On the Accuracy of Language Trees

Simone Pompei[1,2], Vittorio Loreto[1,3], Francesca Tria[1]*

[1] Complex Systems Lagrange Lab, Institute for Scientific Interchange (ISI), Torino, Italy, [2] Department of Physics, Università di Torino, Torino, Italy, [3] Department of Physics, Sapienza Università di Roma, Roma, Italy

Abstract

Historical linguistics aims at inferring the most likely language phylogenetic tree starting from information concerning the evolutionary relatedness of languages. The available information are typically lists of homologous (lexical, phonological, syntactic) features or characters for many different languages: a set of parallel corpora whose compilation represents a paramount achievement in linguistics.

From this perspective the reconstruction of language trees is an example of inverse problems: starting from present, incomplete and often noisy, information, one aims at inferring the most likely past evolutionary history. A fundamental issue in inverse problems is the evaluation of the inference made. A standard way of dealing with this question is to generate data with artificial models in order to have full access to the evolutionary process one is going to infer. This procedure presents an intrinsic limitation: when dealing with real data sets, one typically does not know which model of evolution is the most suitable for them. A possible way out is to compare algorithmic inference with expert classifications. This is the point of view we take here by conducting a thorough survey of the accuracy of reconstruction methods as compared with the Ethnologue expert classifications. We focus in particular on state-of-the-art *distance-based* methods for phylogeny reconstruction using worldwide linguistic databases.

In order to assess the accuracy of the inferred trees we introduce and characterize two generalizations of standard definitions of distances between trees. Based on these scores we quantify the relative performances of the distance-based algorithms considered. Further we quantify how the completeness and the coverage of the available databases affect the accuracy of the reconstruction. Finally we draw some conclusions about where the accuracy of the reconstructions in historical linguistics stands and about the leading directions to improve it.

Citation: Pompei S, Loreto V, Tria F (2011) On the Accuracy of Language Trees. PLoS ONE 6(6): e20109. doi:10.1371/journal.pone.0020109

Editor: Matjaz Perc, University of Maribor, Slovenia

Received March 23, 2011; **Accepted** April 11, 2011; **Published** June 3, 2011

Copyright: © 2011 Pompei et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: fra_trig@yahoo.it

Introduction

The last few years have seen a wave of computational approaches devoted to historical linguistics [1–3], mainly centred around phylogenetic methods. While the first aim of phylogeny reconstruction is that of classifying a set of species (viruses, biological species, languages, texts), the information embodied in the inferred trees goes beyond a simple classification knowledge. Statistical tools [4–9], for instance, permit to assign time weights to the edges of a phylogenetic tree, giving the opportunity to gather information about the past history of the whole evolutionary process. These techniques have been successfully employed to investigate features of human prehistory [10–15].

The application of computational tools in historical linguistics is not a novel one, since it dates back to the 50's, when Swadesh [16,17] first proposed an approach to comparative linguistics that involved the quantitative comparison of lexical cognates, an approach named *lexicostatistics*. The most important element here is the compilation, for each language being considered, of lists of universally used meanings (hand, mouth, sky, I, ..). The initial set of meanings included 200 items which were then reduced down to 100, including some new terms which were not in his original list. These famous 100-item Swadesh lists still represent the cornerstone of any attempts to reconstruct phylogenies in historical linguistics. Each language is represented by its specific list and different languages can be compared exploiting the similarity of their lists. The similarity is assessed by estimating the level of cognacy between pairs of words. The higher the proportion of cognacy the closer the languages are related. Though originally cognacy decisions was solely based on the work of trained and experienced linguists, automated methods have been progressively introduced (see [18] and for a recent overview [19]) that exploit the notion of *Edit Distance* (or *Levenshtein Distance*) [20] between words, considered as strings of characters. The computation of the Edit Distance between all the pairs of homologous words in pairs of languages leads to the computation of a "distance" between pairs of languages. This value is entered into a $N \times N$ table of distances, where N is the number of languages being compared. This distance matrix can then be submitted to *distance-based* algorithms for the purpose of generating trees showing relationships among languages.

The construction of the distance matrix is of course a crucial step since the reliability of the reconstruction of the evolutionary history, i.e., the outcome of a phylogenetic reconstruction method, strongly depends on the properties of the distance matrix. In particular if the matrix features the property of being *additive*, there are algorithms that guarantee the reconstruction of the unique true tree (see [21] for a recent overview). A distance matrix is said to be additive if it can be constructed as the sum of a tree's branches lengths. When considering experimental data, additivity is almost always violated. Violations of additivity can arise both from experimental noise and from properties of the evolutionary process the data come from. One of the possible sources of violation of additivity is the so-called back-mutation: in particularly long phylogenies a single character can experience multiple mutations. In this case the distances between taxa are no longer proportional to their evolutionary distances. In historical linguistics this would happen if one was considering meanings that change very rapidly. For this reason linguists are typically interested in removing from the lists all the fast-evolving meanings. Of course this is not an easy task, bringing inextricably with itself a fair amount of arbitrariness in the choice. Along the same lines another crucial difficulty in lexicostatistics concerns the rate of change of the individual meanings. Different meanings, represented in each language by different words, evolve with different rates of change. In a biological parallel one would say that the mutation rate, i.e., the rate over which specific words undergo morphological, phonetic or semantic changes, are meaning dependent. This effect again is not easily cured and again different choices of the list composition could lead to different reconstructions. Finally another source of deviations from additivity is the so-called horizontal-transfer. The reconstruction of a phylogeny from data underlies the assumption that information flows vertically from ancestors to offspring. However, in many processes information also flows horizontally. In historical linguistics borrowings represent a well-known confounding factor for a correct phylogenetic inference.

All the fore-mentioned difficulties in the reconstruction of phylogenetic trees strongly call for reliable methods to evaluate the reconstructed phylogenies. Along with this it comes the need of valid benchmarks for determining the reliability of the different methods used to reconstruct phylogenetic trees. The standard way of testing the proposed algorithms is the construction of models to generate artificial phylogenies [21–23], so that the algorithmic results can be directly compared with the true, known, observables of interest. However, in doing that, one makes inevitable assumptions on the evolutionary processes of interest, which can in turn influence the reconstruction performance. To overcome this problem, we consider here an application of phylogenetic tools to historical linguistics. This field offers a good reference point, since classifications made with phylogenetic tools can be compared with catalogues of languages made by experts. We focus in particular on the Ethnologue classification. The Ethnologue can be described as a comprehensive catalogue of the known languages spoken in the world [24], organized by continent and country, being thus a valid reference point to evaluate trees inferred using phylogenetic algorithms (see section *Data* for details).

Here we evaluate trees reconstructed using *distance-based* phylogenetic methods against the Ethnologue trees. To this end it is important to set the tools to compare expert Ethnologue trees and phylogenetically inferred trees. There are several standard ways of measuring the distance between two phylogenetic trees. Here we take in account two of them, the Robinson-Foulds (RF) distance [25], which counts the number of bipartitions on which the two trees differ, and the Quartet Distance (QD) [26], which counts the number of subset of four taxa on which the two trees differ.

A technical problem when comparing Ethnologue classifications and inferred trees is that typically Ethnologue trees are not binary while all the inferred trees are. In order to overcome this difficulty we introduce two incompatibilities scores, which are two generalizations of both the Robinson-Foulds [25] and the Quartet Distance measures [26]. We present results obtained on a wide range of language families. This allows to compare different definitions of distances as well as different reconstruction algorithms.

The outline of the paper is as follows. We first introduce the *Ethnologue* [24] project and both the *Automated Similarity Judgement Program* (*ASJP*) [27] and the *Austronesian Basic Vocabulary Database* (*ABVD*) [28] database we used in our analysis, pointing out some structural and statistical features that will be relevant in our discussion. Next we introduce some mathematical tools. We define both the *Levenshtein Normalized Distance* (*LDN*) and the *Levenshtein Divided Normalized Distance*(*LDND*) [19] to compute a "distance" between lists of word. The quantification of the accuracy of the inference of language trees we present is achieved with the Robinson-Foulds distance (RF) [25] and the Quartet Distance (QD) [26]. These are two standard definitions of distances between trees. We introduce and characterize such mathematical tools and we also present generalizations of these two scores, in order to adapt them for the comparison of binary (inferred) and non-binary (classifications) trees. We then present the results of the comparisons between the Ethnologue classifications and the language trees inferred based on the ASJP database. We first consider the ASJP database in order to perform a worldwide, i.e., large-scale, analysis. Finally we point out how some of the properties of word-lists, such as the completeness and the coverage, affect the accuracy of the reconstruction. To this end we present a comparative analysis on the inference of the Austronesian family, making use of both the ASJP and the ABVD database. File S1 provides an extensive account of the whole set of results we obtained.

Materials and Methods

Data

The **Ethnologue** can be described as a comprehensive catalogue of the known languages spoken in the world [24]. The Ethnologue was founded by R.S. Pittman in 1951 as a way to communicate with colleagues about language development projects. Its first edition was a ten-page informal list of 46 language and language group names. As of its sixteenth edition, Ethnologue has grown in a comprehensive database that is constantly being updated as new information arrives. As of now it contains close to 7000 language descriptions, organized by continent and country, which can be represented as a tree. As already mentioned, this tree is not always fully specified since it contains a lot of non-binary structures, in which the details of the phylogeny are not given due to a lack of certain information. Figure 1 illustrates geographically how the Ethnologue classifications deviate from being purely binary.

The **Automated Similarity Judgement Program** (**ASJP**) [27] includes 100-items word lists of about 50 families of languages throughout the world. These lists are written in a standardized orthography (ASJP code) which employs only symbols of the standard QWERTY keyboard, defining vowels, consonants and phonological features. The full database is available at http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm. Figure 2 (top) reports two statistical measures on the database to quantify its completeness. In particular we report the ranked fraction of languages containing a word for a specific meaning vs. the rank (left panel) and the ranked fraction of pairs of languages sharing a word (not necessarily a cognate) for a specific meaning vs. the rank (right panel). The second measure helps in understanding how accurate is, from a statistical point of view, computing the distance between two languages averaging the Levenshtein distances of all the words for homologous meanings. It is evident the extreme completeness of the database for lists up to 40 meanings.

The **Austronesian Basic Vocabulary Database** (**ABVD**) [28] contains lexical items from 737 languages (as of January 2011) spoken throughout the Pacific region. Most of these languages belong to the Austronesian language family, which is the largest family in the world. Due to the extended and phonetic characters

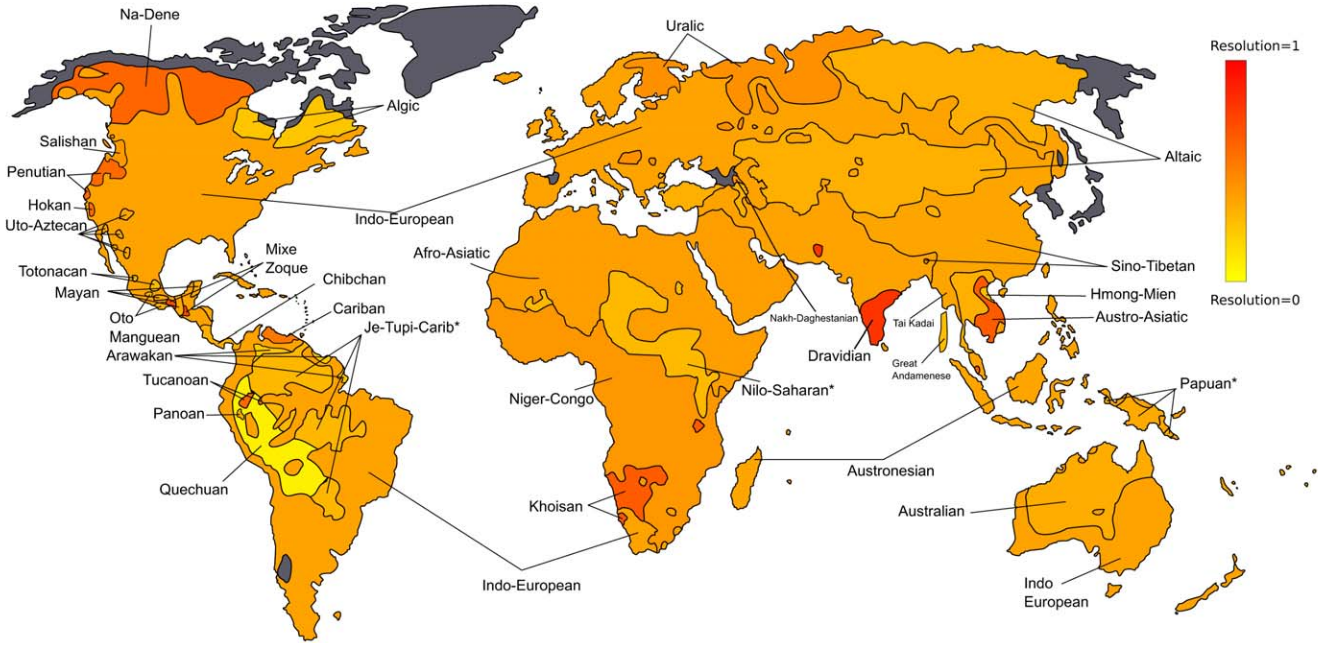

Figure 1. Ethnologue resolution power. This map represents the Ethnologue resolution power in the different world locations. Red areas corresponds to regions where the Ethnologue classification is completely binary, i.e., correspond to a tree in which each internal node has exactly two child nodes. Yellow areas corresponds to fully unspecified trees, featuring only a star structure. Grey areas are those for which no data are present in the databases we consider to reconstruct language trees. Asterisks are for regions which include more than one language family (we report in File S1 the list of such families).

doi:10.1371/journal.pone.0020109.g001

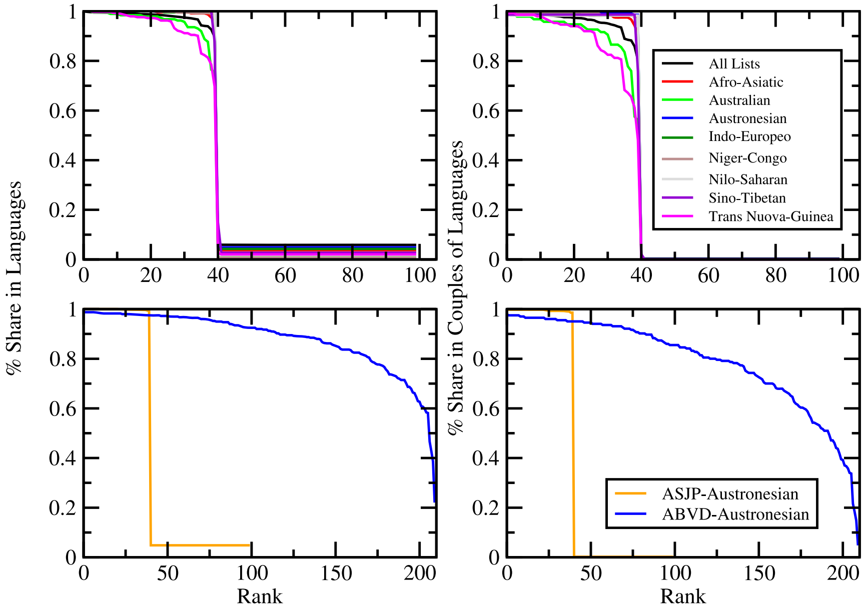

Figure 2. Top: Statistics of the ASJP database. (left panel) Fraction-rank plot: for each word in the lists of words of the Automated Similarity Judgement Project (ASJP), we measured the fraction of languages containing it. The plot reports this fraction vs. its rank. In the 100-items lists in the ASJP database, only 40 meanings are shared by almost 100% of the languages for each family. **(right panel)** Ranked fraction of pairs of languages sharing each specific word vs. rank. Again only 40 meanings are shared by almost 100% of the pairs of languages. Bottom: **Statistical measures on the ABVD database. (left panel)** Fraction-rank plot: for each word in the lists of words of the Austronesian Basic Vocabulary Database (ABVD), we measured the fraction of languages containing it. The plot reports this fraction vs. its rank. **(right panel)** Ranked fraction of pairs of languages sharing each specific word vs. rank. For sake of a rough comparison we also reported the same quantities measured on the Austronesian family of the ASJP database. The ASJP includes 40 words up to a maximum of almost 100% of the languages, whereas in the ABVD the percentage of coverage is at least of 50% for almost all the words in the list. Limited to the 40 most shared words the ASJP database features a slightly larger coverage than the ABVD database.

doi:10.1371/journal.pone.0020109.g002

used for the lexical orthography, all the information is encoded in the Unicode format UTF-8. The web site of the database is http://language.psy.auckland.ac.nz/austronesian/ and we downloaded it on October, the 4th 2010. We focused in particular on a subset of all the available languages composed by 305 languages that are present both in the ASJP database and in the Ethnologue classification. Figure 2 (bottom) reports the same quantities of Figure 2 (top) for the ABVD database. It is evident how, limited to the Austronesian family, the ABVD database features an overall larger (with respect to the ASJP database) number of meanings across all the languages considered. The level of coverage decreases progressively as one increases the number of meanings. A word of caution is in order. It is of course not possible to compare the completeness of the ASJP and the ABVD databases since they refer to two completely different projects with different aims: ASJP aiming at a full coverage of the Swadesh lists on all the world languages and ABVD being focused only on the Austronesian languages. It is nevertheless interesting to compare them only as for the Austronesian family is concerned. We shall come back on this point when we shall compare the accuracy of the reconstructed trees using different databases.

Distance between languages

In our studies we represent a language by its list of words for the different meanings. The distance between two languages is based on the distance between pairs of words corresponding to homologous meanings in the two lists. The distance between two words is computed by means of the Levenshtein distance (LD). The LD is a metric to quantify the difference between two sequences and it is defined as the minimum number of edit operations needed to transform one string into the other, the allowable edit operations being insertion of a character, deletion of a character and substitution of a single character.

Once specified the distance between pairs of words, two different definitions of distances between languages have been introduced [19,29–31]: the *Levenshtein Distance Normalized (LDN)* and a revised interpretation of it named *Levenshtein Distance Normalized Divided (LDND)*. Both these definitions have been introduced to correctly define distances between languages, instead of simply considering an average of the LD distance between words corresponding to homologous meanings in the lists.

According to LDN definition [29,30], given two words α_i and β_j, their distance is given by:

$$LDN(\alpha_i,\beta_j)=\frac{LD(\alpha_i,\beta_j)}{l(\alpha_i,\beta_j)} \tag{1}$$

where $LD(\alpha_i,\beta_j)$ is the LD between the two words and $l(\alpha_i,\beta_j)$ is the number of characters of the longest of the two words α_i and β_j. This normalization has been introduced in order to avoid biases due to long words, giving in this way the same weight to all the words in the lists. Starting from this definition, let us now assume that the number of languages is N and the list of meanings for each language contains M items. Each language in the group is labelled by a Greek letter (say α) and each word of that language by α_i, with $1\leq i\leq M$. Then, two words α_i and β_j in the languages α and β have the same meaning (they correspond to the same meaning) if $i=j$. The LDN between the two languages is thus:

$$LDN(\alpha,\beta)=\frac{1}{M}\sum_i LDN(\alpha_i,\beta_i) \tag{2}$$

Another definition of distance between pair of languages has been introduced in [31] in order to avoid biases due to accidental orthographical similarities in two languages. To this end a new normalization factor has been proposed [31] as follows:

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

The LDND distance between two languages is then defined as:

$$LDND(\alpha,\beta)=\frac{LDN(\alpha,\beta)}{\Gamma(\alpha,\beta)} \tag{4}$$

A comparison of the two definition of distances has been presented in [19]. In the following we consider both these definitions of distances between languages; the dissimilarity-matrices computed according to them will be the starting point for the inference of the family trees, which will be compared with the corresponding Ethnologue classifications.

Robinson-Foulds, Quartet Distance and generalizations

All the conclusions drawn in this work will be based on a quantitative comparison between inferred trees and the Ethnologue classifications. To this end it is important to recall how to measure the distance between two tree topologies. Here we recall in particular the mathematical definitions of two metrics between trees: the Robinson-Foulds distance (RF) [25] and the Quartet Distance (QD) [26].

The Robinson-Foulds (RF) distance between two trees counts the number of bipartitions on which the two trees differ. If we delete an internal edge in a tree, the leaves will be divided in two subsets; we call this division a bipartition. Here we consider a normalized version of the RF distance, which counts the percentage of unshared bipartitions between two trees. More formally, let $T1$ and $T2$ be two trees with the same set of leaves, then:

$$RF(T1,T2)=\frac{i(T1)+i(T2)-2e(T1,T2)}{i(T1)+i(T2)} \tag{5}$$

where $i(T)$ denotes the set of internal edge of T and $e(T1,T2)$ denotes the number of pairs of identical bipartitions in $T1$ and $T2$. The RF distance is a metric in the space of trees, whose value ranges from 0 (if and only if $T1=T2$) to 1.

Another possible distance between two trees is the Quartet Distance (QD). In a tree of N leaves, we can look at the subtrees defined by sets of four taxa (quartets). In the general case of non fully resolved trees, a *butterfly* names a quartet in which the two pairs of leaves are divided by an internal edge and a *star* a quartet in which the leaves are all linked to the same node. The QD between two trees counts the number of non compatible quartets in the two trees. It is defined as:

$$QD(T1,T2)=\frac{q(T1)+q(T2)-2s(T1,T2)-d(T1,T2)}{\mathrm{norm}(N)} \tag{6}$$

where $q(T)$ is the total number of butterflies in T, $s(T1,T2)$ is the number of identical butterflies in $T1$ and $T2$ and $d(T1,T2)$ is the number of different butterflies in the two trees. The normalization factor is the number, $\mathrm{norm}(N)=\binom{N}{4}$, of quartets in a tree of N

taxa. The QD, as well as the RF distance, is a metric in the space of trees, whose value ranges from 0 (if and only if $T1 = T2$) to 1.

In [32,33] a deep analysis of both RF and QD is reported, pointing out the different information the two measures convey. In limiting cases, pairs of trees that have the same RF distance but very different QD, and vice-versa, are also shown. In Fig. 3, quoting an enlightening example in [32,33], we show how the RF and the QD measures weigh a swapping event of two subtrees in a tree. In this case the RF distance is equal to the number of edges in the path between the swapped subtrees, while the QD is sensitive to the size of the subtrees. The RF is then a good measure if we are interested in measuring how far apart subtrees are moved in one tree with respect to another. When we are interested instead in the size of the displaced subtrees, the quartet distance is a more adequate measure.

The Ethnologue classification provides a coarse grained grouping of subsets of languages, often leading to trees that are not fully resolved, i.e., that are not binary. For that reason, it is important to correct the biases suffered by the RF and QD distances while comparing binary with non binary trees.

Figure 4 illustrates a situation when a binary tree (T_i) is compared with a non-binary one (T_e). Both the RF and the QD give a non zero distance between the two trees: some partitions of T_i are in fact not present in T_e. It is important to consider, however, that in the case we are considering (algorithmic inference versus Ethnologue classification) non-binary classification is simply due to a lack of information or details that would lead to a finer classification. We would like to be able to distinguish intrinsic contradictions between reconstructed binary trees and the Ethnologue classifications from errors due to the low level of resolution of the Ethnologue trees. It is with this aim in mind that we introduce a generalization of both the RF distance and the QD.

Let T_e be the Ethnologue (non necessarily binary) tree and T_i the inferred tree, then we define the Generalized Robinson-Foulds (GRF) score as:

$$GRF(T_i, T_e) = \frac{i(T_i) - e_{\mathrm{mod}}(T_i, T_e)}{i(T_i)} \qquad (7)$$

where $i(T_i)$ denotes the number of internal edge of T_i and $e_{\mathrm{mod}}(T_i, T_e)$ the number of bipartitions in T_i compatible with those in T_e. Intuitively, a bipartition in T_i is said to be compatible with a bipartition in T_e if it does not contradict any of the bipartitions

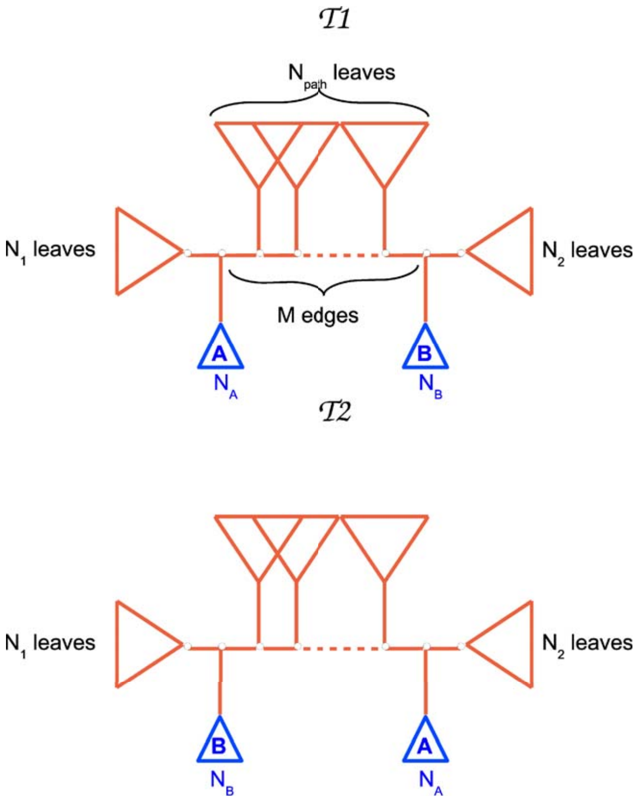

Figure 3. Robinson-Foulds and Quartet Distance: errors due to a displacement of a couple of subtrees. The trees T_1 and T_2 are different because of the swap of the subtrees **A** and **B**. While computing the distance between T_1 and T_2, the **Robinson-Foulds** distance detects all the M edges in the path as errors, regardless of the size of the subtrees attached to them. The number of wrong butterflies quartets counted as errors with the **Quartet Distance** is expressed by $N_1 N_A (N_{path} N_B + N_{path} N_2 + N_B N_2) + N_2 N_B (N_1 N_{path} + N_{path}) N_A$: the QD thus depends on the size of the subtrees.
doi:10.1371/journal.pone.0020109.g003

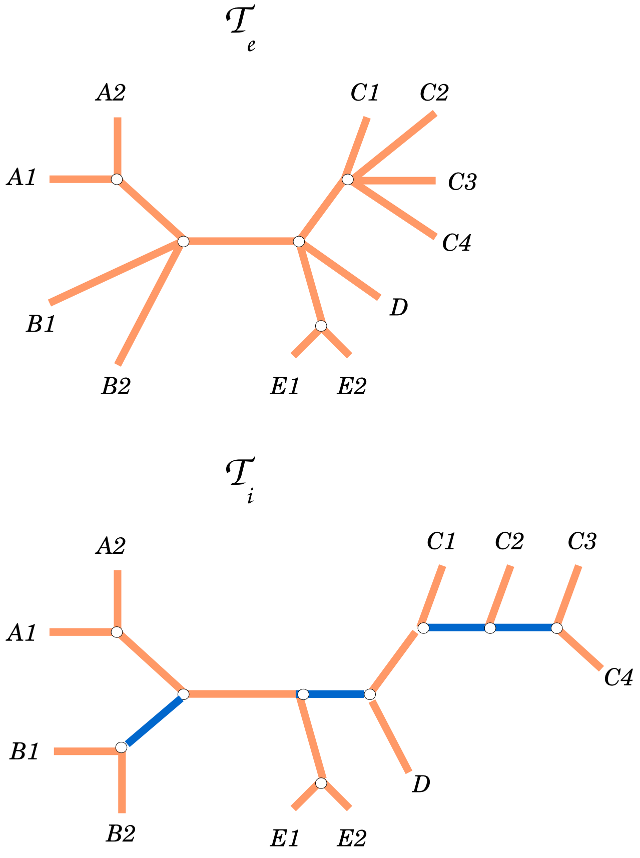

Figure 4. Non-binary nodes: biases of errors. The standard Robinson-Foulds distance and the Quartet Distance have a bias when comparing binary trees with non-binary classifications. The difference between tree T_e and T_i is that T_i shows a more fine grained classification. The two trees, however, are not conflicting, since T_i is simply a refinement of the classification T_e. The RF distance will count every internal edge (blue ones in T_i) of this refinement as errors, since they are not in T_e. The QD will count every quartet including the blue edges as errors, since all these quartets are stars in T_e. The generalized measures we introduce correctly give a null score between T_e and T_i in the example.
doi:10.1371/journal.pone.0020109.g004

induced by cutting an edge in T_e. More rigorously, the compatibility of a bipartition b of T_i with the tree T_e is defined as follows: let us call b_1 and b_2 the two sets defining b, and a_1^i,a_2^i the two sets defining the i-th bipartition of T_e. The partition b is compatible with the tree T_e if for each bipartition i of T_e, the following is true: $b_1 \subseteq a_1^i$, or $b_1 \subseteq a_2^i$, or $b_2 \subseteq a_1^i$, or $b_2 \subseteq a_2^i$. Let us note that the GRF is not symmetric in the two trees: this guarantees that a refinement edge is not counted as an error and the incomplete resolution of T_e does not affect the measure of the reliability of the reconstructed tree. We can verify that the GRF distance between T_i and T_e in figure 4 is zero.

The QD is more straightforwardly generalized. We introduce the Generalized Quartet Distance (GQD) as:

$$GQD(T_i,T_e)=\frac{d(T_i,T_e)}{\mathrm{norm}(T_e)} \tag{8}$$

where $d(Ti,Te)$, as already introduced, denotes the number of different butterflies in T_i and T_e. Again, this definition guarantees that all the star quartets in the Ethnologue trees will not be counted as errors. The normalization factor is equal to the number of butterfly quartets in T_e: $\mathrm{norm}(T_e)=q(T_e)$, recalling the definition of $q(T)$ given in eq. 6.

Let us stress again that both these generalized scores are neither symmetric or metric, since we are simply interested in quantifying the degree of accuracy of a binary tree with respect to an already known classification. With this definition, both the GQD and the GRF score give null scores if a classification tree is compared with one of its possible refinements, while one would get a score of 1 for inferred trees in total disagreement with the classification. In File S1 we report a measure of the correlation of the accuracy of the trees reconstruction with the Ethnologue resolution, as measured both with the standard measures and with the generalized ones, showing how the last ones correctly remove the biases due to the incomplete Ethnologue classification.

Results

Inferred trees vs. Ethnologue

In this section we present the results of the comparison between the Ethnologue classification and the language trees inferred by state-of-the-art distance based algorithms. We first consider the ASJP database in order to perform a worldwide, i.e., large-scale, analysis.

Starting from the word lists of the ASJP project, we first estimated the distance matrices among all the languages in each family. We used both the LDN (2) and the LDND (4) distances, so we had two classes of distance matrices as an input for distance-based algorithms. We use three distance-based algorithms: *Neighbour-Joining (NJ)* [34], *FastME* [35] (belonging to the class of Balanced Minimum Evolution (BME) algorithms) and *FastSBiX* [22,23], a recently introduced Stochastic Local Search algorithm. Each distance matrix was submitted as input to the three algorithms, which gives, for each language family, a total of six possible inferred trees.

To quantify the accuracy of the inferred trees, for each language family we computed the Generalized Robinson-Foulds score (GRF) and the Generalized Quartet Distance (GQD) of the inferred trees with the corresponding Ethnologue classifications. Tables 1 and 2 illustrate in an aggregate way the results obtained using the ASJP database. In particular we report, for each continent, the mean and the variance, across all the language families in that continent, of the values of the GRF and of the GQD between the inferred trees and the corresponding Ethnologue classifications, using both the LDN and the LDND distances. For each continent we considered all the language families present in the ASJP database.

As already mentioned, the GRF and the GQD are two complementary measures of the disagreement between the inferred tree and the expert classification. The GRF quantifies the percentage of wrong edges in the inferred trees, while the GQD counts how many quartets in the Ethnologue tree are different butterflies than in the reconstructed tree. In both cases the performance of the different algorithms always look very similar, though in almost all cases the noise reduction made by FastSBiX corresponds to a slightly better ability in reconstructing the correct phylogenies. FastSBiX features indeed the lowest average scores and, in many cases, the lowest variances. As for the distance matrix, our results show how better performances are obtained, on average, by using the LDND distance (4). The last column of the tables, named "RANDOM", shows the error one would have for a randomly reconstructed tree. This information is useful to correctly appreciate the algorithmic ability of inferring the correct phylogenetic relationships. While in fact we correct the distance measures in order to avoid biases due to non binary classification, it is evident that it is easier to be consistent with a very coarse grained classification than with a finer one. In order to take into account this observation, we can compare the errors made by the reconstruction algorithms with the errors a completely randomly constructed tree (with the same leaves) would feature. The RANDOM columns of tables 1 and 2 report averages over 10 realizations of the GRF and the GQD between a randomly reconstructed tree and the Ethnologue classification.

Figures 5 and 6 report the histograms of the accuracies obtained using the FastSBiX algorithm for each continent and worldwide: large fluctuations exist both within each continent and worldwide (The complete set of results for each language family and for all the accuracy scores is presented in File S1 in Tables S4, S5, S6 and S7).

We finally give a pictorial view of the accuracy of the reconstruction algorithm across the planet. Figure 7 illustrates the Generalized Quartet Distance for the different language families on the world map, normalized with the corresponding random value. More specifically, the color codes, for each family f, the following quantity:

$$X_f=2\frac{GQD(f)}{GQD_{random}(f)} \tag{9}$$

where $GQD_{random}(f)$ represents the mean value of the GQD obtained averaging over 10 randomly reconstructed trees with the same leaves (languages) of the family f. X_f quantifies the level of accuracy of the reconstruction with respect to a null model. The multiplicative factor 2 is included for the sake of better visualization: $X_f=1$ indicates a $GQD(f)$ equal or higher to half of the random tree distance $GQD_{random}(f)$.

Effect of the database completeness and coverage

In this section we consider how the length and the completeness of the lists of words affect the accuracy of the reconstruction. To this end, we restrict our analysis to the Austronesian family for which two different databases are available: the Automated Systematic Judgement Program (ASJP) and the Austronesian Basic Vocabulary Database (ABVD). The two databases mainly differ in two features: ASJP's lists includes at most 100 items for each language, while ABVD's lists includes up to 210 words. In both cases, not all the languages in the family express all the meanings. As we have already pointed out in fig. 2, while in the

Table 1. Accuracy of the reconstructions as measured with the Generalized Robinson-Foulds (GRF).

GENERALIZED ROBINSON-FOULDS SCORE							
	LDN			LDND			
	Neighbour-Joining	FastME	FastSBiX	Neighbour-Joining	FastME	FastSBiX	RANDOM
AFRICA							
Mean	0.2872	0.2845	0.2749	0.2859	0.2743	**0.2729**	0.7888
Variance	0.0327	**0.0322**	0.0329	0.0324	0.0323	0.0332	0.1945
EURASIA							
Mean	0.3152	0.3116	0.2999	0.3056	**0.2930**	0.2998	0.9063
Variance	0.0244	0.0238	0.0138	0.0200	0.0200	0.0108	0.0313
PACIFIC							
Mean	0.1228	0.1271	0.1092	0.1200	0.1178	**0.1083**	0.7282
Variance	**0.0173**	0.0182	0.0181	0.0174	0.0177	0.0177	0.1422
AMERICA							
Mean	0.3084	0.2885	**0.2797**	0.2972	0.3080	0.3023	0.8949
Variance	0.0673	0.0600	**0.0522**	0.0673	0.0726	0.0654	0.0525

For each continent we report the average and the variance of the GRF over all the languages spread on the continent. The different columns correspond to the two different ways of constructing the distance matrix (LDN and LDND) and to the three distance-based algorithms considered. The last column labelled RANDOM reports the results for the null model considered. See the main text for details.

doi:10.1371/journal.pone.0020109.t001

ASJP there are 40 words shared by all the languages and 60 words contained only in a small subset, in the ABVD database each word is shared at least by 50% of the languages in the family.

In order to get a fair comparison, we isolate a subset of 305 lists of words corresponding to languages shared by the two databases. The full list of languages is available in File S1. These two classes of lists are used to infer phylogenetic trees of the corresponding languages to be compared with the Ethnologue classifications. Since the results of the previous section did not show a significant difference between the two definitions of distance matrix, here we only use the *LDN* distance which allows for faster computations. Further, we only consider the *FastSBiX* algorithm to reconstruct phylogenies, being the one that features slightly better performances, as shown in the previous section.

We start by investigating the effect of the length of the word-lists on the accuracy of the inference of evolutionary relationships among languages. To this end, for each of the two databases, we proceed as follows: for each meaning i we compute the fraction f_i of languages which contains a word for i. We sort these values in a decreasing order, obtaining a ranked list of words. We then

Table 2. Accuracy of the reconstructions as measured with the Generalized Quartet Distance (GQD).

GENERALIZED QUARTET DISTANCE							
	LDN			LDND			
	Neighbour-Joining	FastME	FastSBiX	Neighbour-Joining	FastME	FastSBiX	RANDOM
AFRICA							
Mean	0.1379	0.1872	0.1379	0.1094	0.1048	**0.0855**	0.4781
Variance	0.0072	0.0164	0.0069	0.0047	0.0045	**0.0044**	0.0601
EURASIA							
Mean	0.1911	0.1787	0.1721	0.1716	0.1676	**0.1661**	0.6437
Variance	0.0378	0.0387	0.0399	0.0386	0.0385	0.0355	**0.0011**
PACIFIC							
Mean	0.0864	0.0901	**0.0662**	0.0829	0.0858	0.0706	0.4893
Variance	0.0096	0.0091	0.0085	0.0079	0.0109	**0.0070**	**0.0691**
AMERICA							
Mean	0.1595	**0.1536**	0.1569	0.1618	0.1646	0.1600	0.6057
Variance	0.0252	0.0245	**0.0235**	0.0244	0.0281	0.0269	0.0339

For each continent we report the average and the variance of the GQD over all the languages spread on the continent. The different columns correspond to the two different ways of constructing the distance matrix (LDN and LDND) and to the three distance-based algorithms considered. The last column labelled RANDOM reports the results for the null model considered. See the main text for details.

doi:10.1371/journal.pone.0020109.t002

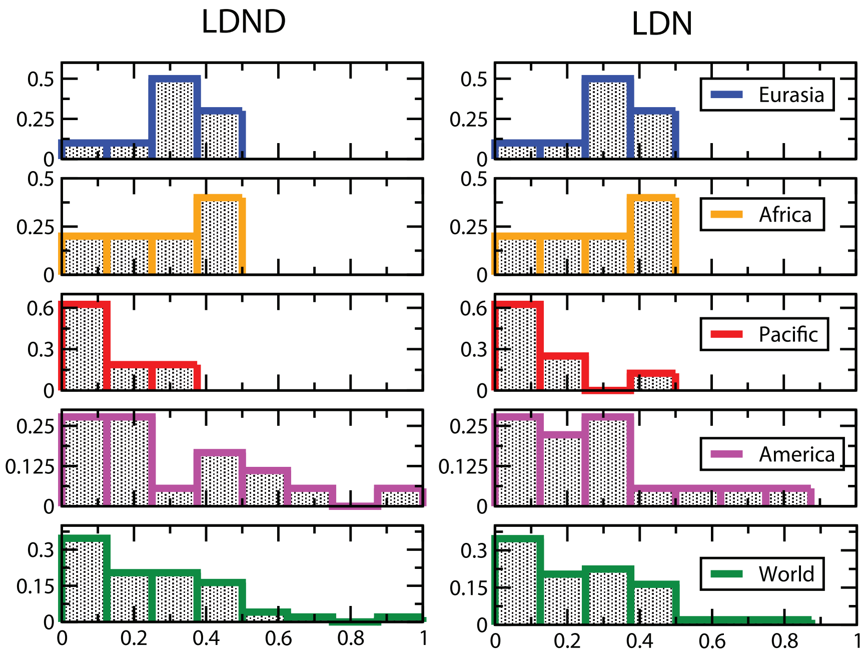

Figure 5. Accuracy histograms as measured with the Generalized Robinson-Foulds score (GRF). For each continent and for the whole world we report the histograms of the GRF as measured over all the families spread on each specific region. We considered here only the FastSBiX algorithm that features slightly better performances with respect to the competing algorithms, and both the the LDN (2) (right panel) and the LDND (4) (left panel) definition of distance. The histograms are always peaked near zero, meaning that the rate of errors are always very low, but the variances are quite large. These distributions do not discriminate the performances of the inference using LDN (2) or LDND (4) definition of distances.

doi:10.1371/journal.pone.0020109.g005

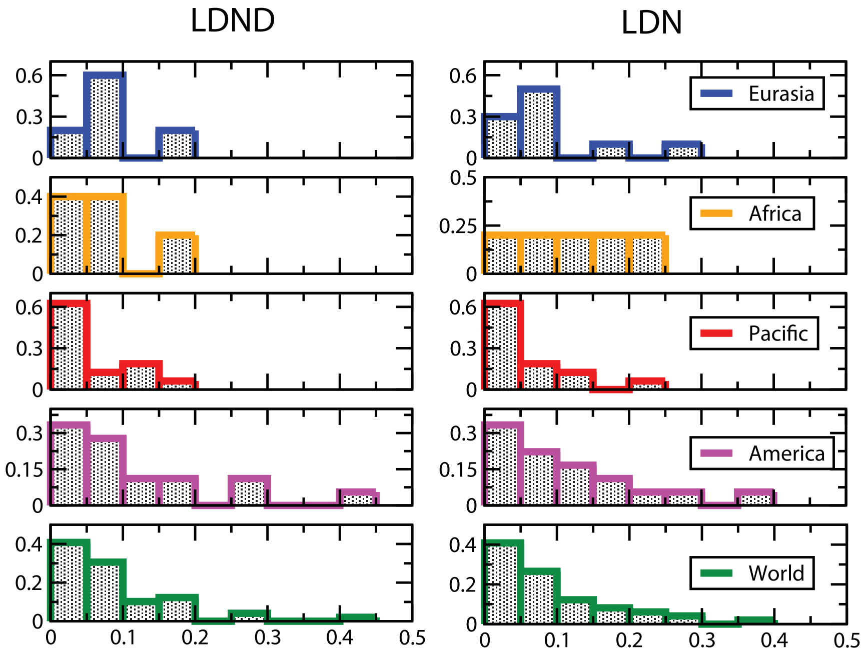

Figure 6. Accuracy histograms as measured with the Generalized Quartet Distance (GQD). For each continent and for the whole world we report the histograms of the GQD as measured over all the families spread on each specific region. We considered here only the FastSBiX algorithm that features slightly better performances with respect to the competing algorithms, both with the LDN (2) (right panel) and the LDND (4) (left panel) definition of distance. The histograms are always peaked near zero, meaning that the rate of errors are always very low. The distributions of the LDN-inferred trees, moreover, display larger variances than the LDND ones, this means that the latter definition allows for better performances in inferring languages trees with a distance-based approach. The overall variances are smaller with respect to the ones in fig. 5.

doi:10.1371/journal.pone.0020109.g006

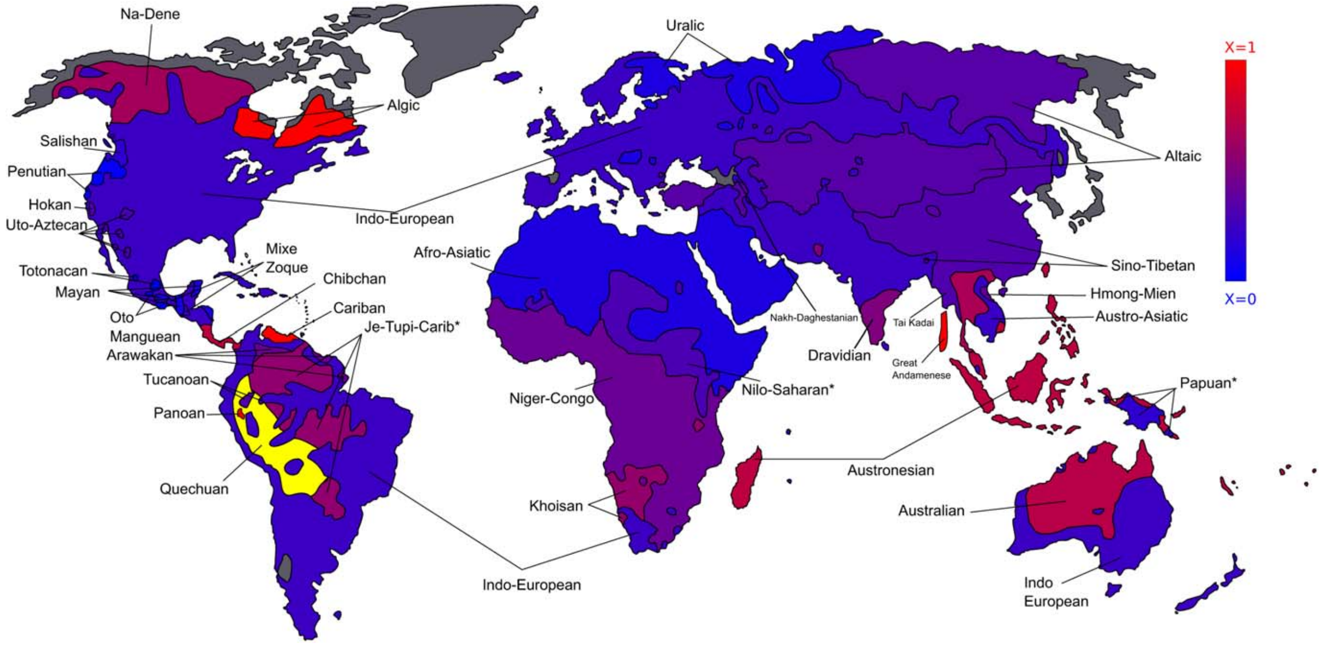

Figure 7. Worldwide accuracy of the inferred language trees. This map represents the level of accuracy of the FastSBiX algorithm on several language families throughout the world. The colors code the values of the Generalized Quartet Distance (GQD) between the trees inferred with the FastSBiX algorithm and the LDND definition of distance for each language family included in the ASJP database and the corresponding Ethnologue classifications. The GQD is normalized with the corresponding random value (see text for details). On the one hand blue regions corresponds to language families for which the inferred trees strongly agree with the Ethnologue classification. On the other hand red regions corresponds to poorly reconstructed language families. Yellow is for the families in which a random reconstruction would get a GQD score of zero, meaning that the Ethnologue classification has a null resolution (the corresponding tree is a star). Grey areas are those for which no data are present in the databases adopted for the reconstruction. Asterisks are for regions which include more than one family of languages. See File S1 for the analogous maps obtained with different algorithms and different definitions of the distance between languages.
doi:10.1371/journal.pone.0020109.g007

consider different word-lists, obtained in the following way: we start with the 10 most frequent words and we progressively add a constant number of words following the ranked list.

We compute the dissimilarity matrices by making use of only the reduced lists constructed as above, and we use those matrices as starting point for the reconstruction algorithm (we use the FastSBiX algorithm for all the results discussed below). Fig. 8 reports the Generalized Robinson-Foulds score (left) and the Generalized Quartet Distance (right) between the inferred trees and the corresponding Ethnologue classifications, as a function of the number M of chosen words, for both the AJSP and the ABVD databases. As a general trend, the number of errors decreases when the size of the word-lists considered increases. Though the large improvement of the accuracy occurs by adding the first 40 or 50 words, a slow improvement of the accuracy is always there if one keeps increasing the word-lists size. This already points in the direction that, in order to improve the accuracy of the phylogenetic reconstruction, one has to increase the size of the word-lists. The accuracy obtained with the ABVD and ASJP databases are very similar when considering the first $M=40$ most shared words. Upon increasing M, ASJP does not feature any improvement while ABVD keeps improving its accuracy, although very slowly, when $M>40$. A possible explanation for this could be related to the presence, in the ASJP database, of meanings with a very low level of sharing (see inset of the left panel of Fig. 8 as well as Fig. 2).

The value of M_{eff} (see inset of the left panel of Fig. 8) takes into account in how many languages a given meaning is expressed through a word. The missing information concerns whether pairs of languages have words for the same meaning. Suppose two languages have words for the same number of meanings. This does not mean that the meaning expressed by words in each language are the same. If paradoxically the sets of meanings covered by the two languages had a null overlap, we wouldn't have data to construct distance matrices. It is thus interesting to measure the degree of overlap between the list of words of pairs of languages. To this end, we define each language i as a binary vector $\vec{l_i}$ whose generic entry l_i^a is 1 if a word exists in that language for the meaning a and 0 otherwise. The overlap of two languages l_i and l_j is thus given by $\sum_a l_i^a l_j^a$. We define as level of coverage for a database the average overlap between all pairs of languages:

$$\text{Coverage} = \frac{2}{N(N-1)} \sum_{i \neq j} \sum_a l_i^a l_j^a, \qquad (10)$$

where N is the total number of languages considered, the index a runs over all the meanings while the indices i and j run over the different languages. In this way the maximal value of the coverage is given by the total number of meanings M we are considering. The inset of the right panel of Figure 8 reports the curves for the Coverage as a function of M. It is evident a strong correlation between M and the Coverage both in the ASJP and ABVD databases. Notice that the maximal observed values of the coverage are well below the theoretical maximum (100) in the ASJP database and below the maximum (210) in the ABVD database.

The above results can be summarized by saying that the accuracy of the reconstructions strongly depends on the completeness (quantified by M_{eff}) as well as on the level of Coverage of

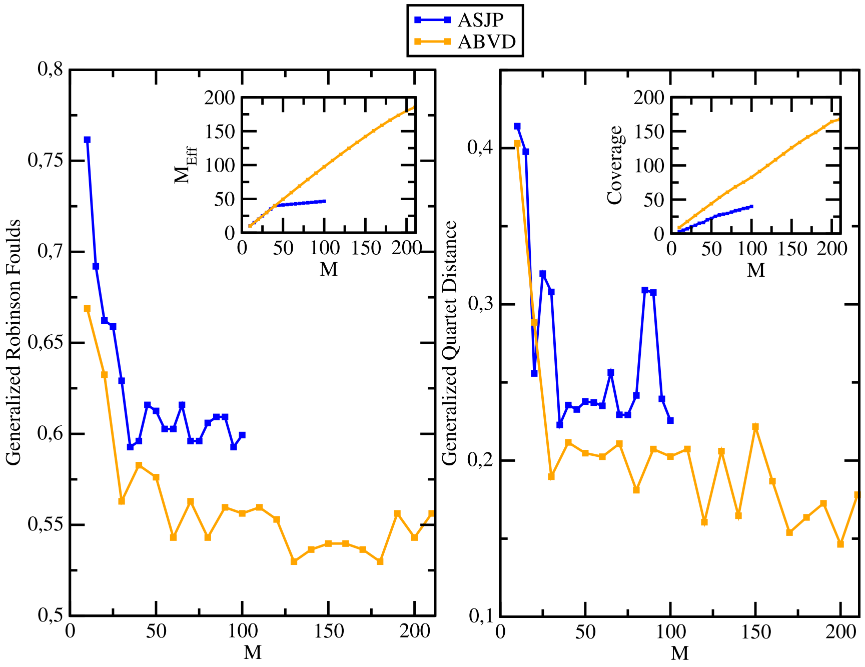

Figure 8. Role of the word-list completeness and coverage. (left) the Generalized Robinson-Foulds (GRF) score between the inferred trees and the corresponding Ethnologue classification for the Austronesian family, vs. the number M of most shared words, both for the ASJP and the ABVD databases. The inset reports the behaviour of M_{eff}, the effective number of most shared words, defines as follows. For each list M_{eff} is the sum of all the value of f_i for all the meanings in the list. In this way M_{eff} quantifies the effective number of most shared meanings. There is a strong correlation between M and M_{eff} for $M<40$. For $M>40$ M_{eff} does not increase anymore in the ASJP database. This explains why the GRF does not decrease for $M>40$ for the ASJP database. **(right)** the Generalized Quartet Distance (GQD) between the inferred trees and the corresponding Ethnologue classification for the Austronesian family, vs. the number M of most shared words, both for the ASJP and the ABVD databases. The inset reports the behaviour of the Coverage, which measures the degree of alignment of the word-lists for the different languages considered, vs. M (see text for details about the definition of Coverage). Again there is a strong correlation between the Coverage and M. The distance-based algorithm used is FastSBiX with the LDN definition of distance.

doi:10.1371/journal.pone.0020109.g008

the database considered. In the ASJP and ABVD databases M, M_{eff} and the Coverage are strongly correlated and one observes a first substantial improvement of the accuracy for $M<40$ and a continuous, though slower, improvement for $M>40$ in the ABVD database, where M_{eff} and the Coverage keeps increasing with M.

Discussion

In this work we presented a quantitative investigation of the accuracy of distance-based methods in recovering evolutionary relations between languages. The quantification of the accuracy rests upon the computation of suitable distances between the inferred trees and the classifications made by experts (in our case the Ethnologue).

We introduced two generalized scores, the Generalized Robinson-Foulds score (GRF) and the Generalized Quartet Distance (GQD), which successfully allow for the comparison of binary trees and expert classifications. The generalizations were made necessary in order to take into account the biases due to the presence of non-binary nodes in the Ethnologue classifications, which came from a non fine-grained groupings of the languages. Our scores do not count every refinement as an error, while properly take in account every displacement of a language or wrong groupings with respect to the classifications. These scores are generalizations of standard measures; on the one hand the RF, which is a good measure if we are interested in measuring how far displaced pairs of subtrees have been moved around in one tree compared to another; on the other hand the QD is a more adequate measure whenever it is important to quantify the size of displaced subtrees. Our generalized scores inherit all these properties. Moreover, while in the GRF the stress is on the inferred trees, counting the percentage of wrong bipartitions in the reconstructed tree, in the GQD the stress is on the classification, since we are computing the percentage of correctly inferred quartets in the reconstructed tree.

Once properly defined the tools for the comparison, we conducted a thorough evalution of the accuracy of distance based methods on all the language families listed in the ASJP database. The analysis was carried out by adopting state-of-the art distance-based algorithms as well as two different definitions of distance between lists of words, the LDN (2) and the LDND (4). In all the cases we obtained very robust results, which enabled us to draw some general conclusions. The two different definitions of distances between word-lists, LDN and LDND, almost guarantee the same accuracy for the inference of the trees of languages, with the LDND definition allowing for a slightly better accuracy (detailed results are reported in File S1). The LDN, on the other hand, because of its lower computational complexity, allows for faster computations without a considerable loss of accuracy. The length of the lists used to compute the distances between the languages strongly affects the accuracy of the reconstruction. The comparison between the two databases for the Austronesian family, the ASJP [27] and the ABVD [28] provides very important hints. The accuracy of the reconstruction always worsens if words with a low level of sharing are included; from this perspective it is always better to restrict the analysis to the meanings with an high Coverage instead of using all of them.

Fig. 7 summarizes the accuracy of distance-based reconstruction algorithms for the different language families on the world map. It is evident how at present the accuracy is satisfactory though highly heterogeneous across the different language families. Once removed the obvious bias due to the finite Ethnologue resolution power, this heterogeneity has to be presumably ascribed to a non homogeneous level of completeness and coverage of the word-lists for specific language families.

In conclusion we provided the first extensive account of the accuracy of distance-based phylogenetic algorithms applied to the recontruction of worldwide language trees. The overall analysis shows as the effort devoted so far to the compilation of large-scale linguistic databases [27,28] already allows for very good reconstructions. We hope our survey could be an important starting point for further progress in the field, especially for language families for which the available databases are still incomplete or the corresponding Ethnologue classification still poorly resolved.

Supporting Information

File S1
(PDF)

Acknowledgments

The authors wish to warmly thank Soeren Wichmann for having provided support for the use of the ASJP database as well as for very interesting discussions. At the same time the authors wish to thank Simon J. Greenhill for having granted the permission of using the ABVD database.

Author Contributions

Conceived and designed the experiments: SP VL FT. Performed the experiments: SP VL FT. Analyzed the data: SP VL FT. Contributed reagents/materials/analysis tools: SP VL FT. Wrote the paper: SP VL FT.

References

1. Renfrew C, McMahon A, Trask L, eds (2000) Time Depth in Historical Linguistics. The McDonald Institute for Archeological Research.
2. Joseph BD, Janda RD, eds (2004) The handbook of historical linguistics Blackwell Publishing.
3. Wichmann S, Grant AP, eds (2010) Quantitative Approaches to Linguistic Diversity, volume 27 of Special Issue of Diachronica Commemorating the centenary of the birth of Morris Swadesh. John Benjamins Publishing company.
4. Kishino H, Thorne JL, Bruno WJ (2001) Performance of a divergence time estimation methods under a probabilistic model of rate of evolution. Mol Biol Evol 18: 352–361.
5. Langley CH, Fitch WM (1974) An estimation of the constancy of the rate of molecular evolution. J Mol Evol 3: 161–177.
6. Rambaut A, Bromham L (1998) Estimating divergence data from molecular sequences. Mol Biol Evol 15: 442–448.
7. Sanderson MJ (2002) A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 19: 101–109.
8. Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol 19: 101–109.
9. Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of evolution. Mol Biol Evol 15: 1647–1657.
10. Gray RD, Atkinson Q (2003) Language-tree divergence times support the anatolian theory of indo-europian origin. Nature 426: 435–439.
11. Bryant D, Filimon F, Gray RD (2005) Untangling our past: Languages, trees, splits and networks. In: RMace C, SShennan, eds. The evolution of cultural diversity: phylogenetic approaches, UCL press. pp 67–84.
12. Atkinson PMQ, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout indo-european history. Nature 449: 717–720.
13. Atkinson Q, Meade A, Venditti C, Greenhill S, Pagel M (2008) Languages evolve in punctuational bursts. Science 319: 588.
14. Dunn M, Levinson SC, Lindström E, Reesink G, Terrill A (2008) Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. Language 84: 710–759.
15. Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in pacific settlement. Science 323: 479–483.
16. Swadesh M (1952) Lexico-statistic dating of prehistoric ethnic contacts. Proceedings of the National American Philosophical Society 96: 453–463.
17. Swadesh M (1955) Towards greater accuracy in lexicostatistic dating. International Journal of American Linguistics 21: 121–137.
18. Nerbonne J, Heeringa W, Kleiweg P (1999) Comparison and classification of dialects. In: Proceedings of the 9th Meeting of the European Chapter of the Association for Computational Linguistics. pp 281–282.
19. Wichmann S, Holman EW, Bakker D, Brown CH (2010) Evaluating linguistic distance measures. Physica A 389: 3632–3639.
20. Petroni F, Serva M (2010) Binary codes capable of correcting deletions, insertions, and reversals. Physica A 389: 2280–2283.
21. Pompei S, Caglioti E, Tria F, Loreto V (2010) Distance-based phylogenetic algorithms: new insights and applications. Mathematical Models and Methods in Applied Sciences (M3AS) 20: 1511–1532.
22. Tria F, Caglioti E, Loreto V, Pagnani A (2010) A stochastic local search algorithm for distancebased phylogeny reconstruction. Molecular Biology and Evolution 27: 2587–2595.
23. Tria F, Caglioti E, Loreto V, Pompei S (2010) A fast noise reduction driven distance-based phylogenetic algorithm. Proceedings of BIOCOMP2010 - The 2010 International Conference on Bioinformatics & Computational Biology.
24. Lewis M, ed (2009) Ethnologue: Languages of the World, Sixteenth edition. Dallas, Texas. SIL International. Online version: http://www.ethnologue.com/.
25. Robinson D, Foulds L (1981) Comparison of phylogenetic trees. Mathematical Biosciences 53: 131–147.
26. Bryant D, Tsang J, Kearney PE, Li M (2000) Computing the quartet distance between evolutionary trees. Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms. pp 258–286.
27. Holmann EW, Wichmann S, Brown CH, Velupillai V, Muller A, et al. (2008) Explorations in automated language classification. Folia Linguistica 42: 331–354.
28. Greenhill SJ, Blust R, Gray RD (2008) The austronesian basic vocabulary database: From bioinformatics to lexomics. Evolutionary Bioinformatics 4: 271–283.
29. Serva M, Petroni F (2008) Indo-european languages tree by levenshtein distance. Europhysics Letters 81: 68005.
30. Levenshtein VI (1966) Measures of lexical distance between languages. Soviet Physics Doklady 10: 707–710.
31. Bakker D, Mller A, Velupillai V, Wichmann S, Brown C, et al. (2009) Adding typology to lexicostatistics: a combined approach to language classification. Linguistic Typology 13: 167–179.
32. Christensen C, Mailund T, Pedersen CNS, Randers M (2005) Computing the quartet distance between trees of arbitrary degree. In: Proceedings of the 5th Workshop in Algorithms in Bioinformatics (WABI 2005). Springer, volume 3692 of Lecture Notes in Computer Science. pp 77–88.
33. Randers M (2006) Computing the Quartet Distance Between Trees of Arbitrary Degrees. Master's thesis, University of Aarhus.
34. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
35. Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. Journal of Computational Biology 9: 687–705.