lambda_1 = 0$ corresponds to the vector of stationary distribution of random walks and thus contains no information about components, we have used 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 the vocabulary meanings in these languages are similar.

5 Geometric Representation of the IE 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. 3, 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,

$$r_i = \sqrt{q_{2,i}^2 + q_{3,i}^2 + q_{4,i}^2}, \quad \theta_i = \arccos\left(\frac{q_{4,i}}{r_i}\right), \quad \phi_i = \arctan\left(\frac{q_{3,i}}{q_{2,i}}\right), \tag{6}$$

rather than the Cartesian system.

The principal components of the IE family reveal themselves in Fig. 3 by four well-separated spines representing the four biggest traditional IE language groups: Romance and 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. 4a and b, respectively.

The Greek, Romance, Celtic and Germanic languages form a class characterized by approximately the same azimuthal angle (Fig. 4a), thus belonging to one plane in the three-dimensional geometric representation shown in Fig. 3, while the Indo-Iranian, Balto-Slavic, Armenian and Albanian languages form another class, with respect to the inclination (zenith) angle (Fig. 4b).

It is remarkable that the division of IE languages with respect to the azimuthal and zenith angles evident from the geometric representation in Fig. 3 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 [21]. 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 IE languages [22], it is not accorded much significance by modern linguists as being just one of many other isoglosses crisscrossing all IE languages [23]. 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 centre of the diagram in Fig. 3. Being eloquently different from others, these languages can be

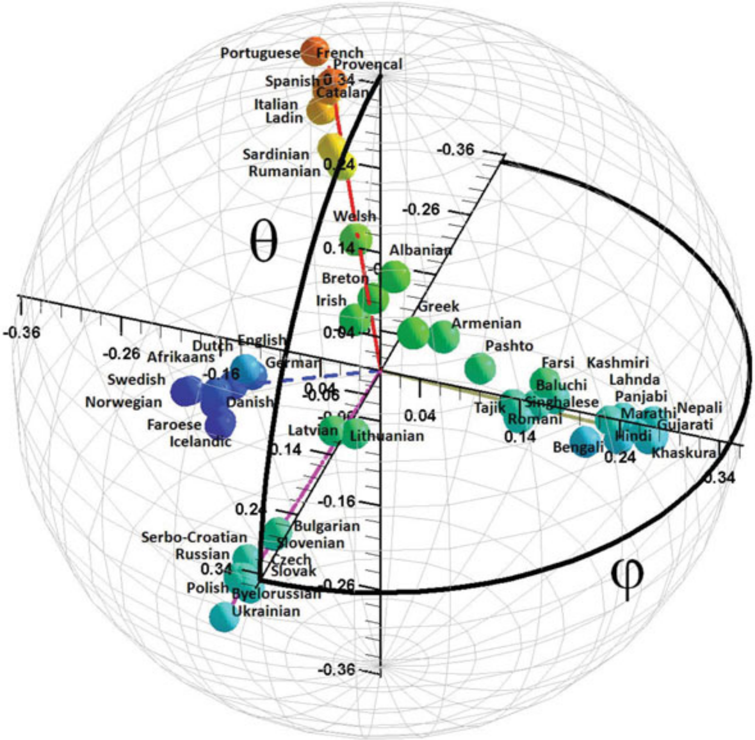

Fig. 3 The three-dimensional geometric representation of the Indo-European language family in space of the major data traits (q_2, q_3, q_4) colour coded. The *origin* of the graph indicates the centre of mass, $q_1 = \pi$, of the matrix of lexical distances $d(l_i, l_j)$, not the Proto-Indo-European 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 *centre* of the graph, the inclination angle θ and the azimuth angle φ

resolved with the use of some minor components q_k, $k > 3$. Remarkably, the Greek and Armenian languages always remain proximate confirming the Greeks' belief that their ancestors had come from Western Asia [24].

6 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 [2], 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

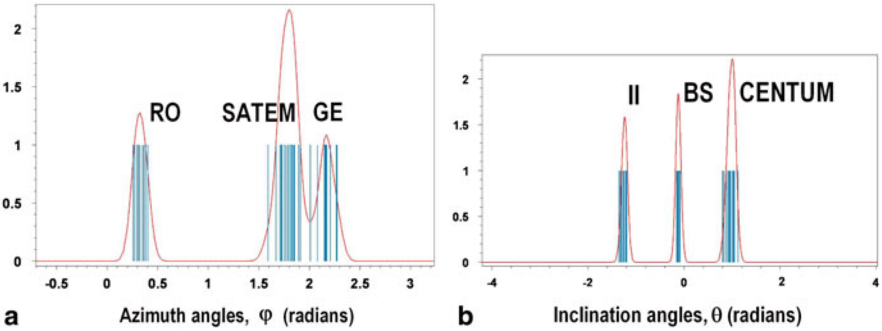

Fig. 4 **a** The kernel density estimates of the distributions of azimuthal angles in the three-dimensional geometric representation of 50 languages of the Indo-European language family, together with the absolute data frequencies. Romance (*RO*), Germanic (*GE*) and the satem languages (*SATEM*) are 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 Indo-European 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

measuring the age of organic materials was rejected by mainstream linguists, considering a language as a social phenomenon driven by unforeseeable sociohistorical events not stable over time. 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 (6)) from the origin of the graph shown in Fig. 3, 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 varying 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. 3, show very good agreement with univariate normality, as seen from the normal probability plots in Fig. 5a–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.

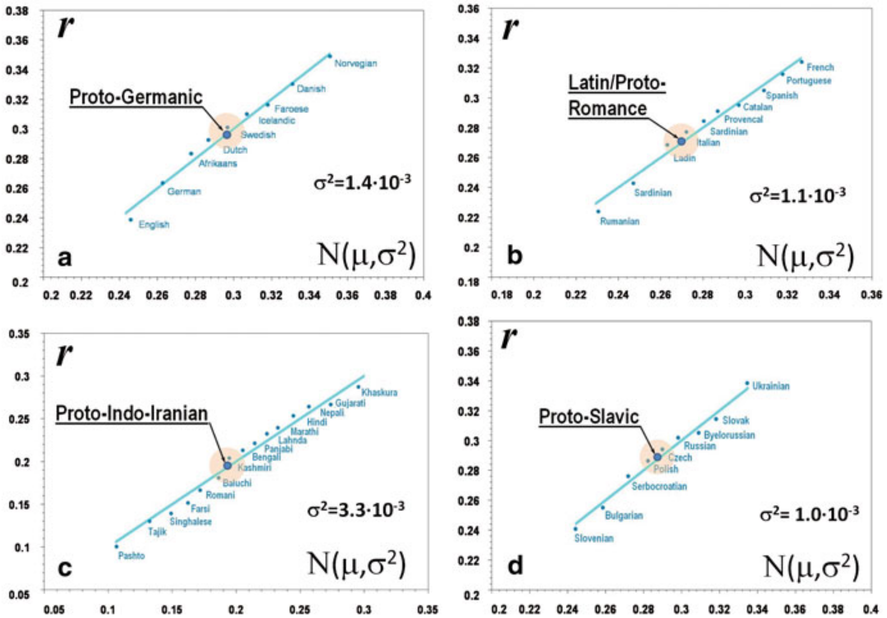

Fig. 5 The *panels A–D* show the normal probability plots fitting the distances r of language points from the centre of mass to univariate normality. The data points were ranked and then plotted against their expected values under normality, so that departures from linearity signify departures from normality. 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*

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. The constant increment rates of variance of radial positions of languages in the geometrical representation, Fig. 3, has nothing to do with the traditional glottochronological assumption about the steady borrowing rates of cognates [25]. It is also important to mention that the values of variance σ^2 calculated for the languages over the individual language groups (see Fig. 5a–d) do not correspond to physical time 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 behaviour, 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 [26]. We do not aspire to put dates on such a fuzzy process, 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 [27]:

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

$$\frac{t}{\sigma^2} = (1.367 \pm 0.002) \times 10^6 \tag{7}$$

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

The time–variance ratio (7) 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 2400 BC, in a good agreement with the migration dates from the early Andronovo archaeological horizon [28]
2. The end of common Balto-Slavic history as early as by 1400 BC, in support of the recent glottochronological estimates [29] well agreed with the archaeological dating of Trziniec–Komarov culture, localized from Silesia to Central Ukraine
3. The separation of Indo-Aryans from Indo-Iranians by 400 BC, probably as a result of Aryan migration across India to Ceylon, as early as in 483 BC [30]
4. The division of Persian polity into a number of Iranian tribes migrated and settled in vast areas of Southeastern Europe, the Iranian plateau and Central Asia by 400 BC, shortly after the end of Greco–Persian wars [31].

7 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 so far. For instance, the *Kurgan* scenario postulating that the people of an archaeological "Kurgan culture" (early fourth millennium BC) in the Pontic steppe were the most likely speakers of the proto- IE language, is widely accepted [32]. 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 eighth and sixth millennia BC [22].

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 re-examined 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. 5a–d. Let us note that we did not include the Baltic languages into the Slavic group when computing the Proto-Slavic centre 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 behaviour 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.

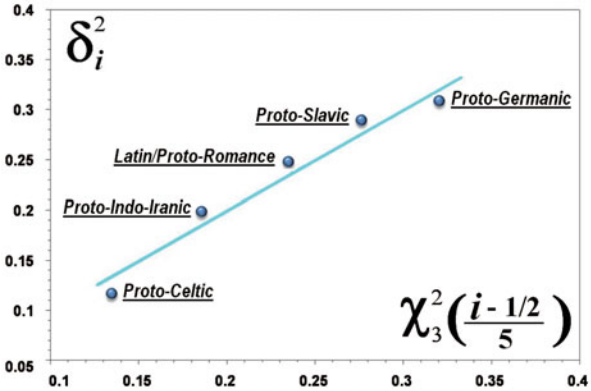

Fig. 6 The graphical test to check three-variate normality of the distribution of the distances δ_i of the five proto-languages from a statistically determined central point is presented by extending the notion of the normal probability plot. The integer parameter i specifies the number of degrees of freedom. 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

In Fig. 6, 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. 6, 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 (7) then dates the initial break-up of the Proto-Indo-Europeans back to 7000 BC pointing at the early Neolithic date, to say nothing about geography, in agreement with the Anatolian hypothesis of the early Indo-European origin [7, 21, 22, 24, 33].

The linguistic community estimates of dating for the proto-IE language lie between 4500 and 2500 BC, a later date than the Anatolian theory predicts. These estimations are primarily based on the reconstructed vocabulary (see [34] 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 favouring 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 [35]. Let us note that our dating (2400 BC) for the migration from the Andronovo archaeological horizon (see Sect. 6) and the early break-up of the proto-Indo-Iranian continuum estimated by means of the variance (see Fig. 5c) 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.

8 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. 7).

The distribution of azimuth angles shown in Fig. 8a 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 [36], 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 [6]. As the Dayak speakers formed the majority in the initial settler group, in agreement with the genetic parental lineages found in Madagascar [37], 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. 8b). On the distribution of azimuth angles shown in Fig. 8a, 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. 8b). By the way, the MP-offset descends from the

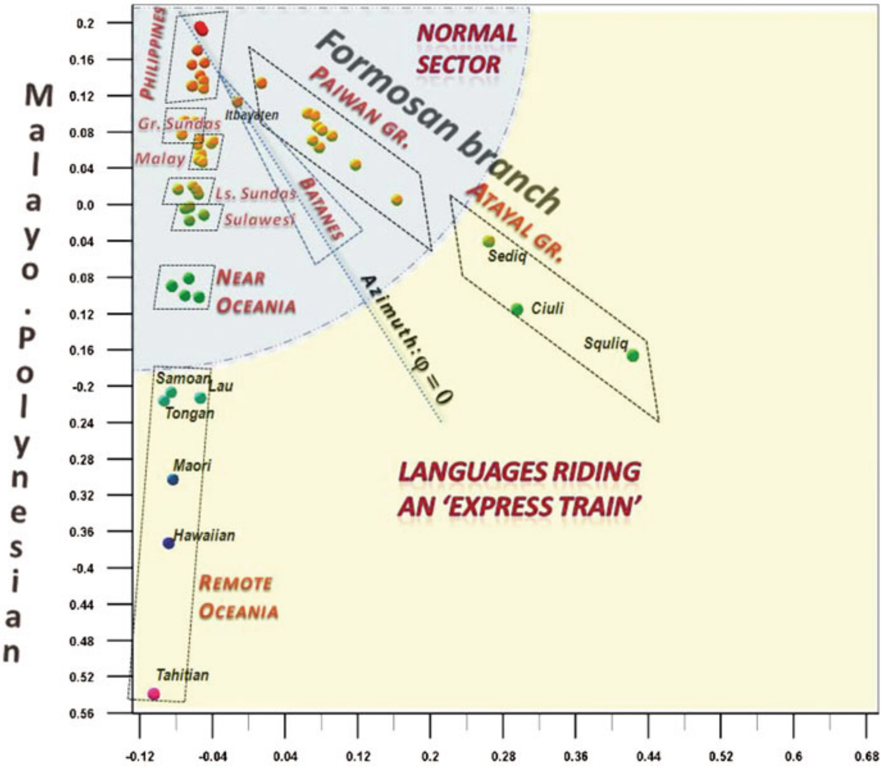

Fig. 7 The geometric representation of the 50 AU languages in space of the major data traits (q_2, q_3) shows the remarkable geographic patterning. It is convenient to use the polar coordinates: the radius from the centre of the graph, $r_i = \sqrt{q_{2,i}^2 + q_{3,i}^2}$, and the azimuth angle $\varphi = \arctan\left(\frac{q_{3,i}}{q_{2,i}}\right)$, 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 Sect. 9) 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*

northern Philippines (the northern Luzon Island) and springs forth eastward through the Malay Archipelago across Melanesia culminating in Polynesia (Fig. 9); in accordance with the famous "express train" model of migrations peopled the Pacific [38]. 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 [38], but then evolved mostly independently from

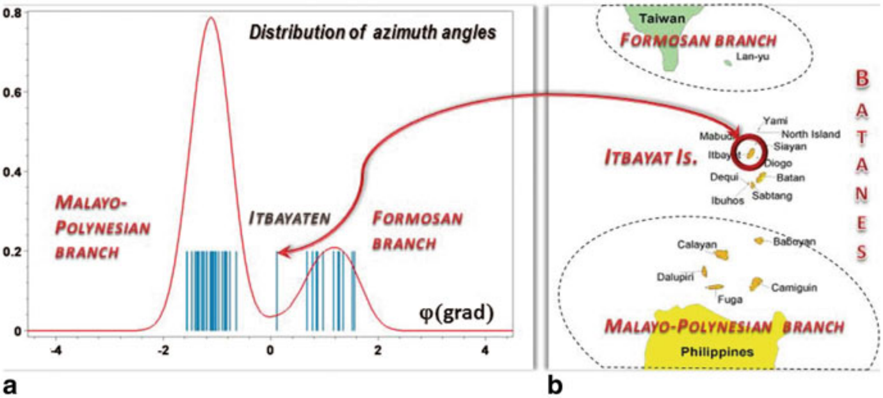

Fig. 8 **a** The distribution of azimuth angles in the geometric representation of the 50 AU languages shown in Fig. 7. **b** The Itbayaten language is pretty close to the azimuth, $\varphi = 0$, bridging over the language family branches lexically and geographically

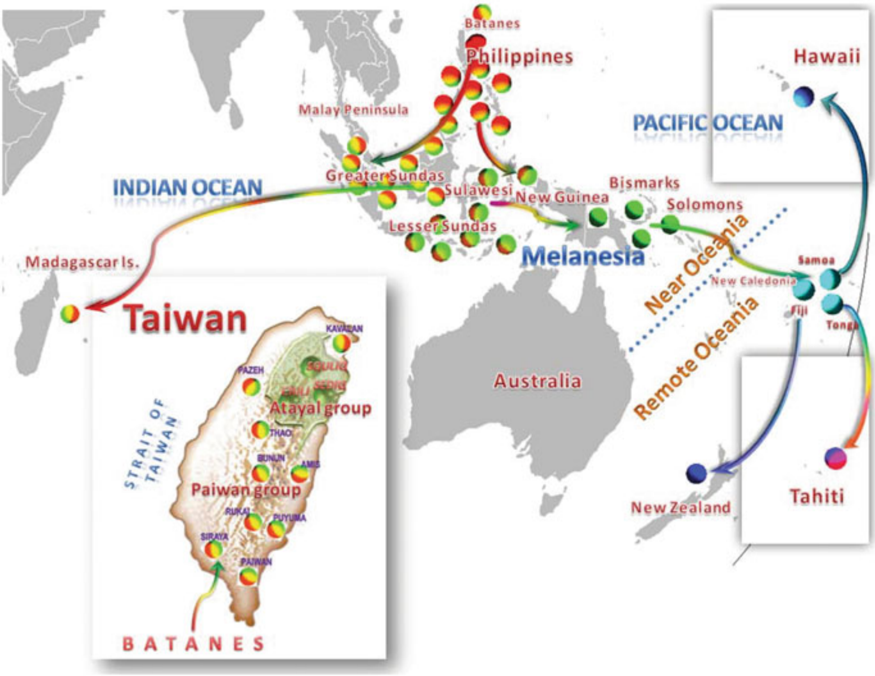

Fig. 9 The geometric representation of the 50 AU languages (Fig. 7) projected onto the geographic map uncovers the possible route of Austronesian migrations

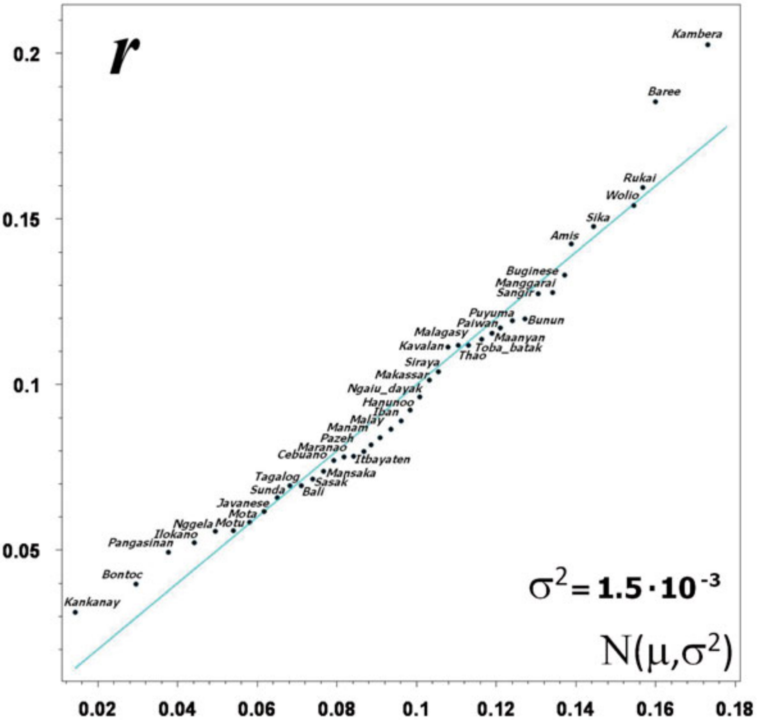

Fig. 10 The normal probability plot fitting the distances r of language points from the "centre of mass" of the geometrical representation of the AU language family to univariate normality. The data points for languages belonging to the "normal sector" shown in Fig. 7 were ranked and then plotted against 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 $\sigma^2 = 1.5 \times 10^{-3}$

each other, on evidence of the Y-chromosome haplotype spread over Taiwanese and Polynesian populations [39].

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 centre of the diagram representing the AU language family in Fig. 7 conforms to univariate normality (see Fig. 10) 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 [40, 41] and translocation of farm animals [42, 43] 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 (7) 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 600–1200 AD while descendants from Melanesia settled in the distant apices of the Polynesian triangle as evidenced by archaeological records [44–46].

9 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 centre of the geometric representation shown in Fig. 7 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 [38] consistent with the multiple evidences on comparatively reduced genetic variations among human groups in Remote Oceania [47–49].

In order to obtain reasonable chronological estimates, an alternative mechanism 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. 7) 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. 11, 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. 11).

The radial coordinates of the languages at the distant margins of the AU family diagram shown in Fig. 11 may be deduced as evolving in accordance with the simple differential equation

$$\dot{r} = ar \tag{8}$$

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

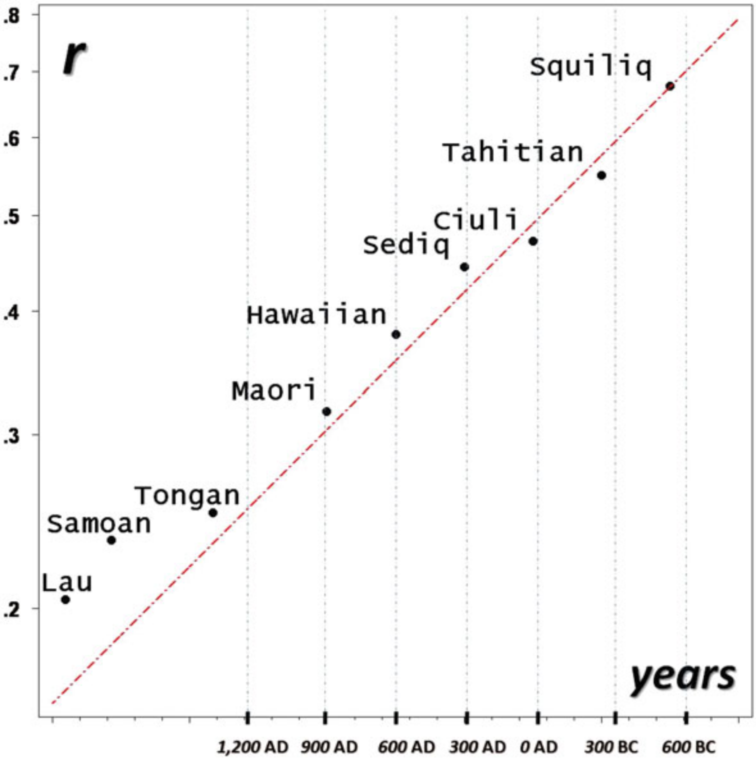

Fig. 11 The log-linear plot fitting the distances *r* to remote languages riding an "express train" in the geometric representation (see Fig. 7) to an exponential distribution. The radial coordinates of the languages were ranked and then plotted against their expected values under the exponential distribution. As usual, the departures from linearity signify departures from the tested distribution (given by the *dash-dotted line*)

train in space of the major traits of the AU family. In the proposed model of language evolution, it is suggested that in absence of contact borrowings the orthographic realizations of Swadesh's meanings would accumulate emergent variations in spellings, so that the radial coordinate indicating the divergence of a remote language from the rest of the group can grow unboundedly with isolation time.

A simple equation mathematically similar to (8) has been proposed by Swadesh [2] in order to describe the change of cognates in time, in the framework of the glottochronological approach. In our previous work [6], another similar equation has been suggested for the purpose of modeling the time evolution of normalized edit distances between languages. However, we have to emphasize that the statistical model (8) has a direct relation to neither the percentage of cognates (as in the traditional glottochronological approach), nor the edit distance itself.

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 (8) as

$$t_1 - t_2 \;=\; \frac{1}{a} \cdot \ln \frac{r_1}{r_2} \tag{9}$$

where $r_2 > r_1$ are the radial coordinates of the languages from the centre of the sample diagram shown in Fig. 7.

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 [44], the date we placed the incipience of the Tahitian society. According to many archaeological reconstructions [44–46], descendants from West Polynesia had spread through East Polynesian archipelagos and settled in Hawaii by 600 AD and in New Zealand by 1000 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

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

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. 11. 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 [50] 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 [51], the Sediq dialect subgroup could have split off from the rest of the Atayal groups about 1600 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 1750 AD [52]. 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 1000 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 1820 AD [52]. 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.

10 Conclusion

We have presented the new paradigm for the analysis of language phylogeny. The proposed method is fully automated; it avoids subjectivity since all results can be replicated by other scholars assuming that the database is the same. Furthermore, it allows for rapid comparison of items of a very large number of languages.

We applied here our method to the IE and AU families of languages considering 200 items lists of words according to the original choice of Swadesh. The output was a stability measure for all items computed separately for the two families. The ranking plots show that the two families behave in the same way, with the higher stability items deviating from the linear interpolation because of their very large values. We are convinced that this phenomenology we observe, both for IE and AU languages, should be a universal characteristic of stability distributions, common to all families. On the contrary, it turns out that the most stable items are not the same even if there is a positive correlation between the stability computed for IE and AU groups.

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 multi-dimensional 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. Recently, we have applied the developed method in order to investigate the detailed historical configuration of Malagasy dialects spoken on Madagascar [53].

Acknowledgements We profoundly thank R.D. Gray for the permission to use the Austronesian Basic Vocabulary Database [54] containing lexical items from languages spoken throughout the Pacific region.

References

1. D'Urville, D.: Sur les îles du Grand Océan. Bull. Soc. Goégr. **17**, 1–21 (1832)
2. Swadesh, M.: Lexicostatistic dating of prehistoric ethnic contacts. Proc. Am. Philos. Soc. **96**, 452–463 (1952)
3. Pagel, M., Atkinson, Q.D., Meadel, A.: Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature **449**, 717–720 (2007).
4. Nichols, J., Warnow, T.: Tutorial on computational linguistic phylogeny. Lang. Linguist. Compass **2**(5), 760–820 (2008)
5. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. Sov. Phys. Dokl. **10**, 707–710 (1966)
6. Petroni, F., Serva, M.: Language distance and tree reconstruction. J. Stat. Mech. Theory Exp. **2008**, P08012 (2008)
7. Serva, M., Petroni, F.: Indo-European languages tree by Levenshtein distance. Europhys. Lett. **81**, 68005 (2008)
8. Petroni, F., Serva, M.: Lexical evolution rates derived from automated stability measures. J. Stat. Mech. **2010**, P03015 (2010)
9. Dyen, I., Kruskal, J., Black, P.: Comparative Indo-European Database collected by Isidore Dyen. 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 (1997)
10. McMahon, A., Heggarty, P., McMahon, R., Slaska, N.: Swadesh sublists and the benefits of borrowing: An Andean case study. Trans. Philol. Soc. **103**(2), 147–170 (2005)
11. Greenhill, S.J., Blust, R., Gray, R.D.: The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics. Evol. Bioinform. **4**, 271. The Austronesian Basic Vocabulary Database. http://language.psy.auckland.ac.nz/austronesian (2008)
12. The database modified by the authors is publicly available on-line at http://univaq.it/∼serva /languages/languages.html
13. Petroni, F., Serva, M.: Automated words stability and languages phylogeny. J. Quant. Linguist. **18**(1), 53–62 (2011)
14. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley, New York (2001)
15. Jolliffe, I.T.: Principal Component Analysis, 2nd ed. Springer Series in Statistics, vol. XXIX. Springer, New York (2002)
16. Blanchard, P., Volchenkov, D.: Intelligibility and first passage times in complex urban networks. Proc. R. Soc. A **464**, 2153–2167 (2008)
17. Blanchard, P., Volchenkov, D.: Mathematical Analysis of Urban Spatial Networks. Understanding Complex Systems, vol. XIV. Springer, Berlin (2009)
18. Volchenkov, D.: 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 (2010)
19. Schölkopf, B., Smola, A.J., Müller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10**, 1299–1319 (1998)
20. Blanchard, P., Petroni, F., Serva, M., Volchenkov, D.: Geometric representations of language taxonomies. Comput. Speech Lang. **25**(3), 679–699 (2011).
21. Gamkrelidze, T.V., Ivanov, V.V.: (1995) Indo-European and the Indo-Europeans: A Reconstruction and historical analysis of a proto-language and a proto-culture. Trends in Linguistics: Studies and Monographs, vol. 80. de Gruyter Berlin
22. Renfrew, C.: Archaeology and Language: The Puzzle of Indo-European Origins. Cambridge University Press, New York (1987)
23. Baldi, P.: The Foundations of Latin. Mouton de Gruyter Series Trends in Linguistics: Studies and Monographs, vol. 117. de Gruyter, Berlin (2002)
24. Gamkrelidze, T.V., Ivanov, V.V.: The early history of Indo-European languages. Sci. Am. **262**(3), 110–116 (1990)

25. Embelton, S.M.: Statistics in Historical Linguistics. Bochum, Brockmeyer (1986)
26. Heggarty, P.: Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data and to dating language? In: Forster&, P., Renfrew, C. (eds.) Phylogenetic Methods and the Prehistory of Languages, p. 183. McDonald Institute for Archaeological Research, Cambridge (2006)
27. Fouracre, P.: The New Cambridge Medieval History. Cambridge University Press (1995–2007)
28. Bryant, E.: The Quest for the Origins of Vedic Culture: The Indo-Aryan Migration Debate. Oxford University Press (2001)
29. Novotná, P., Blažek, V.: Glottochronolgy and its application to the Balto-Slavic languages. Baltistica **XLII**(2), 185–210 (2007)
30. Mcleod, J.: The History of India. Greenwood Publishing Group (2002)
31. Green, P.: The Greco-Persian Wars. University of California Press, Berkeley (1996)
32. Gimbutas, M.: Old Europe in the fifth millenium B.C.: The European situation on the arrival of Indo-Europeans. In: Polomé, E.C. (ed.) The Indo-Europeans in the Fourth and Third Millennia. Karoma Publishers, Ann Arbor (1982)
33. Renfrew, C.: Time depth, convergence theory, and innovation in proto-Indo-European. Proceedings of the conference languages in prehistoric Europe, p. 227, Eichstätt University, 4–6 October 1999, Heidelberg (2003)
34. Mallory, J.P.: In Search of the Indo-Europeans: Language, Archaeology, and Myth. Thames & Hudson, London (1991)
35. Krell, K.S.: Gimbutas' Kurgan-PIE homeland hypothesis: A linguistic critique. In: Blench, R., Spriggs, M. (eds.) Archaeology and Language, II, p. 267, London, Routledge (1998)
36. Dahl, O.C.: Avhandlinger utgitt av Egede-Instituttet **3**, 408, Arne Gimnes Forlag (1951)
37. Hurles, M.E., Sykes, B.C., Jobling, M.A., Forster, P.: The dual origins of the Malagasy in island Southeast Asia and East Africa: Evidence from maternal and paternal lineages. Am. J. Hum. Genet. **76**, 894 (2005)
38. Diamond, J.M.: Express train to Polynesia. Nature **336**, 307–308 (1988)
39. Su, B., et al.: Polynesian origins: Insights from the Y chromosome. Proc. Natl. Acad. Sci. U S A **97**(15), 8225–8228 (2000)
40. Bellwood, P., Koon, P.: Lapita colonists leave boats unburned! Antiquity **63**(240), 613–622 (1989)
41. Kirch, P.V.: The Lapita Peoples: Ancestors of the Oceanic World. Blackwell, Cambridge (1997)
42. Matisoo-Smith, E., Robins, J.H.: Origins and dispersals of Pacific peoples: Evidence from mtDNA phylogenies of the Pacific rat. Proc. Natl. Acad. Sci. U S A **101**(24), 9167–9172 (2004)
43. Larson, G., et al.: Phylogeny and ancient DNA of Sus provides insights into neolithic expansion in island Southeast Asia and Oceania. Proc. Natl. Acad. Sci. U S A **104**(12), 4834–4839 (2007)
44. Kirch, P.V.: On the road of the winds: An archaeological history of the Pacific islands before European contact. University of California Press, Berkley (2000)
45. Anderson, A., Sinoto, Y.: New radiocarbon ages for colonization sites in East Polynesia. Asian Perspect. **41**, 242–257 (2002)
46. Hurles, M.E., et al.: Untangling Pacific settlement: The edge of the knowable. Trends Ecol. Evol. **18**, 531–540 (2003)
47. Lum, J.K., Jorde, L.B., Schiefenhovel, W.: Affinities among Melanesians, Micronesians, and Polynesians: A neutral, biparental genetic perspective. Hum. Biol. **74**, 413–430 (2002)
48. Kayser, M., et al.: Melanesian and Asian origins of polynesians: mtDNA and Y chromosome gradients across the Pacific. Mol. Biol. Evol. **23**, 2234–2244 (2006)
49. Friedländer, J.S., et al.: Genetic structure of Pacific islanders. PLoS Genet. **4**(1), e19 (2008) (Public Library of Science)
50. Utsurikawa, N.: A Genealogical and Classificatory Study of the Formosan Native Tribes. Toko shoin, Tokyo (1935)
51. Li, P.J.: Types of lexical derivation of men's speech in Mayrinax. Bull. Inst. Hist. Philol. **54**(3) 1–18 (1983) (Academia Sinica)

52. Li, P.J.: The dispersal of The Formosan aborigines in Taiwan. Lang. Linguist. **2**(1), 271–278 (2001)
53. Volchenkov, D., Filippo, P., Maurizio, S., Søren, W.: Malagasy dialects and the peopling of madagascar, Journal of Royal Soc. Interface, p. 1–14, doi:10.1098/rsif.2011.0228 (2011)
54. The Austronesian Basic Vocabulary Database by R.D. Gray is publicly available on -line at http://language.psy.auckland.ac.nz/austronesian/