JOURNAL
OF
THE ROYAL
SOCIETY
Interface

*J. R. Soc. Interface* (2012) **9**, 54–67
doi:10.1098/rsif.2011.0228
*Published online* 1 June 2011

# Malagasy dialects and the peopling of Madagascar

**Maurizio Serva[1], Filippo Petroni[2,*], Dima Volchenkov[3] and Søren Wichmann[4]**

[1]*Dipartimento di Matematica, Università dell'Aquila, I-67010 L'Aquila, Italy*
[2]*Facoltà di Economia, Università di Cagliari, I-09123 Cagliari, Italy*
[3]*Center of Excellence Cognitive Interaction Technology, Universität Bielefeld, 33501 Bielefeld, Germany*
[4]*Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany*

The origin of Malagasy DNA is half African and half Indonesian, nevertheless the Malagasy language, spoken by the entire population, belongs to the Austronesian family. The language most closely related to Malagasy is Maanyan (Greater Barito East group of the Austronesian family), but related languages are also in Sulawesi, Malaysia and Sumatra. For this reason, and because Maanyan is spoken by a population which lives along the Barito river in Kalimantan and which does not possess the necessary skill for long maritime navigation, the ethnic composition of the Indonesian colonizers is still unclear. There is a general consensus that Indonesian sailors reached Madagascar by a maritime trek, but the time, the path and the landing area of the first colonization are all disputed. In this research, we try to answer these problems together with other ones, such as the historical configuration of Malagasy dialects, by types of analysis related to lexicostatistics and glottochronology that draw upon the automated method recently proposed by the authors. The data were collected by the first author at the beginning of 2010 with the invaluable help of Joselinà Soafara Néré and consist of Swadesh lists of 200 items for 23 dialects covering all areas of the island.

**Keywords: dialects of Madagascar; language taxonomy; lexicostatistic data analysis; Malagasy origins**

## 1. INTRODUCTION

The genetic make-up of Malagasy people exhibits almost equal proportions of African and Indonesian heritage [1]. Nevertheless, as already suggested by Houtman [2], Malagasy and its dialects have relatives among languages belonging to what is today known as the Austronesian linguistic family. This was firmly established in van der Tuuk [3] and Dahl [4] pointed out a particularly close relationship between Malagasy and Maanyan of southeastern Kalimantan, which share about 45 per cent their basic vocabulary [5]. But Malagasy also bears similarities to languages in Sulawesi, Malaysia and Sumatra, including loanwords from Malay, Javanese and one (or more) language(s) of south Sulawesi [6]. Furthermore, it contains an African component in the vocabulary, especially as regards faunal names [7]. For this reason, the history of Madagascar peopling and settlement is subject to alternative interpretations among scholars. It seems that Indonesian sailors reached Madagascar by a maritime trek some 1000−2000 years ago (the exact time is subject to debate), but it is not clear whether there were multiple settlements or just a single one. Additional questions are raised by the fact that the Maanyan speakers live along the rivers of Kalimantan and have not in historical times possessed the necessary skills for long-distance maritime navigation. A possible explanation is that the ancestors of the Malagasy did not themselves navigate the boat(s) that took them to Madagascar, but were brought as subordinates of Malay sailors [8]. If this is the case, then Malagasy dialects are expected to show influence from Malay in addition to having a component similar to Maanyan. While the origin of Malagasy is thus not completely clarified there are also doubts relating to the arrival scenario. Some scholars [6] consider it most probable that the settlement of the island took place only after the initial arrival on the African mainland, while others assume that the island was settled directly, without this detour. Finally, to date no satisfactory internal classification of the Malagasy dialects has been proposed. To summarize, it would be desirable to know more about (i) when the migration to Madagascar took place, (ii) how Malagasy is related to other Austronesian languages, (iii) the historical configuration of Malagasy dialects, and (iv) where the original settlement of the Malagasy people took place.

[*] Author for correspondence (fpetroni@gmail.com).

*Received* 12 April 2011
*Accepted* 9 May 2011

Our research addresses these four problems through the application of new quantitative methodologies inspired by, but nevertheless different from, classical lexicostatistics and glottochronology [9–12].

The data, collected during the beginning of 2010, consist of 200-item Swadesh word lists for 23 dialects of Malagasy from all areas of the island. A practical orthography that corresponds to the orthographical conventions of standard Malagasy has been used. Most of the informants were able to write the words directly using these conventions, while a few of them benefited from the help of one ore more fellow townsmen. A cross-checking of each dialect list was done by eliciting data separately from two different consultants. Details about the collection of the vocabulary and about the speakers who furnished the data are provided in appendix D. This dataset probably represents the largest collection available of comparative Swadesh lists for Malagasy (see below for the locations). The lists can be downloaded from the website in Serva & Petroni [13].

The Swadesh list [14,15], rather than being a list of arbitrary words, contains terms that are common across cultures and which concern basic items of the environment, the body, and the activities pertaining to humans. Such vocabulary tends to be stable over time. The use of Swadesh lists in glottochronology has been popular for half a century.

While there are linguistic as well as geographical and temporal dimensions to the issues addressed in this paper, all strands of the investigation are rooted in an automated comparison of words through a specific version of the so-called Levenshtein or 'edit' distance (henceforth LD) [16]. The version we use here was introduced in Serva & Petroni [9] and Petroni & Serva [11] and consists of the following procedure. Words referring to the same concept for a given pair of dialects are compared with a view to how easily the word in dialect A is transformed into the corresponding word in dialect B. Steps allowed in the transformations are: insertions, deletions and substitutions. The LD is then calculated as the minimal number of such steps required to completely transform one word into the other. Calculating the distance measure that we use (the 'normalized Levenshtein distance', or LDN) requires one more operation: the 'raw LD' is divided by the length (in terms of segments) of the longer of the two words being compared. This operation produces LDN values between 0 and 1, and takes into account variable word lengths: if one or both of the words compared happen to be relatively long, the LD is prone to be higher than if they both happen to be short, so without the normalization the distance values would not be comparable. Finally, we average the LDNs for all 200 pairs of words compared to obtain a distance value characterizing the overall difference between a pair of dialects (see appendix A for a compact mathematical definition and a table with all distances).

Thus, the LDN is sensitive to both lexical replacement and phonological change and therefore differs from the cognate counting procedure of classical lexicostatistics even if the results are usually roughly equivalent.

The first use of the pairwise distances is to derive a classification of the dialects. For this purpose, we adopt a multiple strategy in order to extract a maximum of information from the set of pairwise distances. We first obtain a tree representation of the set by using two different standard phylogenetic algorithms, then we perform a structural component analysis (SCA) which, analogously to a principal components approach, represents the set in terms of geometrical relations. The SCA also provides the tool for a dating of the landing of Malagasy ancestors on the island. The landing area is established assuming that a linguistic homeland is the area exhibiting the maximum of current linguistic diversity, which is a well-known idea from biology [17] and linguistics [18]. Diversity is measured by comparing lexical and geographical distances. Finally, we perform a comparison of all variants with some other Austronesian languages, in particular with Malay and Maanyan.

For the purpose of the external comparison of Malagasy variants with other Austronesian languages, we draw upon *The Austronesian Basic Vocabulary Database* [19]. Since the wordlists in this database do not always contain all 200 items on our (and Swadesh) lists they are supplemented by various sources, including the database of the Automated Similarity Judgment Program (ASJP) [20].

If some Austronesian lists remained incomplete, the distance with respect to the Malagasy dialects was computed by averaging over the reduced vocabulary.

## 2. THE INTERNAL CLASSIFICATION OF MALAGASY

### 2.1. Our results

In this section, we present two different classificatory trees for the 23 Malagasy dialects obtained through applying two different phylogenetic algorithms to the set of pairwise distances resulting from comparing our 200-item word lists through the LDN.

The two algorithms used are neighbour joining (NJ) [21] and the unweighted pair group method with arithmetic mean (UPGMA) [22]. The main theoretical difference between the algorithms is that the UPGMA assumes that evolutionary rates are the same on all branches of the tree, while NJ allows differences in evolutionary rates. The question of which method is better at inferring the phylogeny has been studied by running various simulations where the true phylogeny is known. Most of these studies were in biology but at least one [23] specifically tried to emulate linguistic data. Most of the studies (starting with [21] and including [23]) found that NJ usually came closer to the true phylogeny. Since, in our case, the relations among dialects are not necessarily tree-like, it is desirable to test the different methods against empirical linguistic data, which is mainly why trees derived by means of both methods are presented here.

The input data for the UPGMA tree are the pairwise separation times obtained from lexical distances by a rule [9] which is a simple generalization of the

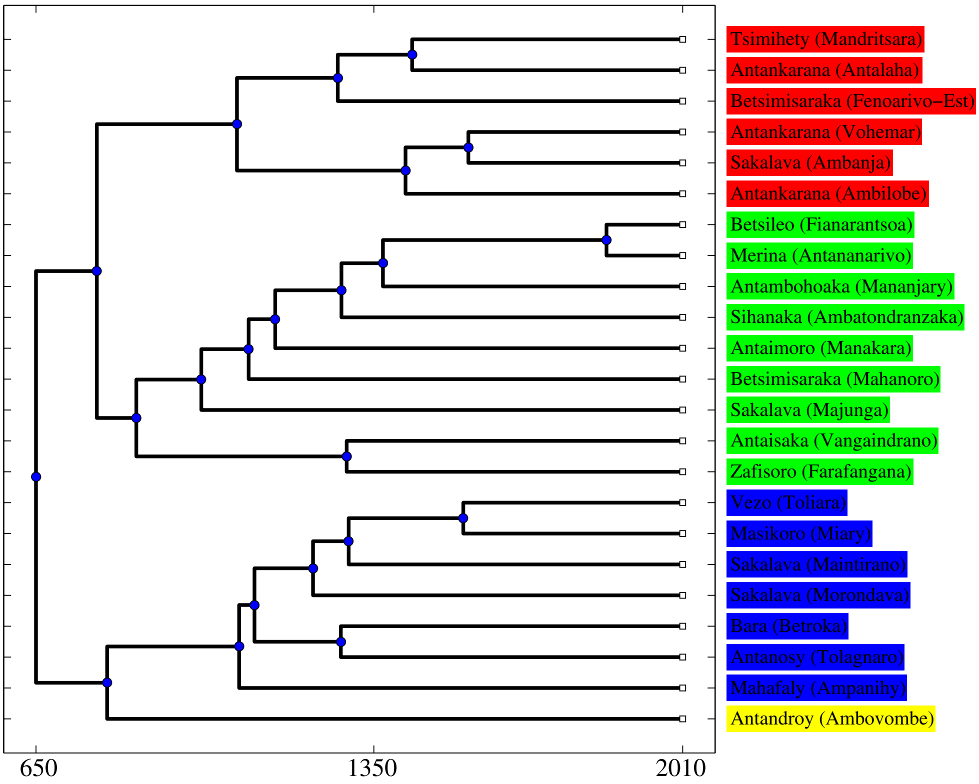

Figure 1. UPGMA tree for 23 Malagasy dialects, with hypothetical separation times. Variants are named by traditional dialect names followed by locations in parentheses. The four main branches are coloured distinctively. The main separation of Malagasy dialects is centre–northeast versus southwest.

fundamental formula of glottochronology. The absolute time scale is calibrated by the results of the SCA analysis (see below), which indicate a separation date of AD 650. While the scale below the UPGMA tree (figure 1) refers to separation times, the scale below the NJ tree (figure 2) simply shows lexical distance from the root.

The unit of the scale of figure 2 is chosen such that the LDN between two language variants is roughly equal to the sum of their lexical distance from their closest common node. For example, two dialects that are maximally distant, such as Antandroy and Tsimihety, are about $0.27 + 0.24 = 0.51$ LDN distant.

Since UPGMA assumes equal evolutionary rates, the ends of all the branches line up on the right-hand side of the UPGMA tree. The assumption of equal rates also determines the root of the tree on the left-hand side. NJ allows unequal rates, so the ends of the branches do not all line up on the NJ tree. The extent to which they fail to line up indicates how variable the rates are. The tree is rooted by the midpoint (the point in the network equidistant from the two most distant dialects) but we also checked that the same result is obtained following the standard strategy of adding an outgroup. In particular, we added Italian but also tried with Maanyan, always obtaining the same tree.

There is a good fit between the geographical position of the dialects (figure 3) and their position in both the UPGMA (figure 1) and NJ trees (figure 2). In both trees, the dialects are divided into two main groups (coloured blue and yellow versus red and green in figure 1) even if we found differences, which will be discussed below.

Given the consensus between the two methods, the result regarding the basic split can be considered solid. Geographically, the division corresponds to a border running from the southeast to the northwest of the island, as shown in figure 3 where the UPGMA and NJ main separation lines are drawn. A major difference concerns the Vangaindrano, Farafangana and Manakara dialects, which have shifting allegiances with respect to the two main groups under the different analyses. Additionally, there are minor differences in the way that the two main groups are configured internally. Most strikingly, we observe that, in the UPGMA tree, Majunga is grouped with the central dialects while, in the NJ tree, it is grouped with the northern ones. This indeterminacy would seem to relate to the fact that the town of Majunga is at the geographical border of the two regions.

Dialects from close regions are usually perceived as being similar by Malagasy people while distant dialects usually have a low degree of mutual intelligibility. Most of the people are able to understand the Merina dialect, which is the official language, but outside of the Imerina region only cultivated people are able to speak it. Thus, diglossia is quite limited. French is only used as a bureaucratic or teaching language and is practically never used in everyday conversation. There are quite a lot of loanwords from French in Malagasy but almost none in the Swadesh

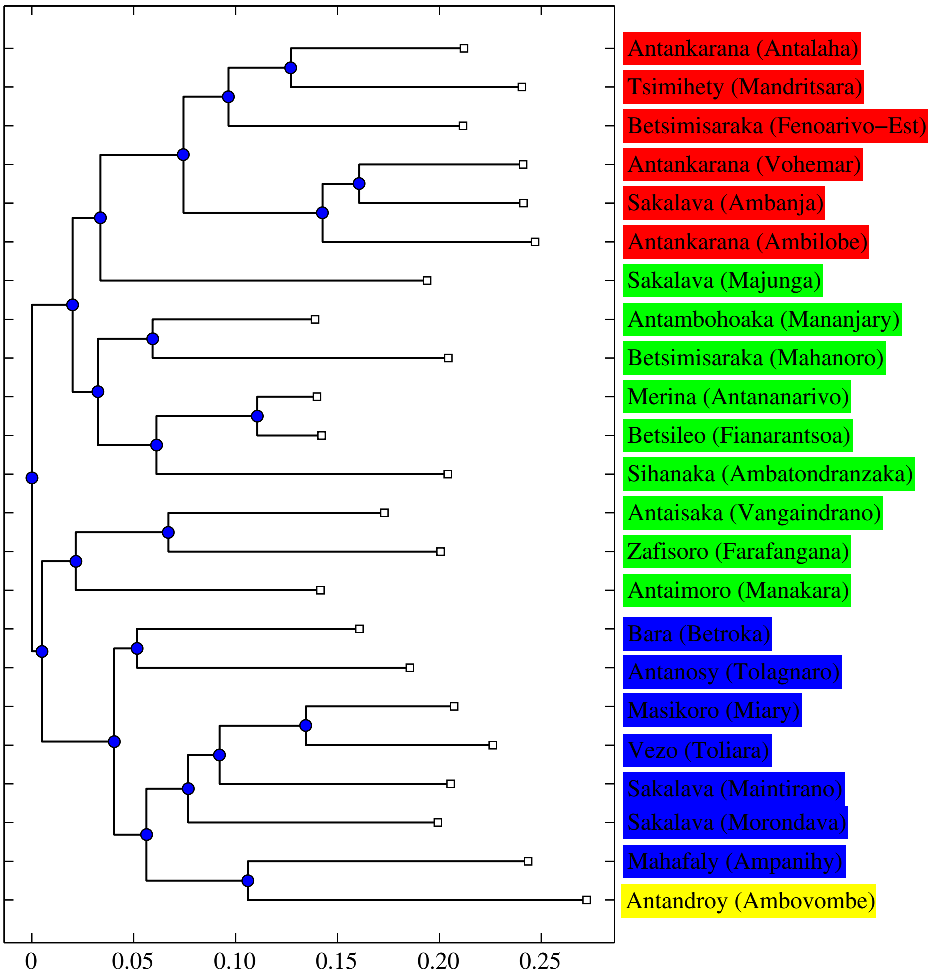

Figure 2. NJ tree for 23 Malagasy dialects. Colours compare with the UPGMA tree in figure 1. The graph confirms the main centre–northeast versus southwest division. The main difference is that three dialects at the linguistic border are grouped differently. Colours facilitate a rapid comparison.

lists (we only registered French loanwords for ‘ice’ or ‘snow’ in a few dialects).

Another difference is that, in the UPGMA tree, the Ambovombe variant of the dialect traditionally called *Antandroy* is quite isolated, whereas, in the NJ tree, Ambovombe and the Ampanihy variant of *Mahafaly* group together. Since the UPGMA algorithm is a strict bottom-up approach to the construction of a phylogeny, where the closest taxa are joined first, it will tend to treat the overall most deviant variant last. This explains the differential placement of Ambovombe in the two trees. The length of the branch leading to the node that joins Ambovombe and Ampanihy in the NJ tree shows that these two variants have quite a lot of similarities, but, in the UPGMA method, these similarities in a sense ‘drown’ in the differences that set Ambovombe apart from other Malagasy variants *as a whole*.

As further confirmation of this analysis, we also computed the average LDN distance from each dialect to all the others. Antandroy has the largest average distance, confirming that it is the overall most deviant variant (something which is also commonly pointed out by other Malagasy speakers). We further note that the smallest average distance is for the official variant, that of Merina. This may be explained, at least in part, as an effect of the convergence of other variants towards this standard.

### 2.2. *The results of Vérin* et al. [24]

Our classification results, including the grouping of the dialects in a southwest and a centre–east–north cluster, differ from Vérin *et al.*’s [24] interpretation of their results, according to which there is a major split between the dialects on the northern tip of the island and all the rest.

This difference is somewhat surprising, so let us look into the way that Vérin *et al.* proceeded. There are some differences in the way that their and our datasets were constructed and the coverage. Vérin *et al.* used a 100-item Swadesh list, while we used a larger set of 200 words. We included locations that Vérin *et al.* did not cover. Moreover, following Gudschinsky [25], Vérin *et al.* ([24], p. 35) excluded Bantu loanwords from consideration, whereas we treated loanwords on a par with inherited words (in practice, however, Vérin *et al.* only seemed to identify one form as Bantu, namely *amboa* ‘dog’). Finally, a major difference is that Vérin *et al.* evaluated distances by the standard glottochronological approach based on cognate counting whereas we used the LDN measure.

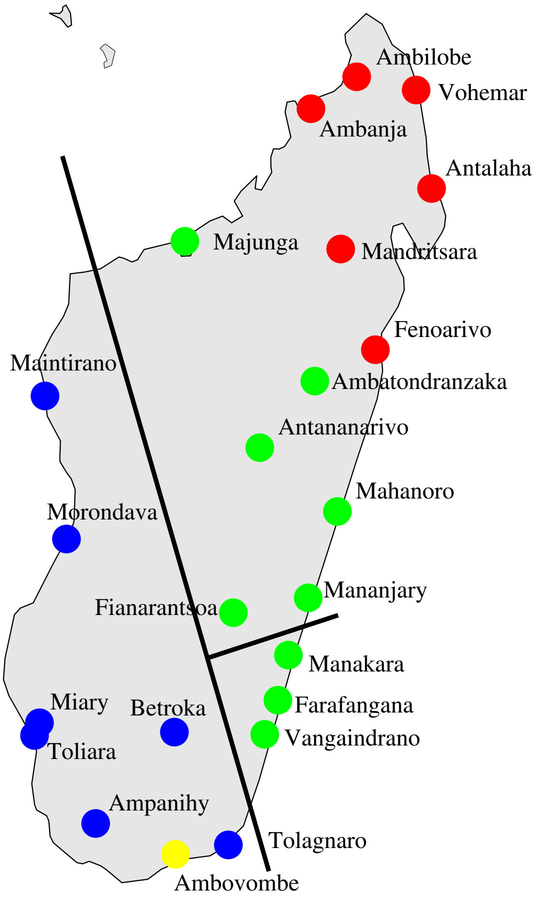

Figure 3. Geographical locations of the 23 dialects studied, with colours showing the main dialect branches according to figure 1. The straight line running from the southeast to the northwest of the island corresponds to the basic split in the UPGMA tree while the NJ grouping is similar but Manakara, Farafangana and Vaingandrano are grouped with the southwest.

In spite of these differences, our results are in reality quite similar to those of Vérin *et al.*, the differences mainly relating to the interpretation of their results. The great leap from results to interpretation is due to the fact that Vérin *et al.* did not have the kinds of sophisticated phylogenetic methods at their disposal for deriving a classification from a matrix of cognacy scores that are available today. Their method for constructing trees goes something like this: cluster the closest dialects first, using some threshold. Then move the threshold and join dialects or dialect groups under deeper nodes. Different trees can be constructed from using different thresholds. One of the problems with this approach, not addressed by the authors, is that it assumes a constant rate of change. For instance, in one of their trees (their chart 1 on p. 59) Merina, Sihanaka and Betsileo Ambositra are joined under one node attached at the 92 per cent cognacy level. The actual percentages, however, do not fit a constant rate scenario (a.k.a. 'ultrametricality'): Sihanaka and Merina share 92 per cent cognates, Betsileo Ambositra and Merina also share 92 per cent, but Sihanaka and Betsileo Ambositra share only 86 per cent. No solution to this problem is given (and, indeed, it is a problem for any phylogenetic algorithm that cannot be 'solved' but at least needs to be addressed). Instead, violations of ultrametricality seem to be dealt with in an ad hoc way. In the case of the example just given, Sihanaka and Betsileo Ambositra are treated as if they also shared 92 per cent cognates. Since the principles used by Vérin *et al.* to derive their trees are unclear, there is no need to discuss their trees in detail. Moreover, each tree in their article differs from the next, making it difficult to summarize the claims embodied in these trees. Some generalizations, however, do emerge. The Antankarana dialect in the far north constitutes its own isolated branch in all three trees, and in all three trees there are three sets of dialects that always belong together on different nodes: (i) Merina, Sihanaka, Betsileo Ambositra, Betsimisaraka, (ii) Taimoro, Antaisaka, Zafisoro and (iii) Mahafaly, Antandroy 1. Other dialects have no particularly close relationship to any other dialect, or else exhibit shifting allegiances.

In figure 4*a*, we subject the distance data of Vérin *et al.* to NJ (for later comparison with our results we consider only the variants also included in our dataset). Using this method, each of the clusters (1–3) also appears, but joined by other dialects that could not be safely placed at any deeper level of embedding by Vérin *et al.* Thus, their clustering method essentially produces so much information that only about half of the dialects become meaningfully classified. The most problematical aspect of their interpretation, however, is that there is supposed to be a fundamental split between the Antankarana dialect in the far north of the island and all other dialects. As we demonstrate in figure 4*a*, this is not borne out by the data, but is an artefact of the clustering method.

The NJ interpretation of the results of Vérin *et al.* (figure 4*a*) may be compared with our own results obtained from the LDN distances evaluated using our own data (figure 4*b*). Only variants belonging to the intersection of the two datasets are included. The Betsimisaraka list from our data is the one from Mahanoro and the Antankarana list is the one from Vohemar.

The two trees have similar topologies, in particular the main partition in both cases separates centre–northeast from southwest dialects. Therefore, our results and those of Vérin *et al.* coincide with respect to this point, although Vérin *et al.*'s interpretation of their own results is different. It is remarkable that the differences between the two trees are so minor considering differences both in the data and in the methods for calculating differences among dialects.

Figure 4*b* was produced by using the same input LDN distances and the same NJ algorithm as used for figure 2. Comparing the two trees, we observe that the simple reduction in the number of input dialects has the effect of modifying the position of Farafangana, Vankaindrano and Manakara variants (compare figure 4*b* with figure 2). Indeed, the NJ tree in figure 4*b* based on 15 dialects shows the same main branching as the UPGMA tree in figure 1, which differs from that of the NJ tree in figure 2 based on 23 dialects. This instability of tree topology caused by the number of input dialects and the differences in algorithms (UPGMA versus NJ) shows that a tree structure is not optimal for capturing

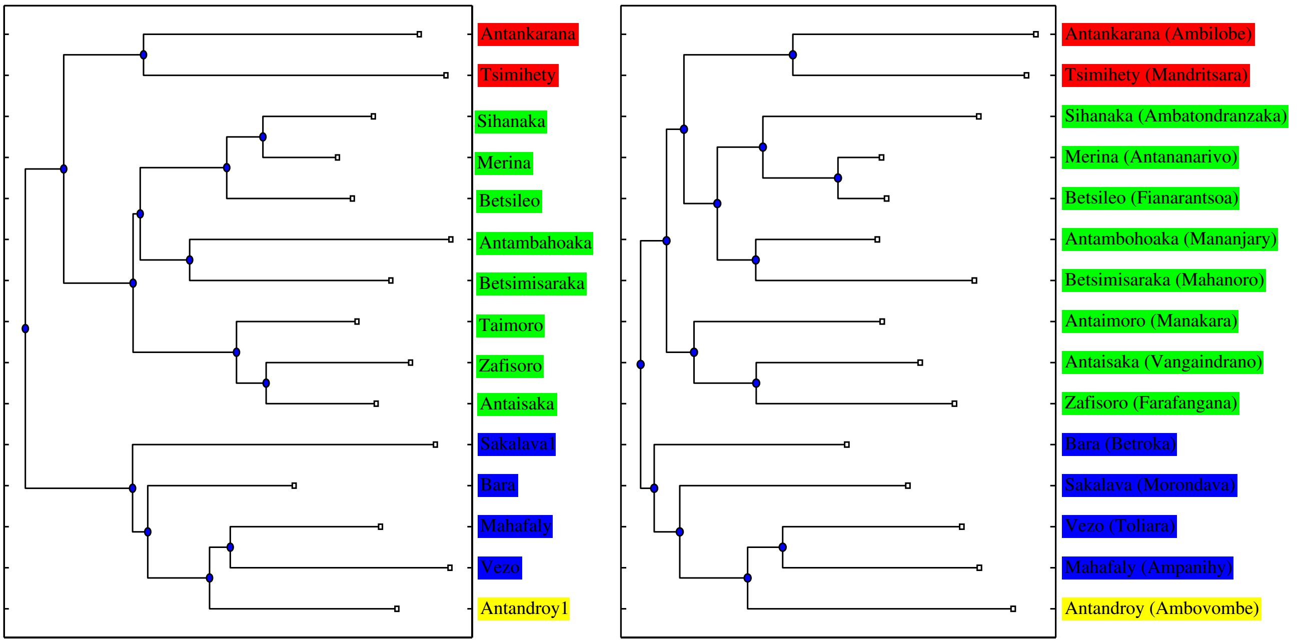

Figure 4. Comparison between NJ trees based on, respectively, data collected (*a*) by Vérin *et al.* [24] and (*b*) by ourselves. The tree with Vérin data (figure 4*a*) is obtained by the standard lexicostatistical approach while the tree with our data (figure 4*b*) uses LDN distances. Names of variants for the tree in (*a*) are those of Vérin *et al.* [24], but the correspondence with our naming scheme, which makes use of both dialect names and towns, is evident.

all the information contained in the set of lexical distances. Thus, we consider a different, geometrically based approach, presented in the following section, necessary for a verification of the classification results.

## 3. GEOMETRIC REPRESENTATION OF MALAGASY DIALECTS

Although tree diagrams have become ubiquitous in representations of language taxonomies they fail to reveal the full complexity of affinities among languages. The reason is that the simple relation of ancestry, which is the single principle behind a branching family tree model, cannot grasp the complex social, cultural and political factors moulding the evolution of languages [26]. Since dialects within a group interact with each other and with the languages of other families in ‘real time’, it is obvious that historical developments in languages cannot be described only in terms of pairwise interactions, but reflect a genuine higher order influence, which can best be assessed by SCA. This is a powerful tool which represents the relationships among different languages in a language family geometrically, in terms of distances and angles, as in the Euclidean geometry of everyday intuition. Being a version of the kernel principal component analysis (PCA) method [27], it generalizes PCA to cases where we are interested in principal components obtained by taking all higher order correlations between data instances. It has so far been tested through the construction of language taxonomies for 50 major languages of the Indo-European and Austronesian language families [28]. The details of the SCA method are given in appendix B.

In figure 5, we show the three-dimensional geometric representation of 23 dialects of the Malagasy and Maanyan languages, which is closely related to Malagasy. The three-dimensional space is spanned by the three major data traits ($\{q_2, q_3, q_4\}$, see appendix B for details) detected in the matrix of linguistic LDN distances.

The clear geographical patterning is perhaps the most remarkable aspect of the geometric representation. The structural components reveal themselves in figure 5 as two well-separated spines representing both the northern (red) and the southern (blue) dialects. It is remarkable that all Malagasy dialects belong to a single plane orthogonal to the data trait of the Maanyan language ($q_2$). The plane of Malagasy dialects is attested by the sharp distribution of the language points in Cartesian coordinates along the data trait $q_2$. This colour point of Malagasy dialects over their common plane is shown in figure 6, in which a reference azimuth angle $\phi$ is introduced in order to underline the evident symmetry. It is important to mention that, although the language point of Antandroy (Ambovombe) is located on the same plane as the rest of the Malagasy dialects, it is situated far away from them and obviously belongs to neither of the major dialect branches; for this reason it is not reported in the next figure to be discussed (figure 6). This clear SCA isolation of Antrandroy is compatible with its position in the tree in figure 1.

The distribution of language points supports the main conclusion following from the UPGMA and NJ methods (figures 1 and 2) of a division of the main group of Malagasy dialects into three groups: north (red), southwest (blue) and centre (green). These clusters are evident from the representation shown in figure 6. However, with respect to the

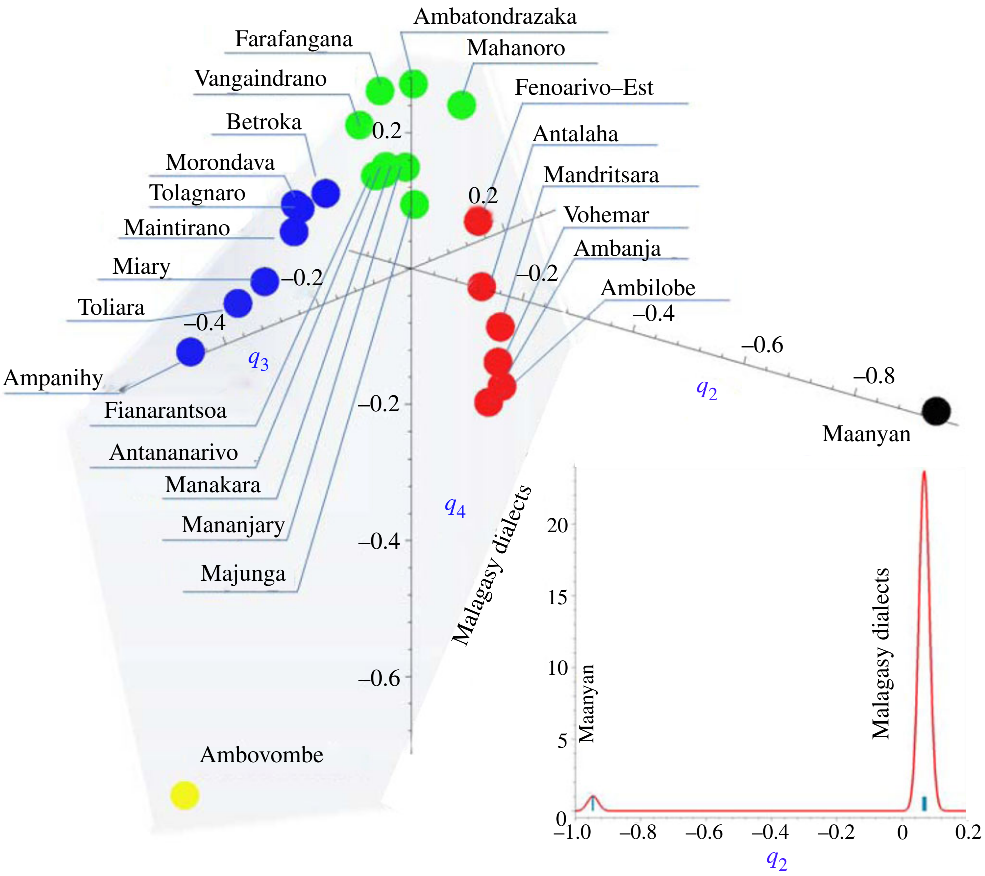

Figure 5. The three-dimensional geometric representation of the Malagasy dialects and the Maanyan language in the space of major data traits ($q_2$, $q_3$, $q_4$) shows a remarkable geographical patterning separating the northern (red) and the southern (blue) dialect groups, which fork from the central part of the island (the dialects spoken in the central part are coloured green, while Antandroy is yellow). The kernel density estimate of the distribution of the $q_2$ coordinates, together with the absolute data frequencies, indicates that all Malagasy dialects belong to a single plane orthogonal to the data trait of the Maanyan language ($q_2$).

classification of some individual dialects, the SCA method differs from the UPGMA and NJ results. Since their azimuthal coordinates better fit the general trend of the southern group, the Vangaindrano, Farafangana and Ambatondranzaka dialects spoken in the central part of the island are now grouped with the southern dialects (blue) rather than the central ones (see §2). Similarly, the Mahanoro dialect is now classified in the northern group (red), since it is best fitted to the northern group azimuth angle. The remaining five dialects of the central group (green) are characterized by the azimuth angles close to a bisector ($\phi = 0$).

## 4. THE ARRIVAL IN MADAGASCAR

### *4.1. Dating the arrival*

The radial coordinate of a dialect is simply the distance of its representative point from the origin of coordinates in figure 5. It can be verified that the position of Malagasy dialects along the radial direction is remarkably heterogeneous, indicating that the rate of change in the Swadesh vocabulary was anything but constant. In fact, if the rate of change during the evolution of the variants from the proto-language had been the same or not very different, the radial coordinates would also have been almost identical.

The radial coordinates have been ranked and then plotted in figure 7 against their expected values under normality, such that departures from linearity signify departures from normality. The dialect points in figure 7 show very good agreement with univariate normality with the value of variance $\sigma^2 = 0.99 \times 10^{-3}$, which results from the best fit of the data. This normal behaviour can be justified by the hypothesis that the dialect vocabularies are the result of a gradual and cumulative process in which many small, independent innovations have emerged and to which they have additively contributed.

The SCA is based on the statistical evaluation of differences among the items of the Swadesh list. A complex nexus of processes behind the emergence and differentiation of dialects is described by the single degree of freedom (as another degree of freedom, the azimuth angle, is fixed by the dialect group) along the radial direction [28,29].

The univariate normal distribution of the radial coordinates (figure 7) can be assessed in the framework of the diffusive model of language evolution [28], in

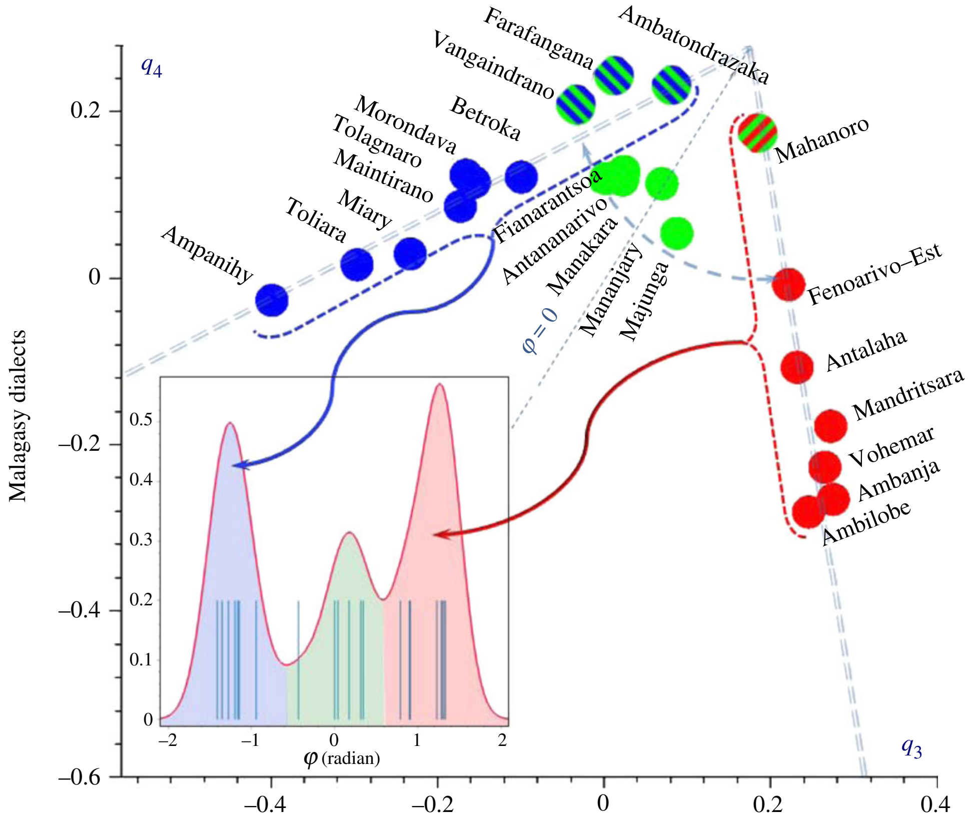

Figure 6. The plane of Malagasy dialects ($q_3$, $q_4$); Antandroy (Ambovombe) is excluded. The kernel density estimate of the distribution over azimuth angles, together with the absolute data frequencies, allows the rest of the Malagasy dialects to be classified into the three groups: north (red), southwest (blue) and centre (green).

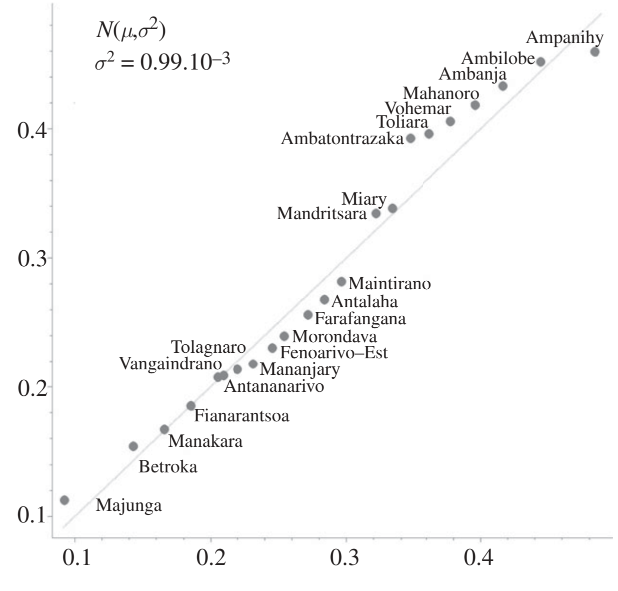

Figure 7. The radial coordinates are ranked and then plotted against their expected values under normality. Departures from linearity, which signify departures from normality, are minimal.

which the evolution is viewed as driven by independent, petty events. Within such a model, a homogeneous diffusion time evolution in one dimension is implied, under which variance $\sigma^2 \propto t$ grows linearly with time.

We stress that the constant rate of increase in the variance of radial positions of languages in the geometrical representation (figure 5) has nothing to do with the traditional glottochronological assumption about the constant replacement rate of cognates assumed by the UPGMA method. It is also important to mention that the value of variance $\sigma^2 = 0.99 \times 10^{-3}$ calculated for the Malagasy dialects does not correspond to physical time but rather gives a statistically consistent estimate of age for the group of dialects. In order to assess the pace of variance changes with physical time and to calibrate the dating method, we have used historically attested events. Although the lack of documented historical events makes the direct calibration of the method difficult, we suggest (following [28]) that variance evaluated over the Swadesh vocabulary proceeds approximately at the same pace uniformly for all human societies. For calibrating the dating mechanism in Blanchard *et al.* [28], we have used the following four anchoring historical events (see [30]) for the Indo-European language family: (i) the last Celtic migration (to the Balkans and Asia Minor) (by 300 BC), (ii) the division of the Roman Empire (by AD 500), (iii) the migration of German tribes to the Danube River (by AD 100), and (iv) the establishment of the Avars Khaganate (by AD 590) causing the spread of the Slavic people. It is remarkable that all of the events mentioned uniformly indicate a very slow variance pace of a millionth per year, $t/\sigma^2 = (1.367 \pm 0.002) \times 10^6$. This time–age

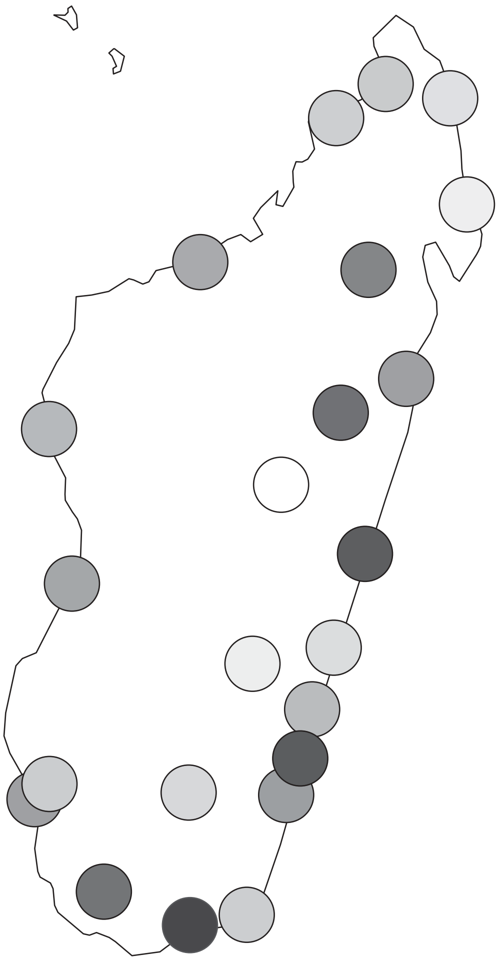

Figure 8. The homeland of Malagasy dialects as determined through diversity measures. The towns with darkest grey circles have the highest diversity values while those with lightest grey have the lowest. The most diverse area is on the southeast coast, where landing would have occurred; the least diverse area is in the north, indicating that this area was settled last.

ratio returns $t = 1353$ years if applied to the Malagasy dialects, suggesting that landing in Madagascar was around AD 650. This is in complete agreement with the prevalent opinion among scholars, including the influential one of Adelaar [6].

### *4.2. The landing area*

In order to hypothetically infer the original centre of dispersal of Malagasy variants, we here use a variant of the method of Wichmann *et al.* [31]. This method draws upon a well-known idea from biology [17] and linguistics [18]—that the homeland of a biological species or a language group corresponds to the current area of greatest diversity. In Wichmann *et al.* [31] this idea is transformed into quantifiable terms in the following way. For each language variant, a diversity index is calculated as the average of the proportions between linguistic and geographical distances from the given language variant to each of the other language variants (see [31] for more detail). The geographical distance is defined as the great-circle distance (i.e. as the crow flies) measured by angle radians. In this paper, we adopt a variant of the method described in more detail in appendix C.

The result of applying this method to Malagasy variants is that the best candidate for the homeland is the southeast coast where the three most diverse towns, i.e. Ambovombe, Farafangana and Mahanoro, are located, and where the surrounding towns are also highly diverse. The northern locations are the least diverse and they must have been settled last.

A convenient way of displaying the results on a map is shown in figure 8, where locations are indicated by means of circles with different gradations of grey. The greater the diversity of a location is, the darker the grey. The figure suggests that the landing would have occurred somewhere between Mahanoro (central part of the east coast) and Ambovobe (extreme south of the east coast), the most probable location being in the centre of this area, where Farafangana is situated. Finally, we have checked that, if the entire Greater Barito East group is considered, the homeland of Malagasy stays in the same place, but becomes secondary with respect to the south Borneo homeland of the group.

The identification of a linguistic homeland for Malagasy on the southeastern coast of Madagascar receives some independent support from unexpected types of evidence. According to Faublée [32], there is an Indian Ocean current that connects Sumatra with Madagascar. When Mount Krakatoa exploded in 1883, pumice was washed ashore on Madagascar's east coast where the Mananjary River opens into the sea (between Farafangana and Mahanoro). During the Second World War, the same area saw the arrival of pieces of wreckage from ships sailing between Java and Sumatra that had been bombed by the Japanese airforce. The mouth of the Mananjary River is where the town of Mananjary is presently located, and it is in the highly diverse southeast coast as shown in figure 8. To enter the current that would eventually carry them to the east coast of Madagascar, the ancestors of today's Malagasy people would probably have passed by the easily navigable Sunda strait.

In his studies on the roots of Malagasy, Adelaar found that the language has an important contingent of loanwords from Sulawesi (Buginese) [6,8]. We have also compared Malagasy (and its dialects) with various Indonesian languages. While we unsurprisingly found that Maanyan is the closest language, we also found that the second closest language is Maranao (Buginese is the third) but for some Malagasy dialects Buginese is the second (see also [11]). The similarity with Buginese appears to be a further argument in favour of the southern path through the Sunda strait to Madagascar. In fact, if the Malay sailors recruited their crew in Borneo and, to a limited extent, in Sulawesi, they probably crossed this strait before starting their navigation in the open waters.

Furthermore, we found that the dialects of Mananjary, Manakara, Antananarivo and Fianarantsoa are noticeably closer both to Maanyan and Malay with respect to the other variants. Mananjary and Manakara are both in the identified landing area on the southeast coast while Antananarivo and Fianarantsoa are in the central highlands of Madagascar. This may suggest that landing was followed shortly after by a migration to the interior of the island.

## 5. CONCLUSION AND OUTLOOK

All results in this paper rely on two main ingredients: a new dataset from 23 different variants of the language and an automated method to evaluate lexical distances. Analysing the distances through different types of phylogenetic algorithms (NJ and UPGMA) as well as through a geometrical approach, we find that all approaches converge on a result where dialects are classified into two main geographical subgroups: southwest versus centre–northeast. It is not clear, at this stage, whether this main division is caused by geography or by an early splitting of the population into two different subpopulations or even by a colonization history with more than one founding nucleus. The last hypothesis, however, is somewhat unlikely given the relative uniformity of the dialects.

An output of the geometric representation of the distribution of the dialects is a landing date of around AD 650, in agreement with a view commonly held by students of Malagasy. Furthermore, by means of a technique that is based on the calculation of differences in linguistic diversity, we propose that the southeast coast was the location were the first colonizers landed. This location also suggests that the path followed by the sailors went from Borneo, through the Sunda strait, and, subsequently, along major oceanic currents, to Madagascar.

Finally, we measured the distance of the Malagasy variants to other Indonesian languages and found that the dialects of Mananjary, Manakara, Antananarivo and Fianarantsoa are noticeably closer to most of them than the other dialects.

A larger comparison of Malagasy variants with Indonesian (and possibly African) languages is desirable. Although Malagasy is assigned to the Greater East Barito group, it has many loanwords from other Indonesian languages, such as Javanese, Buginese and Malay, especially in the domains of maritime life and navigation [6–8,33,34]. It has also been observed that it is unlikely that Maanyan-speaking Dayaks were responsible for the spectacular migrations from Kalimantan to Madagascar since they are forest dwellers with river navigation skills only. Furthermore, many manifestations of Malagasy culture cannot be linked to the culture of the Dayaks of the southeast Barito area. For example, the Malagasy people use outrigger canoes, whereas southeast Barito Dayaks never do; some of the Malagasy musical instruments are very similar to musical instruments in Sulawesi; and some of the Malagasy cultigens (wet rice) cannot be found among Barito river inhabitants. In contrast, some funeral rites, such as the *famadihana* (second burial), are similar to those of Dayaks. Nevertheless, it should be observed that it is not clear whether the above cultural traits are specific to a region or a people or whether they are generic traits that can be found sporadically in other Austronesian cultures.

Non-Maanyan cultural and linguistic traits raise several questions concerning the ancestry of the Malagasy people. Assuming that Dayaks were brought as subordinates together with a few other Indonesians by Malay seafarers, they formed the majority in the initial group and their language constituted the core element of what later became Malagasy. In this way, Malagasy would have absorbed words of the Austronesian languages of the other slaves and of the Malay seafarers. Is this a sufficient explanation, or are things more complicated? For example, may we hypothesize two or more founding colonies with different ethnic compositions? And is it possible that later specific contacts altered the characteristics of some local dialects?

In order to answer these questions, it is necessary to make a careful comparison of all Malagasy variants with all Austronesian languages. A dialect may provide information about the pre-migratory composition of its speakers and also about further external contributions owing to successive landings of Indonesian sailors.

Furthermore, the island was almost surely inhabited before the arrival of Malagasy ancestors. Malagasy mythology portrays a people, called the *Vazimba*, as the original inhabitants, and it is not clear whether they were part of a previous Austronesian expansion or a population of a completely different origin (Bantu, Khoisan?). Is it possible to track the aboriginal vocabulary into some of the dialects, such as Mikea (see [34])?

These questions call for a new look at the Malagasy language, not as a single entity but as a constellation of variants whose histories are still to be fully understood.

We thank Philippe Blanchard and Eric W. Holman for comments on an earlier version of this paper. Further, M. Serva wishes to thank Sanhindou Amady, Clement ‘Zazalahy’, Beatrice Rolla, Renato Magrin, Corto Maltese and Gianni Dematteo for logistical support during his stay in Madagascar.

## APPENDIX A

The lexical distance [9,11] between two languages, $l_i$ and $l_j$, is computed as the average of the normalized Levenshtein (edit) distance [16] over the vocabulary of 200 items,

$$D(l_i, l_j) = \frac{1}{200}\sum_{\alpha=1}^{200} \frac{\| w_i(\alpha), w_j(\alpha) \|}{\max(|w_i(\alpha)|, |w_j(\alpha)|)}, \qquad (\text{A } 1)$$

where the Swadesh item is indicated by $\alpha$, $\| w_i(\alpha), w_j(\alpha) \|$ is the standard LDN between the words $w_i(\alpha)$ and $w_j(\alpha)$ and $|w_i(\alpha)|$ is the number of characters in the word $w_i(\alpha)$. The sum runs over all the 200 different items of the Swadesh list. Assuming that the number of languages (or dialects) to be compared is $N$, then the distances $D(l_i, l_j)$ are the entries of a $N \times N$ symmetric matrix $\mathbf{D}$ (obviously $D(l_i, l_i) = 0$).

The matrix, with entries multiplied by 1000, is the following:

| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | | | | | | | | | |
| 2 | 323 | | | | | | | | | | | | | | | | | | | | | |
| 3 | 246 | 276 | | | | | | | | | | | | | | | | | | | | |
| 4 | 322 | 240 | 295 | | | | | | | | | | | | | | | | | | | |
| 5 | 302 | 281 | 309 | 345 | | | | | | | | | | | | | | | | | | |
| 6 | 227 | 318 | 275 | 359 | 266 | | | | | | | | | | | | | | | | | |
| 7 | 413 | 386 | 390 | 418 | 314 | 370 | | | | | | | | | | | | | | | | |
| 8 | 280 | 386 | 342 | 401 | 356 | 245 | 436 | | | | | | | | | | | | | | | |
| 9 | 366 | 424 | 379 | 412 | 405 | 375 | 450 | 409 | | | | | | | | | | | | | | |
| 10 | 411 | 396 | 416 | 440 | 318 | 366 | 249 | 456 | 482 | | | | | | | | | | | | | |
| 11 | 207 | 326 | 260 | 362 | 286 | 061 | 383 | 201 | 374 | 384 | | | | | | | | | | | | |
| 12 | 362 | 343 | 345 | 387 | 292 | 328 | 289 | 397 | 435 | 330 | 324 | | | | | | | | | | | |
| 13 | 303 | 369 | 330 | 381 | 384 | 329 | 454 | 362 | 256 | 487 | 318 | 407 | | | | | | | | | | |
| 14 | 343 | 302 | 331 | 355 | 243 | 317 | 303 | 403 | 423 | 314 | 336 | 301 | 419 | | | | | | | | | |
| 15 | 397 | 453 | 394 | 462 | 392 | 375 | 342 | 463 | 485 | 304 | 383 | 405 | 471 | 388 | | | | | | | | |
| 16 | 368 | 391 | 385 | 416 | 392 | 390 | 448 | 406 | 320 | 474 | 383 | 429 | 325 | 418 | 486 | | | | | | | |
| 17 | 400 | 350 | 369 | 390 | 280 | 358 | 165 | 433 | 427 | 278 | 373 | 240 | 439 | 261 | 358 | 410 | | | | | | |
| 18 | 322 | 376 | 325 | 374 | 391 | 337 | 426 | 381 | 198 | 473 | 339 | 412 | 234 | 406 | 461 | 264 | 414 | | | | | |
| 19 | 358 | 407 | 376 | 417 | 408 | 394 | 440 | 419 | 292 | 481 | 387 | 431 | 325 | 422 | 472 | 161 | 408 | 243 | | | | |
| 20 | 297 | 388 | 359 | 430 | 356 | 299 | 400 | 346 | 386 | 433 | 275 | 375 | 363 | 375 | 455 | 348 | 394 | 349 | 355 | | | |
| 21 | 386 | 341 | 370 | 385 | 290 | 344 | 262 | 403 | 422 | 321 | 348 | 250 | 404 | 306 | 403 | 401 | 213 | 416 | 417 | 383 | | |
| 22 | 225 | 389 | 332 | 394 | 382 | 316 | 471 | 319 | 385 | 475 | 287 | 421 | 296 | 431 | 480 | 382 | 467 | 348 | 387 | 356 | 441 | |
| 23 | 379 | 424 | 407 | 424 | 398 | 380 | 443 | 433 | 315 | 466 | 380 | 412 | 351 | 420 | 472 | 203 | 395 | 288 | 202 | 351 | 409 | 406 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |

where the number-variant correspondence is (name of location is in brackets):

1 Antambohoaka (Mananjary), 2 Antaisaka (Vangaindrano), 3 Antaimoro (Manakara), 4 Zafisoro (Farafangana), 5 Bara (Betroka), 6 Betsileo (Fianarantsoa), 7 Vezo (Toliara), 8 Sihanaka (Ambatondranzaka), 9 Tsimihety (Mandritsara), 10 Mahafaly (Ampanihy), 11 Merina (Antananarivo), 12 Sakalava (Morondava), 13 Betsimisaraka (Fenoarivo−Est), 14 Antanosy (Tolagnaro), 15 Antandroy (Ambovombe), 16 Antankarana (Vohemar), 17 Masikoro (Miary), 18 Antankarana (Antalaha), 19 Sakalava (Ambanja), 20 Sakalava (Majunga), 21 Sakalava (Maintirano), 22 Betsimisaraka (Mahanoro), 23 Antankarana (Ambilobe).

## APPENDIX B

The lexical distance (A 1) between two languages, $l_i$ and $l_j$, can be interpreted as the average probability to distinguish them by a mismatch between two characters randomly chosen from the orthographic realizations of the vocabulary meanings. There are infinitely many matrices that match all the structures of **D**, and therefore contain all the information about the relationships between languages [28]. It is remarkable that all these matrices are related to each other by means of a linear transformation,

$$\left.\begin{aligned}&\mathbf{T} = \Delta^{-1}\mathbf{D},\\ &\Delta = \mathrm{diag}\left(\sum_{k=1}^{N} D(l_1, l_k)\ldots\sum_{k=1}^{N} D(l_N, l_k)\right),\end{aligned}\right\} \qquad \text{(B 1)}$$

which can be interpreted as a random walk [28,29] defined on the weighted undirected graph determined by the matrix of lexical distances **D** over the $N$ different languages. We have to emphasize that the appearance in our approach of random walks does not carry any particular assumption regarding evolutionary processes in language (as we are not concerned with the problems of modelling diffusion between populations or the spread of information through a society), nor does it relate to the Bayesian analysis used previously [35,36] to construct self-consistent tree-like representations of linguistic phylogenies. They only concern the *unique* linear transformation (in the class of stochastic matrices) consistent with all of the structures of the matrix of lexical distances calculated with respect to Swadesh's list of meanings. Random walks defined by the transition matrix (B 1) describe the statistics of a sequential process of language classification. Namely, while the elements $T(l_i, l_j)$ of the matrix **T** evaluate the probability of successful differentiation of the language $l_i$ provided the language $l_j$ has been identified with certainty, the elements of the squared matrix $\mathbf{T}^2$ ascertain the successful differentiation of the language $l_i$ from $l_j$ through an intermediate language, the elements of the matrix $\mathbf{T}^3$ give the probabilities of differentiating 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 (A 1) is uncovered by the von Neuman series estimating the characteristic time of successful classification for any two languages in the database over a language family,

$$\mathbf{J} = \lim_{n\to\infty}\sum_{k=0}^{n}\mathbf{T}^n = \frac{1}{1-\mathbf{T}}. \qquad \text{(B 2)}$$

The last equality in equation (B 2) is understood as the group generalized inverse [29], being a symmetric, positive semi-definite matrix that 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 $\mathbf{J}$. 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 $\lambda_k \geq 0$, is associated with the virtually independent trait in the matrix of lexical distances $\mathbf{D}$. 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, $l_i \rightarrow (q_{1,i}, q_{2,i}, \ldots q_{N,i})$. Languages that are 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 for analytically. The rank-ordering of data traits $\{q_k\}$, in accordance with their eigenvalues, $\lambda_0 = \lambda_1 < \lambda_2 = \ldots = \lambda_N$, provides us with the natural geometric framework for dimensionality reduction. At variance with the standard PCA analysis [37], where the largest eigenvalues of the covariance matrix are used in order to identify the principal components, while building a 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 $l_i$ in order to build a three-dimensional geometric representation of a language taxonomy. Points symbolizing different languages in the space of the three major data traits are contiguous if the orthographic representations of the vocabulary meanings in these languages are similar.

## APPENDIX C

The lexical distance $D(l_i, l_j)$ between two dialects $l_i$ and $l_j$ was previously defined; their geographical distance $\Delta(l_i, l_j)$ can be simply defined as the distance between the two locations where the dialects were collected. There are different possible measure units for $\Delta(l_i, l_j)$. We simply use the great-circle angle (the angle that the two locations form with the centre of the Earth).

It is reasonable to assume, in general, that larger geographical distances correspond to larger lexical distances and vice versa. For this reason, in Wichmann *et al.* [31], the diversity [17,18] was measured as the average of the ratios between lexical and geographical distance.

This definition implicitly assumes that lexical distances vanish when geographical distances equal 0. Nevertheless, different dialects are often spoken at the same locations, separated by negligible geographical distances. For this reason, and because a zero denominator in the division involving geographical distances would cause some diversity indexes to become infinite, Wichmann *et al.* [31] arbitrarily added a constant of 0.01 km to all distances.

Here, we used a different procedure that is better motivated. We plotted all the $\frac{23\times 22}{2} = 253$ points $\Delta$ $(l_i, l_j)$, $D(l_i, l_j)$ in a bi-dimensional space. The plot is not linear for high geographical distance but can be quite well fitted by the function $1 - a\mathrm{e}^{-b\Delta(l_i,l_j)}$. Nonlinear regression of all the points gives parameters $a = 0.72$ and $b = 0.024$. The results indicate that a lexical distance of 0.28 is expected between two variants of a language spoken in coinciding locations.

The right choice of constants $a$ and $b$ ensures that the ratio between $D(l_i, l_j)$ and $1 - a\mathrm{e}^{-b\Delta(l_i,l_j)}$ is around 1 for any pair of dialects $l_i$ and $l_j$. A large value of the ratio corresponds to a pair of variants that are lexically more distant and vice versa. It is straightforward to define the diversity of a dialect as

$$V(l_i) = \frac{1}{22}\sum_{j\neq i}\frac{D(l_i, l_j)}{1 - a\mathrm{e}^{-b\Delta(l_i,l_j)}}. \qquad (\mathrm{C}\,1)$$

In this way, locations with high diversity will be characterized by a larger $V(l_i)$, while locations with low diversity will have a smaller one.

Notice that the above definition coincides with the one in Wichmann *et al.* [31] for very small geographical distances (when the function $1 - a\mathrm{e}^{-b\Delta(l_i,l_j)}$ can be approximated by $1 - a + ab\Delta(l_i,l_j)$, the main difference being that, instead of adding an arbitrary value, we obtain it through the output of nonlinear regression.

The diversities (in decreasing order), computed with equation (C 1), are the following (name of location in brackets): Antandroy (Ambovombe), 1.13; Zafisoro (Farafangana), 1.09; Betsimisaraka (Mahanoro), 1.08; Sihanaka (Ambatondranzaka), 1.06; Mahafaly (Ampanihy), 1.06; Tsimihety (Mandritsara), 1.04; Antaisaka (Vangaindrano), 1.01; Betsimisaraka (Fenoarivo–Est), 1.00; Vezo (Toliara), 1.00; Sakalava (Morondava), 1.00; Sakalava (Majunga), 0.99; Sakalava (Maintirano), 0.98; Antaimoro (Manakara), 0.97; Antankarana (Ambilobe), 0.96; Masikoro (Miary), 0.96; Antanosy (Tolagnaro), 0.95; Sakalava (Ambanja), 0.95; Bara (Betroka), 0.94; Antambohoaka (Mananjary), 0.94; Antankarana (Vohemar), 0.93; Betsileo (Fianarantsoa), 0.92; Antankarana (Antalaha), 0.92; Merina (Antananarivo), 0.90.

## APPENDIX D

The vocabulary consists of 200-item Swadesh word lists for 23 dialects of Malagasy from all areas of the island. All the lists are complete. The orthographical conventions of standard Malagasy have been used since, with this choice, most of the informants were able to write the words directly. Most of the dialects already have a written form owing to the regional politics of the 1970s and 1980s. Malagasy orthography is entirely adequate for our purposes since it allows for an exact mapping between orthographical representations and phonemes. A cross-checking of each dialect list was done by eliciting data separately from two different consultants. There was about 90 per cent coincidence between the two independent sources, most of the differences being the result

Table 1. People who furnished the data on Malagasy dialects.

| | | |
|---|---|---|
| Merina | [19] | |
| (Antananarivo) | Serva, Maurizio | |
| Antanosy | Soafara, Joselina Nere | 08 Nov 1987 |
| (Tolagnaro) | Etono, Imasinoro Lucia | 18 Feb 1982 |
| Betsimisaraka | Andrea, Chanchette Généviane | 07 Aug 1985 |
| (Fenoarivo–Est) | Razakamahefa, Joachim Julien | 09 Nov 1977 |
| Sakalava | Sebastien, Doret | 26 Nov 1980 |
| (Morondava) | Ratsimanavaky, Christelle J. | 29 Feb 1984 |
| Vezo | Rakotondrabe, Justin | 02 Aug 1972 |
| (Toliara) | Rasoavavatiana, Claudia S. | 28 Jun 1983 |
| Zafisiro | Ralambo, Alison | 11 Jun 1982 |
| (Farafangana) | Razanamalala, Jeanine | 03 Feb 1980 |
| Antaimoro | Razafendralambo, Haingotiana | 24 Jul 1985 |
| (Manakara) | Randriamitsangana, Blaise | 05 Feb 1989 |
| Antaisaka | Ramahatokitsara, Fidel Justin | 24 Apr 1984 |
| (Vangaindrano) | Faratiana, Marie Luise | 17 Aug 1990 |
| Antambohoaka | Rakotomanana, Roger | 04 May 1979 |
| (Mananjary) | Zafisoa, Raly | 20 Apr 1983 |
| Betsileo | Ramamonjisoa, Andrininina Leon Fidelis | 16 Apr 1987 |
| (Fianarantsoa) | Rakotozafy, Teza | 25 Dec 1985 |
| Bara | Randriantenaina, Hery Oskar Jean | 17 Jan 1986 |
| (Betroka) | Nathanoel, Fife Luther | 26 May 1983 |
| Tsimihety | Raezaka, Francis | 23 Dec 1984 |
| (Mandritsara) | Francine, Germaine Sylvia | 04 May 1985 |
| Mahafaly | Velonjara, Larissa | 21 Apr 1989 |
| (Ampanihy) | Nomendrazaka, Christian | 07 Jun 1982 |
| Sihanaka | Arinaivo, Robert Andry | 06 Jan 1979 |
| (Ambatondrazaka) | Rondroniaina, Natacha | 27 Dec 1985 |
| Antankarana | Andrianantenaina, N. Benoit | 06 Aug 1984 |
| (Vohemar) | Edvina, Paulette | 28 Jan 1982 |
| Antankarana | Randrianarivelo, Jean Ives | 24 Dec 1986 |
| (Antalaha) | Razanamihary, Saia | 07 Sep 1985 |
| Sakalava | Casimir, Jaozara Pacific | 03 Apr 1983 |
| (Ambanja) | Zakavola, M. Sandra | 17 Jul 1984 |
| Sakalava | Ratsimbazafy, Serge | 17 May 1978 |
| (Majunga) | Vavinirina, Fideline | 23 Jun 1970 |
| Antandroy | Rasamimanana, Z. Epaminodas | 05 Jun 1983 |
| (Ambovombe) | Malalatahina, Tiaray Samiarivola | 07 Jul 1984 |
| Masikoro | Mahatsanga, Fitahia | 22 Mar 1976 |
| (Antalaha) | Voanghy, Sidonie Antoinnette | 12 Oct 1981 |
| Antankarana | Baohita, Maianne | 21 Aug 1984 |
| (Ambilobe) | Nomenjana Hary, Jean Pierre Felix | 07 Jun 1980 |
| Sakalava | Hantasoa, Marie Edvige | 02 Nov 1985 |
| (Maintirano) | Kotovao, Bernard | 06 Oct 1983 |
| Betsimisaraka | Rasolonandrasana, Voahirana | 24 Sep 1985 |
| (Mahanoro) | Andrianandrasana, Maurice | 03 Apr 1979 |

of different choices between synonyms or near synonyms. In such cases, differences were settled through discussions between the two consultants and eventually through the help of one or more fellow townsmen (a kind of public debate).

The number of speakers for each dialect varies from a few tens of thousands (Masikoro and Zafisoro) to around three million (Merina).

In table 1, we provide information on the people who furnished the data collected by one of us (M.S.) at the beginning of 2010 with the invaluable help of Joselinà Soafara Néré. For any dialect (except for Merina, for which published lists combined with the personal knowledge of M.S. were used), data were elicited independently from two consultants as explained above. Their names and birth dates follow each of the dialect names.

## REFERENCES

1 Hurles, M. E., Sykes, B. C., Jobling, M. A. & Forster, P. 2005 The dual origin of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and

paternal lineages. *Am. J. Hum. Genet.* **76**, 894–901. (doi:10.1086/430051)

2 Houtman, F. 1603 *Spraeckende woord-boeck inde Maleysche ende Madagascarsche talen met vele Arabische ende Turcsche woorden.* Amsterdam, The Netherlands: Jan Evertsz.

3 van der Tuuk, H. N. 1865 Outlines of a grammar of Malagasy language. *J. R. Asiatic Soc.* New Series **1**, 419–446. (doi:10.1017/S0035869X00160976)

4 Dahl, O. C. 1951 *Malgache et Maanjan: une comparaison linguistique.* Oslo, Norway: Egede Instituttet.

5 Dyen, I. 1953 Review of Otto Dahl, Malgache et Maanjan. *Language* **29**, 577–590. (doi:10.2307/409983)

6 Adelaar, A. 2009 Loanwords in Malagasy. In *Loanwords in the world's languages: a comparative handbook* (eds M. Haspelmath & U. Tadmor), pp. 717–746. Berlin, Germany: De Gruyter Mouton. (doi:10.1515/9783110218442.717)

7 Blench, R. M. & Walsh, M. 2009 Faunal names in Malagasy: their etymologies and implications for the prehistory of the East African Coast. In *Proc. 11th Int. Conf. on Austronesian Linguistics* (*11 ICAL*), *Aussois, France, 22–26 June 2009.* See http://www.rogerblench.info/Language%20data/Austronesian/Malagasy/Malagasy%20wild%20animal%20names.pdf.

8 Adelaar, A. 1995 Borneo as a cross-roads for comparative Austronesian linguistics. In *The Austronesians in history* (eds J. F. Bellwood & D. Tryon), pp. 75–95. Canberra, Australia: Australian National University, ANU E Press.

9 Serva, M. & Petroni, F. 2008 Indo-European languages tree by Levenshtein distance. *EuroPhys. Lett.* **81**, 68005. (doi:10.1209/0295-5075/81/68005)

10 Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A. & Bakker, D. 2008 Explorations in automated language comparison. *Folia Linguistica* **42**, 331–354. (doi:10.1515/FLIN.2008.331)

11 Petroni, F. & Serva, M. 2008 Languages distance and tree reconstruction. *J. Stat. Mech. Theory Exp.* **2008**, p08012. (doi:10.1086/430051)

12 Bakker, D. *et al.* 2009 Adding typology to lexicostatistics: a combined approach to language classification. *Linguist. Typol.* **13**, 167–179. (doi:10.1515/LITY.2009.009)

13 Serva, M. & Petroni, F. 2011 Dialects of Malagasy. See http://univaq.it/∼serva/languages/zlist.pdf.

14 Swadesh, M. 1952 Lexicostatistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 452–463.

15 Swadesh, M. 1955 Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21**, 121–137. (doi:10.1086/464321)

16 Levenshtein, V. I. 1966 Binary codes capable of correcting deletions, insertions and reversals. *Soviet Phys. Doklady* **10**, 707–710.

17 Vavilov, N. I. 1926 Centers of origin of cultivated plants. *Trudi po Prikl. Bot. Genet. Selek.* [Transl. *Bull. Appl. Bot. Genet.*] **16**, 139–248.

18 Sapir, E. 1916 *Time perspective in aboriginal American culture, a study in method.* Geological Survey Memoir 90, Anthropological Series, no. 13. Ottawa, ON: Government Printing Bureau.

19 Greenhill, S. J., Blust, R. & Gray, R. D. 2008 The Austronesian basic vocabulary database: from bioinformatics to lexomis. *Evol. Bioinform.* **4**, 271–283. See http://language.psy.auckland.ac.nz/austronesian.

20 Wichmann, S. *et al.* 2010 The ASJP Database (version 13). See http://email.eva.mpg.de/wichmann/languages.htm.

21 Saitou, N. & Nei, M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **40**, 406–425.

22 Sokal, R. R. & Michener, C. D. 1985 A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **38**, 1409–1438.

23 Barbançon, F., Warnow, T., Evans, S. N., Ringe, D. & Nakhleh, L. 2006 An experimental study comparing linguistic phylogenetic reconstruction methods. In *Proc. Conf. Languages and Genes, UC Santa Barbara, 8–10 September 2006.* See http://www.cs.rice.edu/∼nakhleh/Papers/UCSB09.pdf.

24 Vérin, P., Kottak, C. P. & Gorlin, P. 1969 The glottochronology of Malagasy speech communities. *Ocean. Linguist.* **8**, 26–83.

25 Gudschinsky, S. 1956 The ABC's of lexicostatistics (glottochronology). *Word* **12**, 175–210.

26 Heggarty, P. 2006 Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data and to dating language? In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), p. 183, Cambridge, UK: McDonald Institute for Archaeological Research.

27 Schölkopf, B., Smola, A.J. & Müller, K.-R. 1998 Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299–1319.

28 Blanchard, Ph., Petroni, F., Serva, M. & Volchenkov, D. 2011 Geometric representations of language taxonomies. *Comput. Speech Lang.* **25**, 679–699. (doi:10.1016/j.csl.2010.05.003)

29 Blanchard, Ph., Dawin, J.-R. & Volchenkov, D. 2010 Markov chains or the game of structure and chance: from complex networks, to language evolution, to musical compositions. *Eur. Phys. J. Spec. Top.* **184**, 1–82. (doi:10.1140/epjst/e2010-01232-1)

30 Fouracre, P. 1995–2007 *The new Cambridge medieval history.* Cambridge, UK: Cambridge University Press.

31 Wichmann, S., Müller, A. & Velupillai, V. 2010 Homelands of the world's language families. *Diachronica* **27**, 247–276.

32 Faublée, J. 1983 *Mémoire spécial du Centre d'études sur le monde arabe et du Centre d'études sur l'océan occidental.* pp. 21–30, Paris, France: INALCO & Conseil International de la language française.

33 Adelaar, A. 1995 Asian roots of the Malagasy; A linguistic perspective. *Bijdragen tot de Taal-, Land- en Volkenkunde* **151**, 325–356.

34 Blench, R. M. 2010 The vocabularies of Vazimba and Beosi: do they represent the languages of the pre-Austronesian populations of Madagascar? See http://www.rogerblench.info/Language%20data/Isolates/Vazimba%20vocabulary.pdf.

35 Gray, R. D. & Jordan, F. M. 2000 Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052–1055 (doi:10.1038/35016575)

36 Gray, R. D., Drummond, A. J. & Greenhill, S. J. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483. (doi:10.1126/science.1166858)

37 Jolliffe, I. T. 2002 *Principal component analysis.* Springer Series in Statistics XXIX, 2nd edn. New York, NY: Springer.