Gradual Modifications and Abrupt Replacements: Two Stochastic Lexical Ingredients of Language Evolution

Michele Pasquini
Istituto per le Applicazioni del Calcolo
"Mauro Picone" - CNR, Rome, Italy
michele.pasquini@gmail.com

Maurizio Serva
Dipartimento di Ingegneria e Scienze
dell'Informazione e Matematica,
Università dell'Aquila, L'Aquila, Italy
serva@univaq.it

Davide Vergni
Istituto per le Applicazioni del Calcolo
"Mauro Picone" - CNR, Rome, Italy
davide.vergni@cnr.it

The evolution of the vocabulary of a language is characterized by two different random processes: abrupt lexical replacements, when a complete new word emerges to represent a given concept (which was at the basis of the Swadesh foundation of glottochronology in the 1950s), and gradual lexical modifications that progressively alter words over the centuries, considered here in detail for the first time. The main discriminant between these two processes is their impact on cognacy within a family of languages or dialects, since the former modifies the subsets of cognate terms and the latter does not. The automated cognate detection, which is here performed following a new approach inspired by graph theory, is a key preliminary step that allows us to later measure the effects of the slow modification process. We test our dual approach on the family of Malagasy dialects using a cladistic analysis, which provides strong evidence that lexical replacements and gradual lexical modifications are two random processes that separately drive the evolution of languages.

1. The State of the Art

Applying statistics to determine the degree of similarity between two languages is the founding idea of lexicostatistics and can be traced back at least to the first half of the

Action Editor: Minlie Huang. Submission received: 6 April 2022; revised version received: 3 October 2022; accepted for publication: 19 October 2022.

https://doi.org/10.1162/coli_a_00471

19th century, when French admiral Jules Dumont d'Urville collected comparative word lists from several languages of the Pacific area during the naval expedition aboard the *Astrolabe*, the corvette he commanded from 1826 to 1829. Although he mainly dealt with the geographical aspects of the expedition, in his account of the voyage the idea of comparing terms from different languages with the same meaning clearly emerges (D'Urville 1832).

Glottochronology, the application of lexicostatistical methods with the goal of establishing when a language separated into derived languages, was introduced by Morris Swadesh in the 1950s (Swadesh 1950, 1951, 1952, 1954, 1955). Swadesh's approach, as its main formula clearly reveals, was inspired by the success of the carbon-14 dating technique developed at that time: Swadesh's hypothesis was that, just as the radioactive decay of a carbon-14 atom into a more stable one occurs at a constant rate, the replacement of a term with a synonym in a language is a rare but observable event whose probability rate is constant along the centuries.

In a formula:

$$M(t) = M\,\mathrm{e}^{-\lambda t} \tag{1}$$

where M is the initial number of words in a basic list and $M(t)$ the number of words not yet replaced at time t. Swadesh's estimate for the substitution rate was $\lambda \simeq 0.14$ per millennium (Swadesh 1955).

Following the analogy, the logarithm of the fraction of unchanged terms is proportional to the temporal separation between ancestor and descendant languages

$$t = -\frac{1}{\lambda} \ln \frac{M(t)}{M} \tag{2}$$

just as the logarithm of the ratio between carbon-14 to total carbon is proportional to the age of an archaeological finding or a fossil.

Starting from this assumption, Swadesh only needed a way to determine the number of unreplaced words. For this purpose, he introduced a list of M universal concepts (named after him) and prepared word-lists for different languages corresponding to the same concepts (Swadesh 1950). By means of an accurate linguistic analysis, he was able to count the number of cognate pairs in two languages and by a simple probabilistic reasoning he was able to estimate the number $M(t)$ of unreplaced words in a single language. More often the comparisons concern coeval languages with synchronous lists of M concepts, where experts try to evaluate the number $M(t)$ of non-cognate pairs between two languages; therefore a factor $\frac{1}{2}$ has to be added in formula (2), to obtain the temporal separation between the two coeval languages and the first common ancestor.

After Swadesh, many scholars in the following decades continued along his approach—see, for example, Embleton (1986) and McMahon and McMahon (2005)—and the most sensitive point was surely the linguistic comparison between terms of the same concept from different languages: As it is well known, there are many mechanisms that induce changes in a language, even if we limit ourselves to the lexical sphere. Certainly this is a key issue, especially from an epistemological point of view: Could a scientific investigation rely on deep, thoughtful, but ultimately personal and subjective judgments about the cognacy of words? Can linguists with different origin, training, and experience ensure the reproducibility of a linguistic measure?

However, criticisms of Swadesh's method have been directed mainly to other aspects. In particular, it has often been argued that sets of cognates should be cleaned

of linguistic loans (see, e.g., Starostin 2000); the rate of replacement λ is not universal and depends on concepts (van der Merwe 1966; Dyen and Cole 1967); the probability of retention of older words diminishes (rate of replacement λ decreases over time) (Starostin 2000), and, more recently, the stability ranking of concepts varies for different families of languages (Pasquini and Serva 2021).

Apart from strictly linguistic aspects, the remaining tasks of glottochronology consist of methodologies that computational mathematics had already successfully applied in evolutionary biology: matrices of distances, cladistics, family trees. The only appreciable new proposal has been the introduction of automated algorithms: Instead of deciding whether two words are cognates or not, the words are treated as strings and their similarity is fixed by an algorithm, or a well-defined procedure; for instance, the *Levenshtein distance* (LD) (Levenshtein 1966) (also known as *edit distance*), which corresponds to the minimum number of operations (insertions, deletions, or substitutions) required to obtain one string from another, can be used. This is clearly a first response to what we believe is the most important weakness of Swadesh's method: the objectivity and then the reproducibility from independent researchers. Actually, the search for automated methods in comparative linguistics has aroused increasing interest in recent years. In Section 3 the reader can find a quick review of most popular ones.

A linguist's decision about the cognacy between two words is equivalent to assigning a dichotomic distance, 0 if the two terms are cognates, 1 otherwise. By normalizing the LD between words (Nerbonne and Heeringa 1997) to a real number between 0 and 1, we can measure the distance between words (it is a measure in a strictly mathematical sense). In this way an efficient and automated procedure to construct language distance that obtains very accurate phylogenetic reconstructions and family trees (Serva and Petroni 2008; Petroni and Serva 2008, 2010a,b, 2011) has been introduced.

The state of the art suggests that two issues are ready to be further explored: first, an automated cognacy decision may be taken by using more reliable algorithms; second, the model for the process of word change can be improved with respect to the simple Poisson model historically associated with the radioactive decay. Actually, the standard Swadesh approach does not consider all the random events that lead two cognate terms of different languages to slowly diverge from each other through small changes century after century. But, clearly, the random process of gradual lexical modifications beyond the well-known random process of abrupt lexical replacements also plays a crucial role in language evolution.

In this work we study the random process of gradual lexical modification and we introduce a simple but extremely efficient algorithm able to identify the subset of cognate words among a well-represented family of languages or dialects (the collection of lists must cover all varieties of the family). In this way we keep distinct the replacement random process *à la* Swadesh, which induces the separation into different cognate subsets, from the other lexical random process acting inside the subsets of cognates. We are able to quantify the effects of this second process and by means of a statistic and phylogenetic analysis we show that both processes are fundamental ingredients for language evolution. This means that the lexical modifications process also has the potential to reveal enlightening information for understanding the evolution of a language family.

2. Linguistic Context, Data, and Mathematical Tools

Madagascar (Figure 1) is a wide island (about three times the area of Great Britain), colonized by Indonesian sailors about 1,400 years ago (an essential review of the

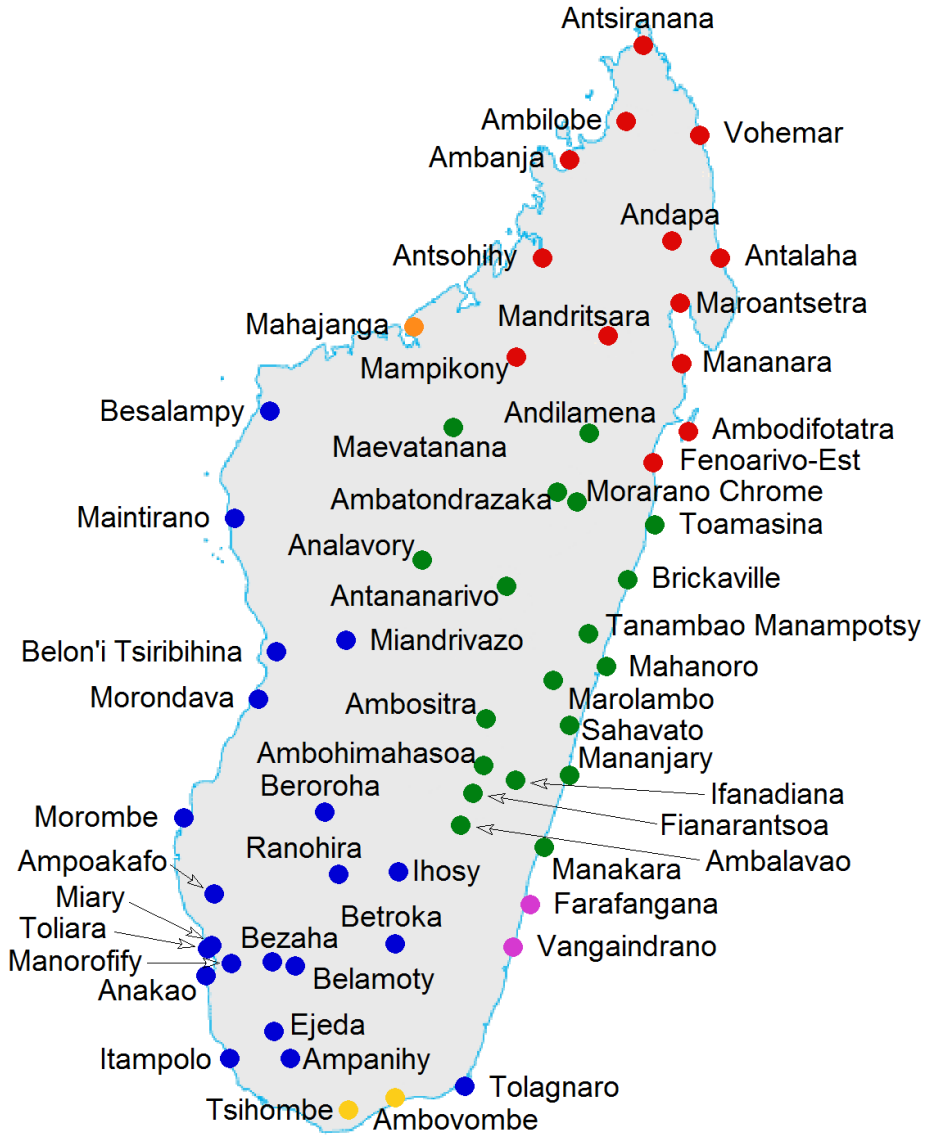

Figure 1
The map of Madagascar with the names of the towns/villages where each Swadesh list was collected. The names of the ethnicities are missing. Colors correspond to the classification proposed in Serva and Pasquini (2020). In both Toliara (south-west coast) and Morondava (west coast), two different dialects coexist, but in Morondava they are close to each other (both are blue), whereas in Toliara they are distant (one is yellow and unreported, the other is blue).

literature on this point can be found in Serva and Pasquini [2020]). Since then, the Austronesian language of the settlers has begun to differentiate into regional varieties as their descendants spread all over the island (Dahl 1938, 1951; Dyen 1953; Dez 1963; Vérin, Kottak, and Gorlin 1969; Hudson 1967; Adelaar 1995, 2006; Beaujard 2003; Blench 2007; Blench and Walsh 2009; Serva 2012). There are two main reasons for choosing Madagascar for our analysis:

(1) The Malagasy language has evolved into dozens of dialects with little external contamination (only rare traces of Bantu words can be found [Dahl 1954; Adelaar 2012; Blench 2008; Serva and Pasquini 2022]);

(2) we have a complete database of Malagasy dialects (*Malagasy Swadesh lists version 1.1 - October 2021*; see Supplementary Material section), covering the entire island with all ethnic groups and all major inhabited centers, consisting of 60 Swadesh lists of 207 terms, entirely collected by one of us (MS) in a two-year span (2018–2019).

Version 1.0 of the database was published in Serva and Pasquini (2020); however, a small number of entries have been updated in the meantime (0.28% of the total, almost all in the Sakalava [Mahajanga] list). The name of each list shows the ethnicity followed by the location in parentheses. We emphasize that in each list every concept has a single corresponding entry, chosen by the interviewer as the most commonly used by the respondents. For the sake of completeness, we also mention that we reduced the Swadesh lists to 205 concepts, neglecting the items *ice* and *snow*, too often misleading for Malagasy speakers.

In this article, N is the number of languages/dialects and M is the number of concepts ($N = 60$ and $M = 205$ for the Malagasy database, respectively); Greek letters $\alpha, \beta, ...$ will always be associated with languages, while roman letters $i, j, ...$ will point to concepts. If we think of Swadesh lists as a series of side-by-side columns, then the database is an $M \times N$ matrix whose generic entry:

$$W_{\alpha,i} \qquad (\alpha = 1, \ldots, N \quad ; \quad i = 1, \ldots, M) \tag{3}$$

is the word used in the language α to represent the i-th concept. The α-th column $\{W_{\alpha,i}\}_{i=1,\ldots,M}$ is the whole Swadesh list for the language α, while the i-th row $\{W_{\alpha,i}\}_{\alpha=1,\ldots,N}$ shows all the different ways in which the i-th concept is expressed in the different dialects.

Following Nerbonne and Heeringa (Nerbonne and Heeringa 1997), we use Normalized Levenshtein Distance (NLD) to measure the degree of similarity between two words $W_{\alpha,i}$ and $W_{\beta,j}$:

$$\mathrm{NLD}(W_{\alpha,i}, W_{\beta,j}) = \frac{\mathrm{LD}(W_{\alpha,i}, W_{\beta,j})}{\max\_\mathrm{length}(W_{\alpha,i}, W_{\beta,j})} \tag{4}$$

that is, the ratio of the $\mathrm{LD}(W_{\alpha,i}, W_{\beta,j})$ between the two words (i.e., the number of insertions, deletions, or substitutions to transform $W_{\alpha,i}$ into $W_{\beta,j}$ or vice versa) and the length of the longest of them. In this way we obtain a real number between 0 (two identical words) and 1 (two completely different words), with all shades in between.

The definition of a distance, $D^{\mathrm{NLD}}_{\alpha,\beta}$, for two languages α and β immediately follows from (4), considering only pairs of words belonging to the same concept ($j = i$) and then averaging over all M concepts:

$$D^{\mathrm{NLD}}_{\alpha,\beta} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{NLD}(W_{\alpha,i}, W_{\beta,i}) \tag{5}$$

This definition of language distance based on NLD has been systematically used since 2008 (Serva and Petroni 2008; Petroni and Serva 2008; Bakker et al. 2009; Petroni and Serva 2010a, 2010b, 2011; Ciobanu and Dinu 2018). This lexical distance can be associated with a genealogical distance $T^{\mathrm{NLD}}_{\alpha,\beta}$—as has been typically done in cladistics studies since the seminal works of Swadesh (Swadesh 1950, 1955)—which is the time depth from the last common ancestor of α and β:

$$T^{\mathrm{NLD}}_{\alpha,\beta} = -\frac{\tau}{2}\ln(1 - D^{\mathrm{NLD}}_{\alpha,\beta}) \tag{6}$$

where τ, measured in millennia, has to be fixed by external information (by historical facts, or taken from another linguistic context [Serva and Petroni 2008; Petroni and Serva 2008]). Notice that $\lambda = \frac{1}{\tau}$ corresponds to the Swadesh number of substitutions per millennium; the extra factor $\frac{1}{2}$ is due to the fact that we are comparing contemporary languages.

Actually, the parameter τ is relevant only if one is interested in determining the time depth of a language family, but can be neglected if the focus is only on phylogenetic reconstruction.

3. Automated Cognate Detection

Attempts to make some aspects of the linguist's work automatic has greatly increased in recent years, especially concerning the detection of cognate pairs. Since the pioneering work of Covington (Covington 1996), many combinations of metrics for word similarity and criteria for cognate set partitioning have been proposed (e.g., see Hauer and Kondrak 2011; List 2012; Ciobanu and Dinu 2014a,b, 2015; List, Lopez, and Bapteste 2016; Jäger, List, and Sofroniev 2017; Rama et al. 2018; Rama and List 2019; and also see List 2014 for an historical overview). Since 2008 NLD distance has been repeatedly used as a metric tool (see List, Greenhill, and Gray 2017 for a recent example), but NLD as well as other metrics are often coupled with partitioning strategies that do not perform particularly well. Some authors limit the decision about cognacy only to the pair under consideration, neglecting how other languages express the same concept; with this approach there is no way to find out that the word *leche* (Spanish) and γάλα (Greek) are actually cognates. Other scholars, more cleverly, take all the decisions about cognacy for the same concept as a single step using partitioning algorithms, such as UPGMA-like variants, where a global threshold or a maximal threshold is used to create the subsets of cognates. Our approach, as it will be clear in the following, is different since the threshold only is used to create links, without any restriction on average distance or maximal distance in a subset. In this way the detection of *leche* and γάλα as cognates only needs a sufficiently large number of intermediate words. This is why a broad collection of the varieties of a language family is a crucial aspect of our methodology.

We also think that NLD is the right metric to be used since it is the only one sufficiently sensitive and precise to grasp the gradual modifications and not confuse

them with abrupt replacements. NLD distance deals with deletions, insertions, and substitutions of elements in a string, which are the genuine essence of lexical modifications; therefore, it is the natural tool to handle this kind of process. On the other hand, when a replacement occurs it is reasonable to expect, at least from a probabilistic point of view, that the new synonym is not a cognate of the previous one. We can provide a couple of examples: classical Latin *caput* (head) replaced by *testa* in late Latin (original meaning: clay pot); classical Latin *caseus* (cheese) replaced by *formaticum* in late Latin, from *forma* (mold) + the noun-forming $-$*aticum* suffix. In both cases the replacing synonym and the original term are not cognates.

To show how NLD deals with pairs of non-cognate words, we use the Malagasy database. We randomly choose pairs of words associated with different concepts from different dialects that surely are not cognate, we compute the NLD distances between these words, and we display the distance statistics. Figure 2 (purple bars) shows the result: Almost all the distribution is concentrated in the right half of the NLD axis ($0.5 \leq \text{NLD} \leq 1$). NLD distribution for pairs of words corresponding to the same concept in different languages (which can or cannot be cognate) is plotted in the inset of Figure 2 (red bars): Apart from about 30% of identical words, the distribution covers more or less uniformly the NLD axis. This distribution is the superposition of the distribution of distances between cognate pairs (still unknown) and the distribution of distances between non-cognate pairs (represented by purple bars in Figure 2). The last

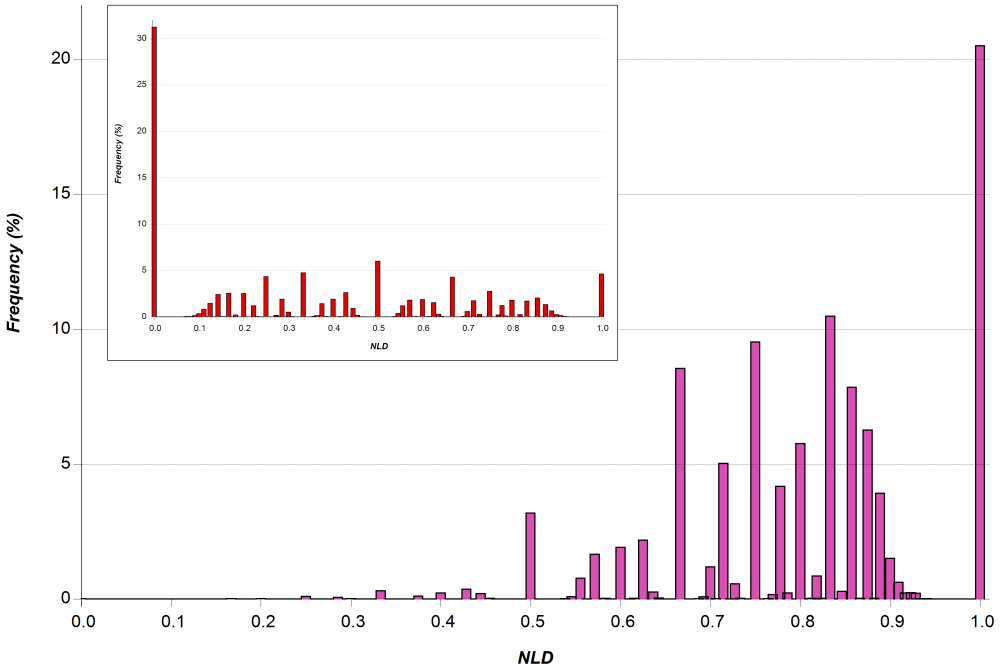

Figure 2
Main picture (purple bars): Percentage frequency distribution of NLD out of more than 600,000 random draws of pairs of terms from the Malagasy database. Each pair is composed of two words, $W_{\alpha,i}$ and $W_{\beta,j}$ which belong to different languages ($\alpha \neq \beta$) and to different concepts ($i \neq j$). The first half of NLD values ($NLD < 0.5$) has a probability less than 1.5%. Inset (red bars): Percentage frequency distribution of NLD for pairs of words of the same concept ($\alpha \neq \beta$, $i = j$).

remark suggests a different approach in introducing a threshold for automated cognacy: Instead of using the threshold to identify which pairs are surely not cognate, since a certain value has overcome the threshold (the usual policy), we can use it just to confirm which pairs are surely cognates. By comparing the two distributions in Figure 2, we can conclude that if a pair has an NLD less than about 0.5, the two words almost certainly are cognates. At the end of this section, a novel, objective methodology for obtaining the optimal threshold for identifying two direct cognate terms will be presented.

Here, an algorithm for automated cognacy detection is introduced. Given a *cognacy threshold* D_T and a given concept i, all possible $N(N-1)/2$ pairs of words are checked and a direct cognacy link, $L^i_{\alpha,\beta}$, is established for those pairs whose NLD is below the threshold:

$$L^i_{\alpha,\beta} = \begin{cases} 1 & \text{if} \quad \text{NLD}(W_{\alpha,i}, W_{\beta,i}) < D_T \\ 0 & \text{if} \quad \text{NLD}(W_{\alpha,i}, W_{\beta,i}) \geq D_T \end{cases} \quad \Leftrightarrow \quad \text{direct cognacy link} \tag{7}$$

In this way, the N words associated with the concept i are split into a certain number of non-overlapping subsets or, more precisely, a certain number of disjoint subgraphs. In fact, an unweighted undirected graph, $G_i = (V_i, E_i)$, can be defined for each given concept i, where V_i, the set of vertices, are the N words, $\{W_{\alpha,i}\}_{\alpha=1,\ldots,N}$, associated with the i-th concept in the different dialects α and E_i, the set of undirected edges, are composed of those paired words, $(W_{\alpha,i}, W_{\beta,i})$, with a direct cognacy link ($L^i_{\alpha,\beta} = 1$; i.e., their NLD is below the threshold D_T).

With this simple link-generation rule, the graph G_i is naturally divided into connected subgraphs whose vertices are naturally identified as the subsets of cognates. In other words, we define cognates as all the words that belong to each given subgraph and, conversely, two words belonging to different subgraphs are not cognate. It is very important to stress that, even if two words are in the same subgraph (that is, are cognates), their NLD is not necessarily less than the cognacy threshold D_T (see the following example of concept *ash* and Figure 1).

The connection with graph theory gives us a powerful context to formalize the interesting quantities to be studied and it will prove to be very useful in the next sections of the article. For example, we can easily handle the cognacy relation between a pair of words $W_{\alpha,i}$ and $W_{\beta,i}$ exploiting definition (7) to introduce a discrete variable $C^i_{\alpha,\beta}$, whose value is 1 if words are cognates (i.e., if vertices α and β belong to the same subgraph), 0 otherwise:

$$C^i_{\alpha,\beta} = \begin{cases} 1 & \text{if exist } \{\gamma_j\}_{j=1,\ldots,n} \text{ such that: } L^i_{\alpha,\gamma_1} \cdot L^i_{\gamma_1,\gamma_2} \cdot \ldots \cdot L^i_{\gamma_{n-1},\gamma_n} \cdot L^i_{\gamma_n,\beta} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

That is, two words are cognates if they are linked by a sequence of direct cognates.

An example of well-known words can help to fix the idea. Consider the concept *ash* in the languages listed in Table 1. Using the above-discussed algorithm, the words are split into two disjoint subgraphs (Germanic family and Romance family); therefore in Figure 3 we have two subsets of cognates: *as, asche, aschen, ash, aska, aske, jiske* (blue) and *cender, cendra, cendre, cenere, ceniza, cenusa, cinza* (red). Each line indicates a direct cognacy link, that is, the NLD distance of pairs of words is below the cognacy threshold $D_T = 0.5$, which results in an undirect edge between the two terms. Figure 3 shows the subgraphs resulting from the application of the automated cognacy procedure, and, in particular, it is important to stress that words belonging to the same subgraph are

Table 1
Concept *ash* expressed in some Germanic and Romance languages commonly spoken in western Europe.

Language	Word	Language	Word
Catalan	*cendra*	Italian	*cenere*
Danish	*aske*	Ladin	*cender*
Dutch	*as*	Luxembourgish	*aschen*
English	*ash*	Portuguese	*cinza*
French	*cendre*	Romanian	*cenusa*
Frisian	*jiske*	Spanish	*ceniza*
German	*asche*	Swedish	*aska*

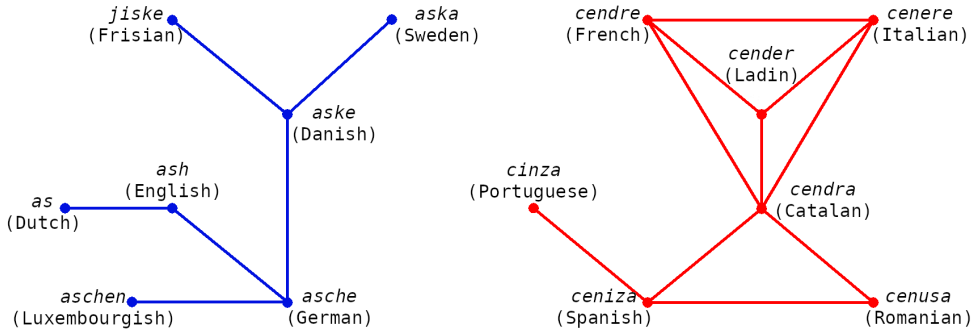

Figure 3
Example of cognacy subgraphs, one for the Germanic family (on the left, in blue) and one for the Romance family (on the right, in red) associated with the concept *ash* ($D_T = 0.5$).

considered cognates but they do not necessarily have a direct link: For example, *aschen* (Luxembourgish) and *jiske* (Frisian) have NLD $= 5/6 \simeq 0.83 > D_T$, and *cenere* (Italian) and *cinza* (Portuguese) have NLD $= 4/6 \simeq 0.67 > D_T$. However, their cognacy relation is assured by the fact that they belong to the same subgraph given the presence of the intermediate words *asche* (German) and *aske* (Danish) in the Germanic family, and *cendra* (Catalan) and *ceniza* (Spanish) in the Romance family.

This small example clearly highlights the peculiarities of our algorithm: It is fast and it has a single parameter D_T, but it needs a large representation of languages. If we did not know the existence of *ceniza* (Spanish) we would have lost the cognate connection between *cinza* (Portuguese) and the rest of the Romance subset. This is the most important reason to use the Malagasy dataset that fully covers all varieties.

An objective method for determining the optimal threshold depending on the language family under consideration will now be discussed. Choosing a value for D_T, the automated algorithm returns back the cognate subsets for each concept i, and the resulting NLD distribution for all non-cognate pairs g(NLD) can be easily computed. We have previously obtained an example of the distribution g(NLD) simulating it in Figure 2 by means of f(NLD) (purple bars), the distribution of pairs of terms taken from different concepts and different languages (which are definitely not cognates). A strategy to fix the only parameter of our approach immediately follows: The optimal D_T minimizes the distance between g(NLD) and f(NLD).

In order to compare these two distributions, we need to slightly modify them. First, we go from percentage frequency to absolute frequency (by simply dividing by 100); then, we split the range of values of NLD ($0 \leq NLD \leq 1$) into a certain number of successive intervals labeled by k, computing the cumulative absolute frequencies $\{g_k\}$ and $\{f_k\}$ for both distributions in all intervals. The interval width is not fixed, but has been made variable so that an adequate number of distances between non-cognate pairs falls in each interval; in other words, we always have values f_k that are not too small to carry out a statistically significant computation. Just to fix the idea, in the first half $0 \leq NLD < 0.5$ we have few large intervals, while in the second half $0.5 \leq NLD \leq 1$ the intervals are much more dense and smaller. Our criterion is that f_k reaches at least the minimal value of 0.5% in each interval ($f_k \geq 0.005\,\forall k$, since the $\{f_k\}$ are absolute frequencies).

The distance between the distribution $\{f_k\}$ (different languages, different concepts, forced non-cognates pairs resulting from the data collected on field) and the distribution $\{g_k\}$ (different languages, same concepts, non-cognate pairs resulting from our algorithm) can be measured with Le Cam distance (Le Cam 1986)—a normalized statistical distance derived from the symmetrized version of Pearson χ^2 divergence—defined as:

$$\chi_{LC} = \sqrt{\frac{1}{2}\sum_k \frac{\left(f_k - g_k\right)^2}{f_k + g_k}} \tag{9}$$

The results for the Malagasy database can be appreciated in Figure 4, where the χ_{LC} distance is plotted at varying of the threshold D_T. As it was already possible to see from Figure 2, the best reconstruction of non-cognate NLD distribution is performed by automated cognate detection around the value 0.5 for the threshold D_T. Indeed, numerical data show that the objective choice is precisely $D_T = 0.5$. Finally, the minimum of Le Cam distance $\chi_{LC} \simeq 0.18$ means a good agreement between real non-cognate NLD distribution and our equivalent automated reconstruction.

In conclusion, the unique parameter D_T is objectively chosen to be 0.5. Using this value, for each concept we can identify all cognate pairs and measure their NLD distances, as we can do with the complementary set of non-cognate pairs. These distances are the ingredient of all analysis and results in next sections. Finally, we would like to mention the work of Rama et al. (2018), where threshold is automatically inferred for each meaning by a graph approach.

4. Madagascar: Analysis of the Results

The application of our automated cognate detection algorithm to the Malagasy dataset with the optimal threshold $D_T = 0.5$ reveals to be a source of considerable non-trivial information.

As a first test, we have plotted the NLD distribution of all cognate pairs, which is the counterpart of the non-cognate NLD distribution of Figure 2 (main figure). The result can be found in Figure 5, where the purple bars show that a non-negligible portion of cognates (about 19%) have NLD equal to or greater than the threshold. This means that the algorithm provides meaningful connections within the set of words associated with the same concept, revealing non-trivial cognate pairs. Notably, some of the cognate terms are totally different (NLD=1), which is the same peculiar result we have seen for the cognate words *leche* and γάλα.

Because all varieties of our dataset are also identified by the geodesic coordinates of the town or village where the list was collected, we have the opportunity to put the cognate subgraphs in a geographical context, and this procedure can provide significant information. We are able, in fact, to detect a significant relation between geographical and linguistic proximity. This is a consequence of the osmosis among the vocabulary of populations that live nearby. This phenomenon has been studied and quantified in Serva et al. (2017) and, in perspective, it can be used to eventually detect migration events that have occurred in the past history of Madagascar.

Some clarifying examples can be found in Figure 6, where the resulting subsets of cognates of four concepts (*fish*, *guts*, *tree*, and *woman*) are represented: Each color corresponds to a different subset of cognates and two towns/villages are joined only if the corresponding terms have a NLD below the threshold D_T (direct cognacy link). Checking the figure carefully, it can be noted that some dots are not directly connected to all other dots of the same color since the NLD distance between the corresponding words is above the threshold, despite belonging to the same cognate subgraph.

The phenomenology shown by figures is another relevant evidence of the goodness of our automated cognate detection technique; in fact, divisions into linguistic subsets for words have a clear geographical equivalent for the corresponding locations on the map. This can be visually perceived for many concepts—a more accurate quantitative test of this correspondence could be an argument of future research.

Quantitative linguistic analysis concerning Malagasy dialects is rare in the literature and limited to a small number of varieties and a few words. Our database is by far the

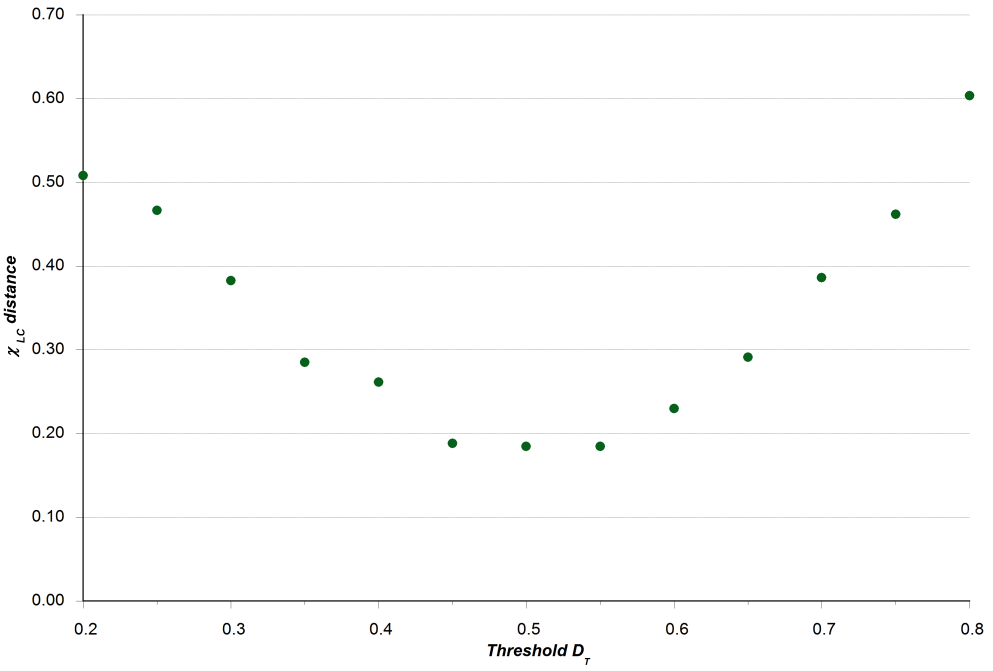

Figure 4
χ_{LC} distance between NLD distributions of forcefully non-cognate pairs $\{f_k\}$ and automated non-cognate pairs $\{g_k\}$, as a function of threshold D_T. The range $0.45 \leq D_T \leq 0.55$ is almost flat, but numerical data show that the minimum is reached at $D_T = 0.5$.

most reliable resource of knowledge about the Malagasy language from the point of view of lexicostatistics. We hope that the sharing of our free dataset arouses the interest of linguistics experts into this kind of investigation so as to compare our automated detection tool with detection performed by experts with traditional techniques. In the meanwhile, the best we can do is to compare our approach to the LexStat-Infomap algorithm (List, Greenhill, and Gray 2017; Rama et al. 2018), one of the most popular and efficient tools for the task of cognate detection, setting its threshold to 0.55, as reported in the above-quoted papers. To evaluate the similarity of the two groups of cognate subsets we compute the B-cubed F-scores (Amigó 2009); this turns out to be 0.94, indicating a high degree of agreement between the two procedures. The comparison between our algorithm and the LexStat-Infomap method will be further explored in the last section, devoted to phylogenetic reconstruction.

5. The Random Process of Gradual Lexical Modifications

As we have already mentioned, Swadesh and subsequent scholars have always exclusively used the lexical replacement process to determine the composition of language families, their temporal depth, and the moments of internal separations. The gradual modification of cognate terms has never been considered as a possible source of information for language evolution.

Let's consider the typical situation of two languages that begin to differentiate due to an elementary cause, physical distance, as for example has happened since

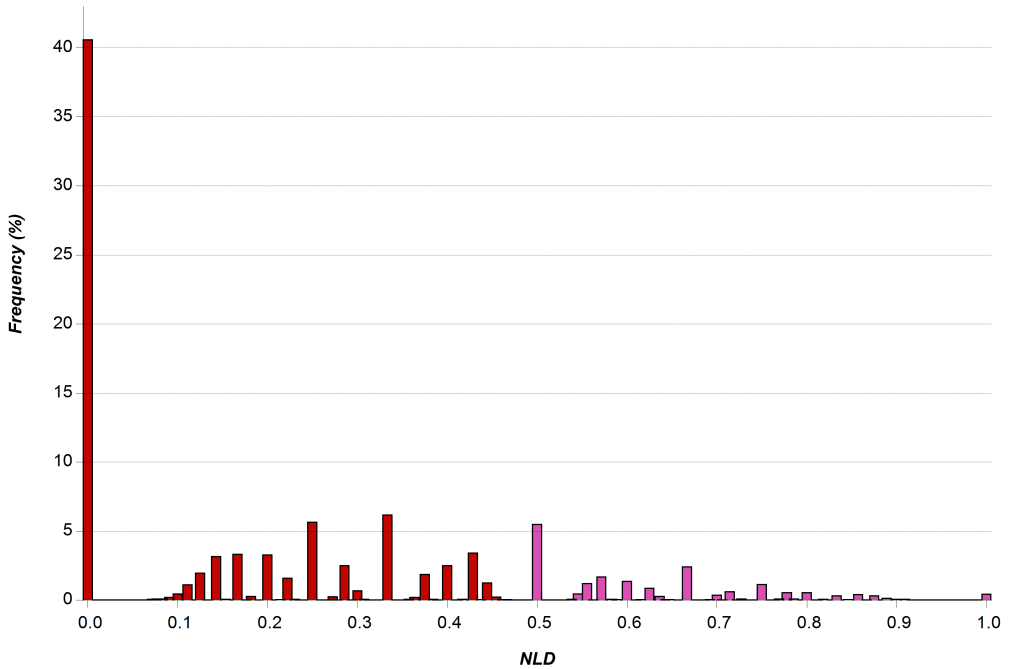

Figure 5
Percentage frequency NLD distribution for cognate pairs according to our automated detection procedure. The red and purple bars distinguish pairs below and over the threshold $D_T = 0.5$. The total amount of percentage frequency over the threshold is about 19%.

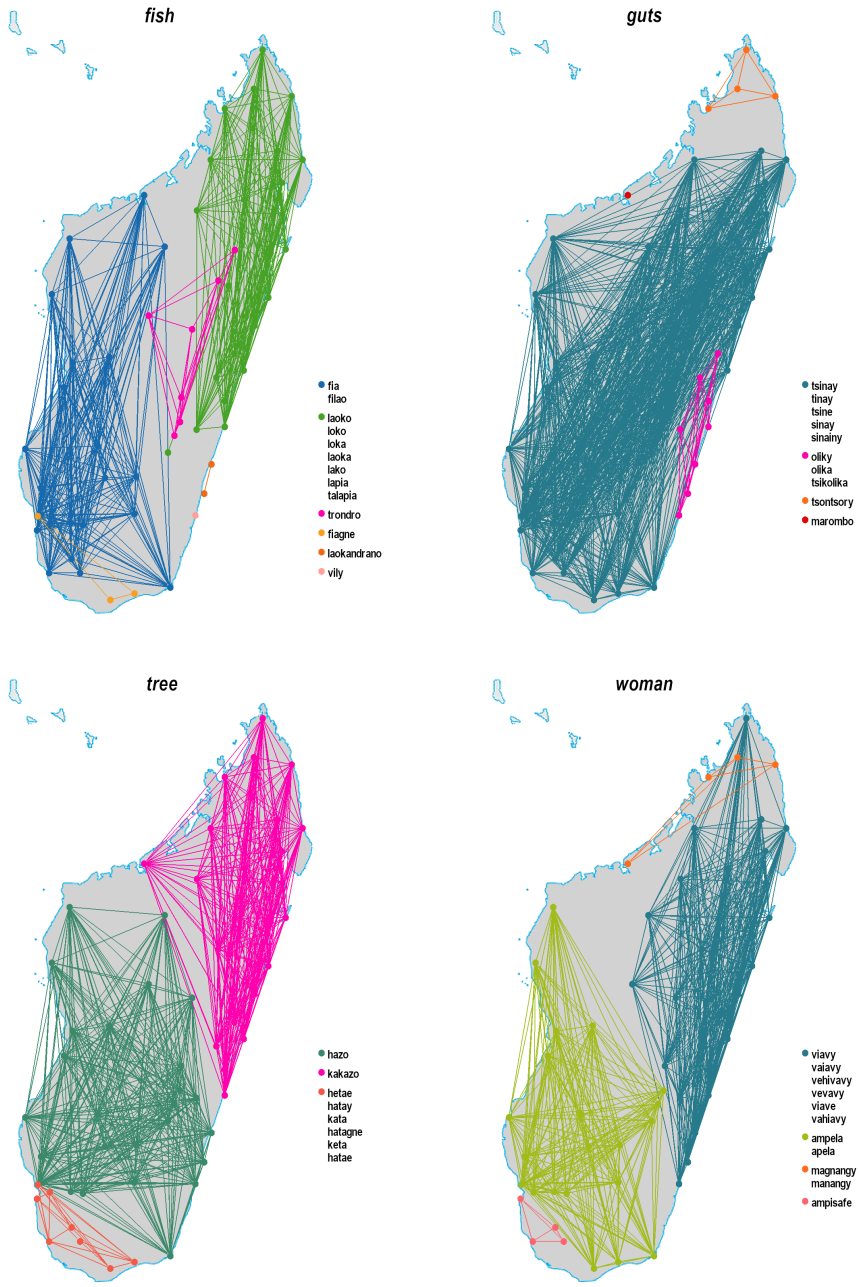

Figure 6
Result of the automated cognate detection for the *fish*, *guts*, *tree*, *woman* concepts of the Malagasy database with threshold $D_T = 0.5$. Each color identifies a different subset of cognates (terms are in the legend). The dots on the map geographically locate where a word of that subset has been collected, while the lines indicate pairs of words whose NLD is less than the direct cognacy threshold D_T.

the 9th century to the old Norse spoken in Scandinavia and the old Norse of the first Scandinavian settlers of Iceland. The classic approach of glottochronology is to identify those terms of the common original language that along the following centuries were replaced in Norway or in Iceland. The timing of a replacement is not *a priori* predictable; every year there is a very small probability that it take place for a given concept. Therefore, the random replacements are well described by a stochastic process (specifically a Poisson process).

The lexical replacement can be considered an event of a certain relevance; but also minor changes, which we call **gradual lexical modifications**, that frequently occur in the lexicon (modification of a vowel, truncations, or final additions, etc.) have a central role, as we will show later in this article. Modifications may be due to different causes, such as vowel reduction, consonant and vowel shift, morphological truncation, consonant lenition, and other causes that are not always identifiable. This phenomenon is the counterpart of the gradual genomic random modification in biology, and it may happen simply because the transfer of vocabulary from one generation to the next is forcefully imperfect.

The term associated with a concept, initially identical for the two populations, can, over the centuries, be altered by little changes either in Scandinavia or Iceland (or both, almost surely in a different way!): A linguist has almost no difficulty in recognizing the close relationship of the two new words and classifiying them as cognates. Even these small changes can be considered a stochastic process, which it is reasonable to assume occurs at a constant rate: The effects are smaller, but at a faster rate than lexical replacements.

In conclusion, we have shown that from the statistical analysis of the differences between cognates, it is reasonable to expect qualitatively similar information to that obtained from the analysis of lexical replacements. But there is an important difference: In lexical replacement, the change takes place (if it occurs) in a definitive way and two words either remain cognates, or not. In this context, the only reasonable measure to assign to their distance is the dichotomic one: 0 for cognate words, 1 otherwise. Conversely, in gradual lexical modifications between cognates, the changes can repeat and add up and a quantitative measure can be introduced to quantify the degree of similarity: In this case, NLD is the right measure to use. Our idea is first to compare the two different lexical distances (dichotomic distance and NLD) for the two different random processes (respectively, abrupt replacements and gradual modifications), then to use both of them in order to get more information with respect to that furnished by the lexical replacements distance alone.

The lexical replacements distance $D^{\mathrm{R}}_{\alpha,\beta}$, between two languages α and β, is simply the ratio between the number of non-cognate terms and the total number of concepts. In fact, for each concept i the distance of the pair $W_{\alpha,i}$ and $W_{\beta,i}$ is fixed to 0 if they are cognates, 1 otherwise; $D^{\mathrm{R}}_{\alpha,\beta}$ immediately follows averaging over all the M concepts. Recalling that $C^{i}_{\alpha,\beta}$ carries the information about the cognacy of two words (see Equation (8)), we have that the above dichotomic distance is $(1 - C^{i}_{\alpha,\beta})$, which implies

$$D^{\mathrm{R}}_{\alpha,\beta} = \frac{1}{M}\sum_{i=1}^{M}\left(1 - C^{i}_{\alpha,\beta}\right) = 1 - \frac{M_{\alpha,\beta}}{M} \tag{10}$$

where

$$M_{\alpha,\beta} = \sum_{i=1}^{M} C^{i}_{\alpha,\beta} \tag{11}$$

is the number of concepts for which the terms of the languages α and β are cognates ($0 \leq M_{\alpha,\beta} \leq M$). Eq. (10) is the traditional lexical distance between languages used in glottochronology since Swadesh's works in the 1950s.

The distance $D^{\mathrm{M}}_{\alpha,\beta}$ associated with the random process of gradual lexical modifications, which uses NLD when words are cognates, reads as follows:

$$D^{\mathrm{M}}_{\alpha,\beta} = \frac{1}{M_{\alpha,\beta}} \sum_{i=1}^{M} C^{i}_{\alpha,\beta} \cdot \mathrm{NLD}(W_{\alpha,i}, W_{\beta,i}) \tag{12}$$

Let us stress that the presence of $C^{i}_{\alpha,\beta}$, both explicit as a multiplicative factor of NLD and implicit in the $M_{\alpha,\beta}$ definition, has the effect of constraining the average only on those concepts where languages are cognates, while lexical replacement is not considered at all. This choice is consistent when considering only gradual lexical modifications, acting within a subset of cognates.

Two different language distances given by two different mechanisms (lexical replacements and gradual lexical modifications) has been introduced and a test of mutual consistency is required. In Figure 7 the distance $D^{\mathrm{M}}_{\alpha,\beta}$ as a function of the distance $D^{\mathrm{R}}_{\alpha,\beta}$ is shown for all the pairs of languages. At a glance, the data reveal a good proportionality between the two distances, well confirmed by the statistical analysis (correlation 0.73). This is a very interesting result: Although the two processes of lexical modification and lexical replacement are very different mechanisms of linguistic evolution, they both contribute, with a good relationship, to the modification of a language. Therefore we can safely affirm that the random process of gradual modification of the language is a real phenomenon and that it is reasonable to expect that a cladistic reconstruction of the evolution of Malagasy from the $D^{\mathrm{M}}_{\alpha,\beta}$ distance will also provide reasonable results.

Because our hypothesis is that the two stochastic processes of gradual lexical modifications and abrupt lexical replacements occur in parallel, it is natural to introduce a new overall measure that merges the two previously introduced measures. Let's take a step back to the definition of distance between two words, $W_{\alpha,i}$ and $W_{\beta,i}$, related to the same concept i. If these terms are cognates ($C^{i}_{\alpha,\beta} = 1$), then the best definition of distance is clearly the NLD, because it is gradual and sensitive to small variations; if they are not ($C^{i}_{\alpha,\beta} = 0$), there is no relationship between the two words and so it is natural to assign the maximum possible distance, that is, 1.

In other words, a new measure of similarity between words can be introduced by merging $D^{\mathrm{M}}_{\alpha,\beta}$ and $D^{\mathrm{R}}_{\alpha,\beta}$, choosing the NLD for the cognates and the dichotomic distance otherwise. All that remains is to average out all the concepts to define a new distance between languages, $D^{\mathrm{MR}}_{\alpha,\beta}$. In symbols:

$$D^{\mathrm{MR}}_{\alpha,\beta} = \frac{1}{M} \sum_{i=1}^{M} \left[C^{i}_{\alpha,\beta} \cdot \mathrm{NLD}(W_{\alpha,i}, W_{\beta,i}) + \left(1 - C^{i}_{\alpha,\beta}\right) \right] \tag{13}$$

Comparing this last distance with the initial $D^{\mathrm{NLD}}_{\alpha,\beta}$ in (5), the improvement obtained is evident. When two words are cognates ($C^{i}_{\alpha,\beta} = 1$), the element in the sum in Equation (5) coincides with the corresponding element in (13), but when they are not, $D^{\mathrm{NLD}}_{\alpha,\beta}$ keeps using NLD, causing loss of information; in fact, we are now handling two unrelated words and their typical $0.5 < \mathrm{NLD} < 1$ value (see Figure 2) is somewhat incidental, due to spurious lexical coincidences, while logically it has to be 1 (as it is according to $D^{\mathrm{MR}}_{\alpha,\beta}$).

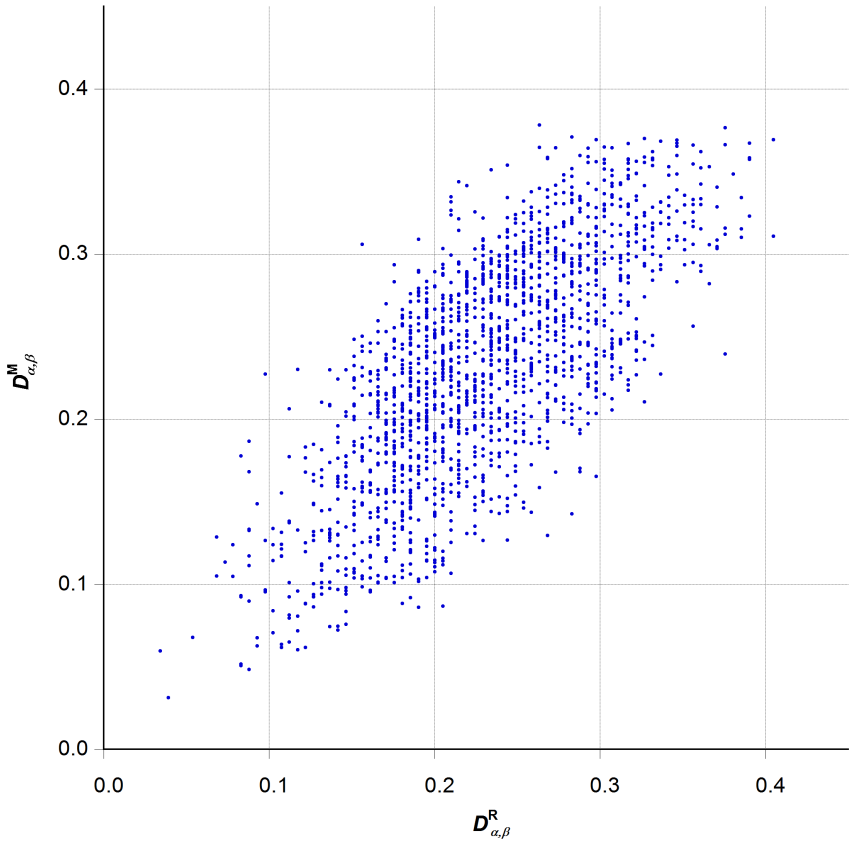

Figure 7
Language distance $D^{M}_{\alpha,\beta}$ as a function of the language distance $D^{R}_{\alpha,\beta}$ for all the $N(N-1)/2 = 1,770$ couples of dialects of the Malagasy database (cognacy threshold $D_T = 0.5$). The data exhibits a very significant correlation of 0.73, confirming the working hypothesis.

6. The Definitive Test: The Comparison on Cladistics

Every linguistic model, every hypothesis on the temporal evolution of languages, must eventually be verified through a realistic application: cladistics, or phylogenetic reconstruction. Up to now we have examined the language distances associated with the processes of lexical replacements ($D^{R}_{\alpha,\beta}$) and gradual lexical modifications ($D^{M}_{\alpha,\beta}$), while for the combination of the two ($D^{MR}_{\alpha,\beta}$) and for the NLD language distance ($D^{NLD}_{\alpha,\beta}$) we have just given the definition.

However, regardless of the chosen distance, the values of distance for each language pair (α, β) has been treated individually and no comparative study of the various languages has been made, that is exactly what cladistics does. Cladistics examines the family of $N(N-1)/2$ distances $D^{X} = \{D^{X}_{\alpha,\beta}\}_{1 \leq \alpha < \beta \leq N}$ as a whole (where X means R, M, MR, or NLD), checking if there is an actual overall coherence, a structural relationship between different language families. The test is significant because the phylogenetic

reconstruction must be consistent with other information available about the populations involved (from history, geography, anthropology, etc.).

In order to perform this analysis, the four families of lexical distances D^{R}, D^{M}, D^{MR}, and D^{NLD}, are first transformed into their equivalent genealogical distances T^{R}, T^{M}, T^{MR}, and T^{NLD}, as in Equation (6). We have chosen to calculate Unweighted Pair Group Method Average (UPGMA) trees, which are best suited for temporal analysis. The cladograms are drawn in Figure 8, where each branch (each Swadesh list of the Malagasy database) is identified by the ethnicity and the location (in parentheses) where the list was collected (the geographical positions can be found in Figure 1).

The NLD case is used as term of comparison; the associated tree was already published in Serva and Pasquini (2020) and the colors of the branches are maintained here (the same as the map in Figure 1). The confidence we place in the quality of the NLD tree is confirmed in its excellent agreement with other external information: The main divisions exactly correspond to the geographical locations (red in the north, green in the center and east, blue in west and south-west, yellow in south); the ethnic groups are all preserved and even minor details are correct, such as, for instance, the partially secluded position of the Mikea (Ampoakafo) leaf that corresponds to the most isolated hunter-gatherer population of Madagascar; the total depth of the tree fixes the root around 650 CE (borrowing for parameter τ the value for Romance family; see Serva and Pasquini [2020] for details), in remarkable agreement with estimates from genetics and archeology. However, there is a minor difference between the T^{NLD} cladogram of Figure 8 and the analogue published in Serva and Pasquini (2020)—that is, the Sakalava (Mahajanga) branch fits a little differently. This is due to the few corrections introduced in the database since the publication of Serva and Pasquini (2020). Although Sakalava (Mahajanga) now seems more linked to the northern group, it continues to maintain an intermediate position between the red and green blocks (not surprisingly it was classified with an exclusive color orange, as Mahajanga is historically a town of maritime trade, inhabited for a long time by different ethnic groups, many arrived from the hinterland of the island).

Returning to our main hypothesis, that the language separation process consists of two distinct independent random mechanisms (gradual lexical modifications and abrupt lexical replacements), let us take a look at T^{R} and T^{M} trees, keeping the T^{NLD} cladogram in mind.

At a glance, it is clear that both reconstructions substantially maintain the main geographical structure and almost always correctly bring together lists belonging to the same ethnic group. The most significant result is that the cladogram based on gradual lexical modifications, T^{M}, turns out to be surprisingly accurate and this is clear evidence that the stochastic process of gradual lexical modifications within cognates is an essential element in language evolution. Conversely, the T^{R}-based UPGMA tree, while maintaining a substantial internal coherence, shows more imperfections. Let us remember that the T^{R} tree relies on lexical replacements, which is the only random process taken into account by Swadesh and subsequent scholars. Surprisingly, the neglected (until now) stochastic process of gradual lexical modifications provides better results.

In summary, cladistic analysis confirms the dual nature of the language evolution process proposed here. Combining both random processes in the T^{MR} tree, the cladogram seems to be almost identical to the T^{NLD} one, made only of NLD distances. This is not that surprising: When words are cognates, both use NLD distance; when they are not, the combined measures shift to 1 while the NLD distance gives a distance between 0.5 and 1, as in Figure 2.

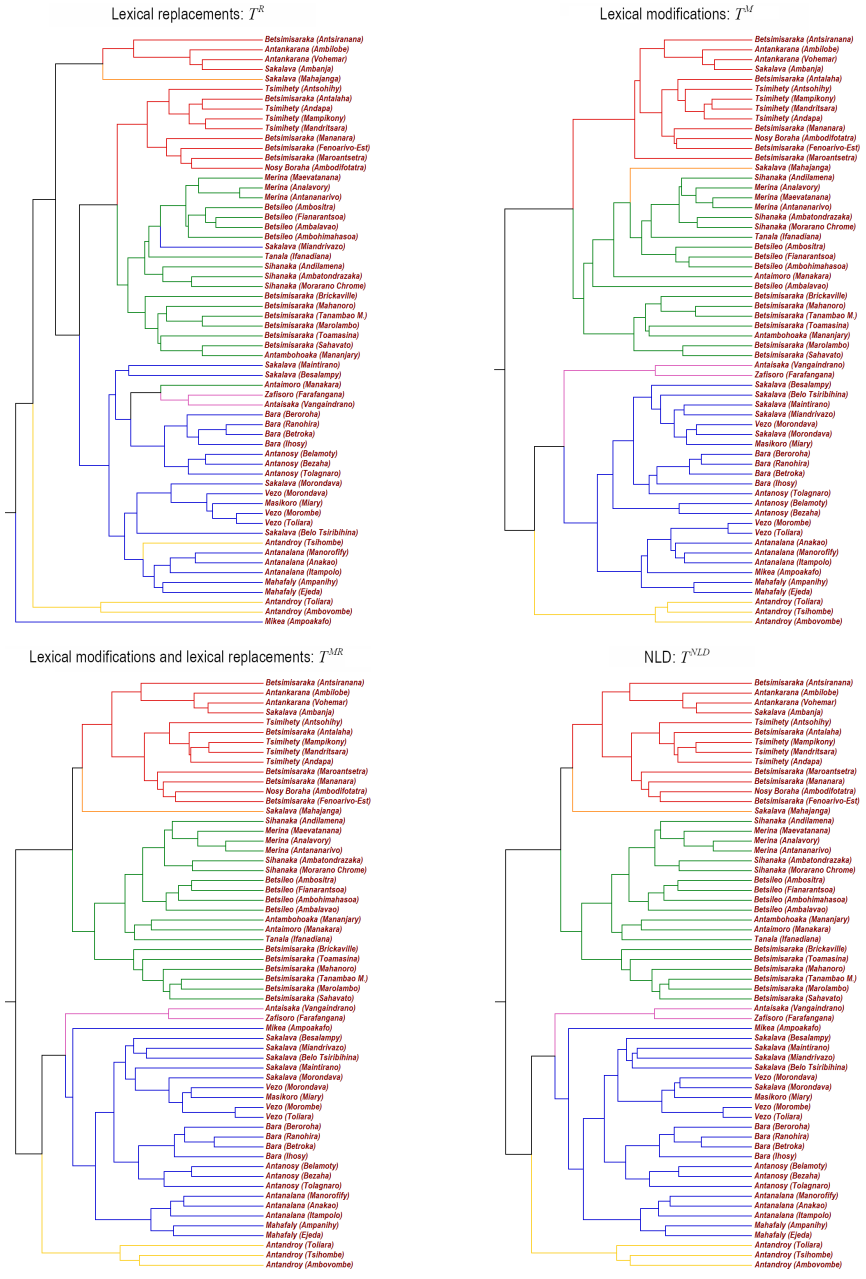

Figure 8
UPGMA trees of Malagasy from the genealogical distances T^{R}, T^{M}, T^{MR}, and T^{NLD}, respectively, related to lexical replacements, gradual lexical modifications, combination of both stochastic processes, and pure NLD distances. Both T^{R} and T^{M} trees give substantially correct phylogenetic reconstructions, only a few misplacements can be seen in both of them. It should be noted that these two trees are generated by the effect of completely separated random phenomena. The remarkable similarity between T^{MR} and T^{NLD} is expected due to the closeness of their definitions.

Nevertheless, the relevant point is not that combining the information of lexical replacements and gradual lexical modifications better describe the cladistics of varieties, but that both random processes can be separately and successfully used to build good quality trees. Incidentally, this is rather an *a posteriori* explanation of why NLD-based language distance gives such good results; it is because it reads the random process involving cognate words well and loses only a small amount of information concerning lexical replacements. In our opinion, this last point is particularly relevant. Comparisons are often reported in literature in which the NLD distance is not always competitive compared to other types of lexical metrics. The reason can be found here: NLD has a high sensitivity with respect to small changes, therefore is the perfect tool when words are cognate (lexical modifications), but when they are not (lexical replacements) its sensitivity captures spurious coincidences altering the result. If the cognacy relation is known, the perfect distance is D^{MR}.

Finally, we use again the results of the LexStat-Infomap algorithm (see Section 4) to build up the equivalent of our T^{R} cladogram (see upper left corner of Figure 8). A direct comparison can be appreciated in Figure 9, where it is evident that the two phylogenetic reconstructions are close to each other, confirming the reliability of our approach. However, both contain a few wrong placements, which, on the contrary,

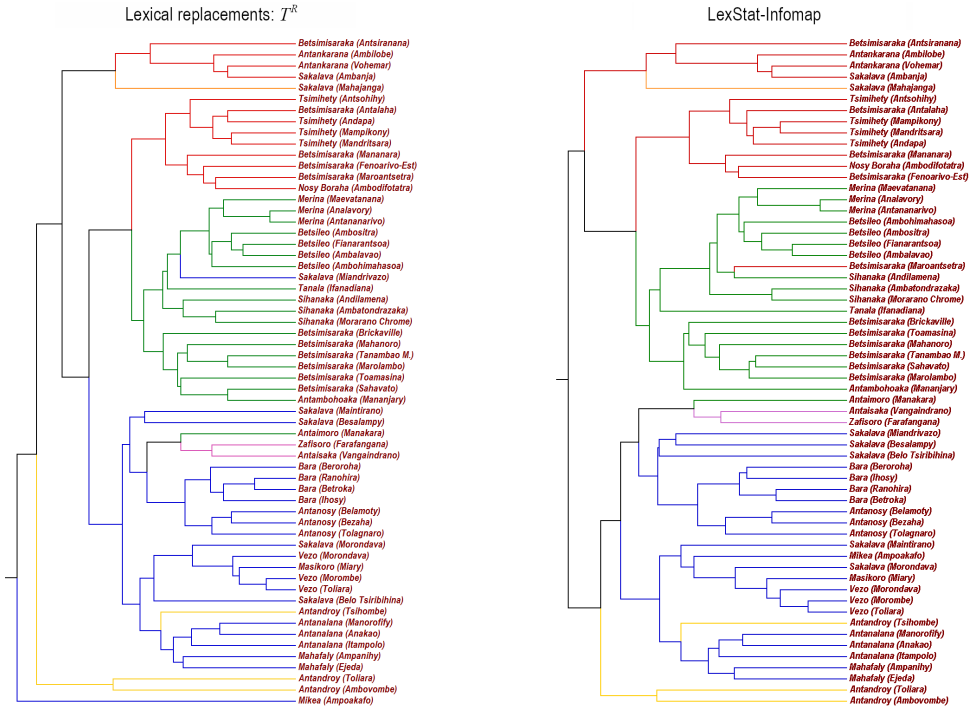

Figure 9
UPGMA trees of the Malagasy database reconstructed from genealogical distances T^{R}, inferred both with our algorithm (left, the same as the one in the upper left corner of Figure 8) and with the LexStat-Infomap method (right). The two cladograms are very similar, showing only minor imperfections. Both trees are less accurate than the two generated by NLD and MR distances (bottom panels of Figure 8).

are correct both in the T^{MR} and T^{NLD} trees. This observation can be quantified by means of generalized quartets distance (GQD) (Pompei, Loreto, and Tria 2011) between pairs of trees as follows (LSI = LexStat-Infomap): $GQD(T^{M}, LSI) = GQD(T^{MR}, LSI) = GQD(T^{NLD}, LSI) = 0.11$, and $GQD(T^{R}, LSI) = 0.45$.

7. Conclusions

The characterization of the stochastic process of gradual lexical modifications is the main innovation introduced in this article. We have seen that this random process is able to give significant information about a family of languages as, for example, providing a valid phylogenetic reconstruction as accurate as, at least, those *à la* Swadesh obtained with the classic tools of glottochronology. The reason for this good performance is simple to understand: Given the lexicon of two languages from the same family, it is easy to find pairs of very different words for the same concept (probably not cognate words), but a comparable or greater number of words that partially looks alike (probably cognate candidates). In other terms, gradual lexical modifications have usually larger statistics than lexical replacements and the increased accuracy in phylogenetic reconstruction is an obvious consequence.

It is worth noting that lexical modifications can be successfully described using NLD distance, a sensitive tool for small changes, while lexical replacements are better evaluated by a dichotomic distance 0/1. We have therefore shown how the right combination of these two metrics gives the appropriate distance for evaluating language similarity.

An appropriate tool for cognate detection is essential to distinguish the effects of lexical replacement, which modifies the lexicon of a language by separating words referring to the same concept into different subsets of cognates, from the effects of lexical modification, which continues to modify cognates within their subsets. We have thus introduced an automated procedure inspired by graph theory for this task, an extremely fast algorithm with a single easy-to-quantify parameter that returns cognate subsets very close to those of the LexStat-Infomap method in the case of the Malagasy dataset. The necessary condition for such good results is to have a very rich database, with a large representation of languages that belong to the family under investigation. At the moment this is still a limitation, but in the last few years the rapid diffusion of digital resources have led to a growing amount of collected data and we expect that good and large datasets for our approach will be more and more available in the near future.

The separation into inferred subsets of cognates shows a non-trivial linguistic structure for many concepts of the Malagasy dataset, well supported by a coherent geographical distribution of the corresponding localities. In other cases, the nature of a dialect family reveals itself in a single predominant subset. Actually, this leads us to believe that our automated cognate detection can be a very versatile tool and it could be used to find a new objective way to answer the old question of whether a pair of varieties are two separated languages or two dialects of the same language. In the first case, the abrupt replacements process is the principal explanation of the lexical differences, while in the second it should be the gradual modification process.

These promising results deserve to be tested and confirmed in a different, articulated context, such as a wide family of languages, much more open to external influences, with much more differentiation and where, additionally, loanwords play a non-negligible role. We think that such an analysis could be the right development for future research.

8. Supplementary Material

- Dataset Malagasy Swadesh lists version 1.1 - October 2021: The complete dataset of 207-item Swadesh lists for 60 Malagasy variants in text format. Version 1.0 - November 2019 has already been published (Serva and Pasquini 2020).
- A Python version of our code that requires the LingPy 2.6.9 package (`https://lingpy.org/`), complete with dataset Malagasy Swadesh list 1.1 in cldf format.

In addition, code and data are available for download via GitHub at `https://github.com/michelepasquini/LexMod_LexRepl`.

Acknowledgments

Michele Pasquini acknowledges the financial support from CNR, Istituto per le Applicazioni del Calcolo "Mauro Picone," Rome, with grant "Studio di linguistica quantitativa e lessicostatistica." Davide Vergni acknowledges the financial support from CNR project DIT.AD021.161.001 "Analisi probabilistica di dataset biologici e network dynamics."

References

Adelaar, K. Alexander. 1995. Borneo as a cross-roads for comparative Austronesian linguistics. In Peter Bellwood, James Fox, and Darrell Tryon, editors, *The Austronesians in History*. Australian National University, ANU E Press, pages 75–95.

Adelaar, K. Alexander. 2006. The Indonesian migrations to Madagascar: Making sense of the multidisciplinary evidence. In Truman Simanjuntak, Ingrid H. E. Pojoh, and Muhammad Hisyam, editors, *Austronesian Diaspora and the Ethnogenesis of People in Indonesian Archipelago*. Lipi Press, Jakarta, pages 205–232.

Adelaar, K. Alexander. 2012. Malagasy phonological history and Bantu influence. *Oceanic Linguistics*, 51:123–159. `https://doi.org/10.1353/ol.2012.0003`

Amigó, Enrique, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486. `https://doi.org/10.1007/s10791-008-9066-8`

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology*, 13:167–179. `https://doi.org/10.1515/LITY.2009.009`

Beaujard, Philippe. 2003. Les arrivées Austronésiennes à Madagascar: Vagues ou continuum? *Études Océan Indien*, 35–36:59–147.

Blench, Roger Marsh. 2007. New palaeozoogeographical evidence for the settlement of Madagascar. *Azania: Archaeological Research in Africa*, 42:69–82. `https://doi.org/10.1080/00672700709480451`

Blench, Roger Marsh. 2008. The Austronesians in Madagascar and their interaction with the Bantu of the East African Coast: Surveying the linguistic evidence for domestic and translocated animals. *Studies in Philippine Languages and Cultures*, 18:18–43.

Blench, Roger Marsh and Martin Walsh. 2009. Faunal names in Malagasy: Their etymologies and implications for the prehistory of the East African Coast. In *Eleventh International Conference on Austronesian Linguistics (11 ICAL)*, 31 pages.

Ciobanu, Alina Maria and Liviu P. Dinu. 2014a. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 99–105. `https://doi.org/10.3115/v1/P14-2017`

Ciobanu, Alina Maria and Liviu P. Dinu. 2014b. An etymological approach to cross-language orthographic similarity. Application on Romanian. In *Proceedings of the 2014 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 1047–1058. https://doi.org/10.3115/v1/D14-1112

Ciobanu, Alina Maria and Liviu P. Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pages 431–437. https://doi.org/10.3115/v1/P15-2071

Ciobanu, Alina Maria and Liviu P. Dinu. 2018. Simulating language evolution: A tool for historical linguistics. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, Association for Computational Linguistics*, pages 68–72.

Covington, Michael A. 1996. An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496.

Dahl, Otto Christian. 1938. Le systeme phonologique du proto-malgache. *Norsk Tidsskrift for Sprogvidenskap*, 10:189–235.

Dahl, Otto Christian. 1951. *Malgache et Maanjan: Une Comparaison Linguistique*. Egede Instituttet (Arne Gimnes Forlag), Olso.

Dahl, Otto Christian. 1954. Le substrat Bantou en Malgache. *Norsk Tidsskrift for Sprogvidenskap*, 17:325–362.

Dez, Jacques. 1963. Apersus pour une dialectologie de langue malgache. *Bulletin de Madagascar*, 204, 205, 206, 210.

D'Urville, Jules Dumont. 1832. Sur les îles du Grand Océan. *Bulletin de la Société de Góegraphie*, 17:1–21.

Dyen, Isidore, A. T. James, and J. W. L. Cole. 1967. Language divergence and estimated word retention rate. *Language*, 43:150–171. https://doi.org/10.2307/411390

Dyen, Isidore. 1953. Review of Otto Dahl, Malgache et Maanjan: Une comparaison linguistique. *Language*, 29(4):577–590. https://doi.org/10.2307/409983

Embleton, Sheila M. 1986. *Statistics in Historical Linguistics*, volume 30. Studienverlag Brockmeyer, Bochum.

Hauer, Bradley and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 865–873.

Hudson, Alfred B. 1967. *The Barito Isolects of Borneo: A Classification Based on Comparative Reconstruction and Lexicostatistics*. Cornell University, Ithaca, New York.

Jäger, Gerhard, Johann-Mattis List, and Pavel Sofroniev. 2017. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Long Papers)*, pages 1204–1215. https://doi.org/10.18653/v1/E17-1113

Le Cam, Lucien. 1986. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, Berlin.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

List, Johann-Mattis. 2012. Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources*, pages 117–125.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf University Press, Düsseldorf.

List, Johann-Mattis, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLoS ONE*, 12(1):e0170046. https://doi.org/10.1371/journal.pone.0170046, PubMed: 28129337

List, Johann-Mattis, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. *Proceedings of the Association for Computational Linguistics*, 2:599–605. https://doi.org/10.18653/v1/P16-2097

McMahon, April and Robert McMahon. 2005. *Language Classification by Numbers*. Oxford University Press.

Nerbonne, John and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of SIGPHON-97: 3rd Meeting of the ACL Special Interest Group in Computational Phonology*, pages 11–18.

Pasquini, Michele and Maurizio Serva. 2021. Stability of meanings versus rate of replacement of words: An experimental test. *Journal of Quantitative Linguistics*, 28:95–116. https://doi.org/10.1080/09296174.2019.1647754

Petroni, Filippo and Maurizio Serva. 2008. Languages distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*,

page P08012. https://doi.org/10.1088/1742-5468/2008/08/P08012
Petroni, Filippo and Maurizio Serva. 2010a. Lexical evolution rates derived from automated stability measures. *Journal of Statistical Mechanics: Theory and Experiment*, 2010:P03015. https://doi.org/10.1088/1742-5468/2010/03/P03015
Petroni, Filippo and Maurizio Serva. 2010b. Measures of lexical distance between languages. *Physica A*, 389:2280–2283. https://doi.org/10.1016/j.physa.2010.02.004
Petroni, Filippo and Maurizio Serva. 2011. Automated world stability and language phylogeny. *Journal of Quantitative Linguistics*, 18:53–62. https://doi.org/10.1080/09296174.2011.533589
Pompei, Simone, Vittorio Loreto, and Francesca Tria. 2011. On the accuracy of language trees. *PLoS ONE*, 6(6):e20109. https://doi.org/10.1371/journal.pone.0020109, PubMed: 21674034
Rama, Taraka and Johann-Mattis List. 2019. An automated framework for fast cognate detection and Bayesian phylogenetic inference in computational historical linguistics. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 6225–6235. https://doi.org/10.18653/v1/P19-1627
Rama, Taraka, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 393–400. https://doi.org/10.18653/v1/N18-2063
Serva, Maurizio. 2012. The settlement of Madagascar: What dialects and languages can tell us. *PLoS ONE*, 7(2):e30666. https://doi.org/10.1371/journal.pone.0030666, PubMed: 22363465
Serva, Maurizio and Michele Pasquini. 2020. Dialects of Madagascar. *PLoS ONE*, 15(10):e0240170. https://doi.org/10.1371/journal.pone.0240170, PubMed: 33007011
Serva, Maurizio and Michele Pasquini. 2022. Linguistic clues suggest that the Indonesian colonizers directly sailed to Madagascar. *Language Sciences*, 93:101497. https://doi.org/10.1016/j.langsci.2022.101497
Serva, Maurizio and Filippo Petroni. 2008. Indo-European languages tree by Levenshtein distance. *EuroPhysics Letters*, 81:68005. https://doi.org/10.1209/0295-5075/81/68005
Serva, Maurizio, Filippo Petroni, Dima Volchenkov, and Søren Wichmann. 2012. Malagasy dialects and the peopling of Madagascar. *Journal of the Royal Society Interface*, 9:54–67. https://doi.org/10.1098/rsif.2011.0228, PubMed: 21632612
Serva, Maurizio, Davide Vergni, Dima Volchenkov, and Aneglo Vulpiani. 2017. Recovering geography from a matrix of genetic distances. *Europhysics Letters*, 118:48003. https://doi.org/10.1209/0295-5075/118/48003
Starostin, Sergei A. 2000. Comparative-historical linguistics and lexicostatistics. In *Time Depth in Historical Linguistics, v. 1*. The McDonald Institute for Archaeological Research, Cambridge, pages 223–265. https://doi.org/10.1515/9781474473316-019
Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16:157–167. https://doi.org/10.1086/464084
Swadesh, Morris. 1951. Diffusional cumulation and archaic residue as historical explanations. *Southwestern Journal of Anthropology*, 7:1–21. https://doi.org/10.1086/soutjanth.7.1.3628647
Swadesh, Morris. 1952. Lexicostatistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96:452–463.
Swadesh, Morris. 1954. Perspectives and problems of Amerindian comparative linguistics. *Word*, 10:306–332. https://doi.org/10.1080/00437956.1954.11659530
Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121–137. https://doi.org/10.1086/464321
van der Merwe, Nikolaas J. 1966. New mathematics for glottochronology. *Current Anthropology*, 7:485–500. https://doi.org/10.1086/200754
Vérin, Pierre, Conrad P. Kottak, and Peter Gorlin. 1969. The glottochronology of Malagasy speech communities. *Oceanic Linguistics*, 8:26–83. https://doi.org/10.2307/3622902