BRILL

LANGUAGE DYNAMICS AND CHANGE 8 (2018) 1–21

brill.com/ldc

Making genealogical language classifications available for phylogenetic analysis

Newick trees, unified identifiers, and branch length

*Dan Dediu**
Max Planck Institute for Psycholinguistics, Nijmegen
dan.dediu@mpi.nl

Abstract

One of the best-known types of non-independence between languages is caused by genealogical relationships due to descent from a common ancestor. These can be represented by (more or less resolved and controversial) language family trees. In theory, one can argue that language families should be built through the strict application of the comparative method of historical linguistics, but in practice this is not always the case, and there are several proposed classifications of languages into language families, each with its own advantages and disadvantages. A major stumbling block shared by most of them is that they are relatively difficult to use with computational methods, and in particular with phylogenetics. This is due to their lack of standardization, coupled with the general non-availability of branch length information, which encapsulates the amount of evolution taking place on the family tree. In this paper I introduce a method (and its implementation in R) that converts the language classifications provided by four widely-used databases (Ethnologue, WALS, AUTOTYP and Glottolog) into

* Many thanks to the authors of the databases used here for making their data freely available, to Balthasar Bickel for computing the AUTOTYP distance and agreeing to making it freely available with this paper, to Luke Maurits for clarifying their "genetic method," to Michael Cysouw for making me aware of an alternative specification of cross-database language identifiers, to Harald Hammarström and Seán Roberts for discussions, to Michael Dunn for various bug reports, to Christian Bentz, Gerhard Jäger, and two anonymous reviewers for very thorough and helpful feedback on the manuscript. Thanks to the organizers of the NIAS-Lorentz workshop "Capturing Phylogenetic Algorithms for Linguistics," 26–30 October 2015, Leiden, the Netherlands, and to the editors of the proceedings in *Language Dynamics and Change*. This work is part of a project funded by the NWO (Netherlands Organisation for Scientific Research) VIDI grant number 016.124.315.

 | DOI: 10.1163/22105832-00801001

the *de facto* Newick standard generally used in phylogenetics, aligns the four most used conventions for unique identifiers of linguistic entities (ISO 639-3, WALS, AUTOTYP and Glottocode), and adds branch length information from a variety of sources (the tree's own topology, an externally given numeric constant, or a distance matrix). The R scripts, input data and resulting Newick trees are available under liberal open-source licenses in a GitHub repository (https://github.com/ddediu/lgfam-newick), to encourage and promote the use of phylogenetic methods to investigate linguistic diversity and its temporal dynamics.

Keywords

phylogenetics – language family – Newick – branch length

1 Introduction

Languages are not independent. This is due to historical processes such as language contact and descent from a common ancestor, and it is crucial to take into account the various types of non-independence between languages (e.g., Ladd et al., 2015; Roberts and Winters, 2013). Probably the most important type of non-independence is due to shared ancestry (Campbell and Poser, 2008): the daughter languages descended from a mother (or proto-)language share characteristics that they inherited from this common ancestor. This type of similarity tends to decrease with the passage of time since separation and is known as "Galton's problem" (this applies more generally to cultural phenomena; Mace and Pagel, 1994). Related languages descending from a shared ancestor form a *language family*, usually represented as a tree where the attested, present-day or recent languages form the *leaves* (or terminal nodes) and the extinct, mostly unattested, languages are *internal nodes*.

The identification of these *genetic relationships* is a complex problem (Campbell and Poser, 2008; Bowern and Evans, 2014) where controversies abound, including on the status of the so-called "macro-families" and on the composition and internal structure of language families such as Indo-European; disagreements concern the languages that belong to the same family, the internal relationships between them—the *tree topology*—and the amount of change that separates nodes in the tree—the *branch length*.

Using such language classifications with modern quantitative methods raises a number of major issues, including (i) the fact that there are several such classifications available, (ii) the fact that these are often presented in a

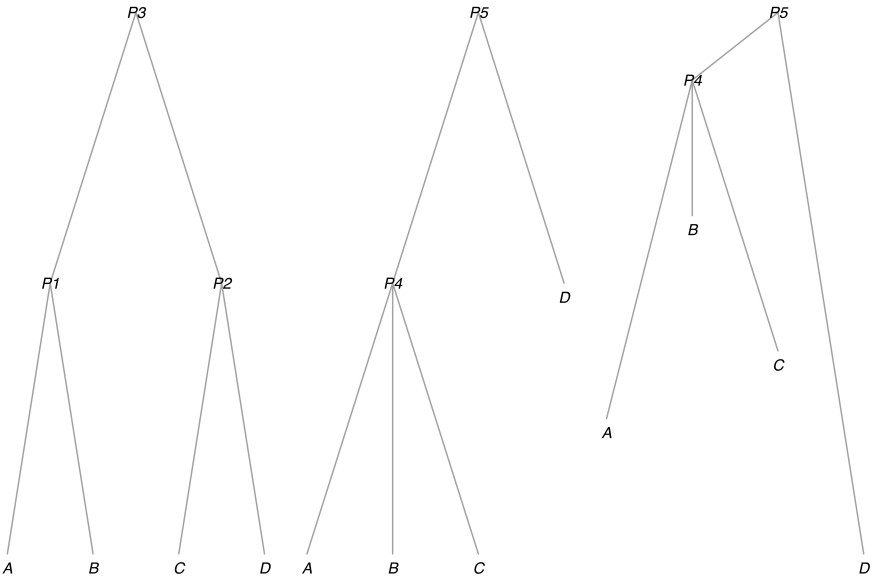

FIGURE 1 *Three language families composed of the same four languages* (A, B, C *and* D) *but with different structures* (*left vs. center*) *and branch length* (*center vs. right*). *Time flows downwards from the proto-language at the top* (P3, P5 *and* P5 *respectively*) *towards the attested languages at the bottom. For example, in the leftmost tree, languages* A *and* B *are more closely related than either of them is to language* C. *In the rightmost tree, language* B *has changed the least from its most recent common ancestor* (P4) *with languages* A *and* B.

non-standard format, and (iii) the problem that methods requiring not only the topology but also the amount of change (such as most modern phylogenetic approaches) cannot be directly applied because, in general, branch length information is lacking (and for good reason, as we generally do not know how to estimate it). The work presented here attempts to offer a solution to these issues by giving the *de facto* standard Newick tree format (http://evolution.genetics.washington.edu/phylip/newicktree.html) representation of language family trees from several classifications, and by adding branch length information estimated using several methods. For a few large families (such as Indo-European, Austronesian, Bantu and Uto-Aztecan), the application of Bayesian phylogenetic methods to basic vocabulary cognacy data and the use of calibration points with known dates resulted in the availability of posterior samples of trees with branch length (e.g. Bouckaert et al., 2012; Dunn et al., 2011), but there are still debates concerning these methods and their results, and the vast majority of language families did not yet receive this treatment, making an approach such as the one introduced here necessary. In this paper I describe the data, the methods and the format in which these trees are available, everything

being freely accessible on the GitHub repository https://github.com/ddediu/lgfam-newick, including the actual primary data (wherever allowed by their respective licenses), the R code (R Core Team, 2014), and the language family trees with branch length in Newick format.

2 Data and methods

2.1 *Primary data*

The main primary data is represented by the four most widely used language classifications. For each one I acquired the classification data in a format dependent on their export capabilities and converted it into Newick trees without branch length information, resulting in a set of language tree topologies. More precisely, each database needs a uniquely tailored approach because they use particular representations of the hierarchical relationships between languages, and my solution was to write a set of R (R Core Team, 2014) types and functions which extend R's own representation of phylogenetic trees (using the class `phylo` from package `ape`; Paradis et al., 2004) and allow the representation and manipulation of language family trees. Specifically, for each of the four language classifications, the data format and procedure were as follows:

- *Ethnologue* (Lewis et al., 2014), denoted in the following as **E**: the language classification data is not directly available for download, but instead the website provides[1] a webpage (http://www.ethnologue.com/browse/families) with the list of all the language families and hyperlinks to their own webpages, which were downloaded and automatically parsed to extract the tree structure of the family, the internal group names, the language names, and their ISO 639-3 codes.
- *World Atlas of Language Structures Online* (Dryer and Haspelmath, 2013) or *WALS*, denoted as **W**: the entire database (containing the language names, codes, geographic coordinates and the values for more than 130 structural features) is freely available for download at http://wals.info/static/download/wals-language.csv.zip under a Creative Commons license (CC BY-NC-ND 2.0 DE; http://creativecommons.org/licenses/by-nc-nd/2.0/de/deed.en), and I used only the fields containing the WALS, ISO 639-3 and Glottolog

1 As of February 2015, under a set of conditions given in the Terms of Use (www.ethnologue.com/terms-use) allowing "portions" of the data to be used for "research or educational purposes."

codes, the languages' names, their "genus" and "family," as this classification is flattened into a mostly three- (but sometimes four) level structure.
– *AUTOTYP* (Nichols et al., 2013), denoted **A**: the family trees are freely available for download at http://www.autotyp.uzh.ch/available.html, and can be used and distributed provided that their source is clearly mentioned; the language families are in a format similar to WALS, with each language being listed with its names, the AUTOTYP LID, the Glottolog and the ISO 639-3 codes, as well as the tree given as the "stock," "mbranch," "sbranch," "ssbranch" and "lsbranch" names, each being a hierarchical level (with the "stock" being the highest, the language family), sometimes with missing intermediate levels.
– *Glottolog* (Hammarström et al., 2014), denoted **G**: provides the family trees already in a standardized Newick format at http://glottolog.org/static/trees/tree-glottolog-newick.txt under a Creative Commons license (CC BY-SA 3.0; http://creativecommons.org/licenses/by-sa/3.0).

Please note that, while this paper concerns particular versions of these resources, I will try to keep the GitHub repository updated and compatible with newer versions and releases.

The methods for inferring branch length implemented here (see below) can use either the tree topology directly, a numeric constant, or a *distance matrix*. While the framework and my actual implementation in R can handle any distance matrix, for this paper I have used the following 11 distances (which fall into five types), respectively based on:

1. *vocabulary*: (1) ASJP16 distance,
2. *geography*: (2) great-circle distance,
3. *WALS*: (3) Gower distance and (4) Euclidean distance, without (3a and 4a) and with (3b and 4b) missing data imputation,
4. *AUTOTYP*: (5) Gower distance with missing data, using only the variables with a single datapoint per language (this distance was computed by Balthasar Bickel), and
5. the *tree topology*: the "genetic method" of Maurits and Griffiths (2014) applied to the WALS (6), Ethnologue (7), Glottolog (8) and AUTOTYP (9) classifications.

(1) represents the distances between languages as given by *The Automated Similarity Judgment Program* version 16 (ASJP16; Wichmann et al., 2013) and the ASJP software (version 2.1), freely available under a Creative Commons license (CC BY 3.0, http://creativecommons.org/licenses/by/3.0) from the authors' website (http://asjp.clld.org), computed as the normalized Levenshtein dis-

tances between standardized short wordlists transcribed with a reduced set of symbols (Bakker et al., 2009). After processing and conversion (manual replacement of some non-ASCII characters in the language descriptors and the 26-character language identifiers exported by ASJP v2.1), I exported these as a 3932 × 3932 distance matrix (with no missing data) between languages identified by their ISO 639-3 codes.

(2) is the geographic (great circle) distances between the languages, computed using R's function `distm()` (package `geosphere`; Hijmans, 2014), resulting in a 7494 × 7494 matrix with no missing data.

(3) and (4) represent distances between languages computed on the feature values in the WALS typological database, using R's function `daisy()` (package `cluster`; Maechler et al., 2015), either method `gower` (3; each feature is standardized between 0 and 1 by subtracting the feature's minimum and dividing by its range; Gower, 1971) or `euclidean` (4; standard Euclidean distance on the feature space). However, there is a lot of missing data in the WALS database (85.1% of the cells), so I have computed these distances using per variable mode data imputation, resulting in the following four 2679 × 2679 distance matrices: Gower (48.9% missing data; 3a), Gower with imputation (no missing data; 3b), Euclidean (48.9% missing data; 4a), and Euclidean with imputation (no missing data; 4b).

(5) is similar to (3) without missing data imputation but using the AUTOTYP typological database, resulting in a 2928 × 2928 distance matrix with 57.6% missing data.

Finally, (6) to (9) are distances between languages belonging to the same family, computed using the family tree topology as described in the "genetic method" of Maurits and Griffiths (2014):[2] languages with n shared intermediate nodes on their path to the root have a distance $d = M - \sum_{i=1}^{n} \alpha^i$ (where M is the maximum possible distance, and α is fixed at 0.69); I implemented it in R, and its application to each of the four classifications resulted in four distance matrices: MG2015 using WALS (2607 × 2607), Ethnologue (7492 × 7492), Glottolog (15772 × 15772), and AUTOTYP (2926 × 2926).

2.2 *Unique identifiers across classifications*

The question of allocating unique persistent identifiers to linguistic entities is essential, and several schemes are currently in wider use. Relevant here are:

2 Thanks to Luke Maurits for his help with clarifying the inner workings of the method; because these clarifications happened in an e-mail exchange during 2015, I denote this distance in the following as MG2015.

ISO 639-3 codes (three letters, denoted in the following as **i**; http://www-01.sil.org/iso639-3), *WALS codes* (three letters, **w**; http://wals.info), *AUTOTYP LIDs* (numeric, **a**; http://www.autotyp.uzh.ch), and *Glottocodes* (alphanumeric, four letters followed by four digits, **g**; http://glottolog.org/glottolog/glottologinformation). The mapping between these schemes is not yet standardized.[3] Here I devise a flexible scheme for uniquely mapping linguistic entities between these four systems. Some databases provide a mapping between their primary identifier and some others: Ethnologue (primary: **i**, secondary: none), WALS (primary: **w**, secondary: **i** and **g**), AUTOTYP (primary: **a**, secondary: **i** and **g**), and Glottolog (primary: **g**, secondary: **i**), allowing the reciprocal mapping (not always unique) between these four systems. With this, the coding scheme (or "UULID," the *Universally Unique Language IDentifier*) is standardized as 'NAME [i-I][w-W][a-A][g-G],' where the optional NAME represents a human-readable language name,[4] followed by a SPACE and the four unique codes I (ISO 639-3), W (WALS), A (AUTOTYP) and G (Glottocode), where any or all of these can be missing (the empty string "") or have multiple values separated by "-". A few examples (taken from the WALS classification of the Indo-European family) are: 'German Zurich [i-gsw][w-gzu][a-1305-1306-1307-1308-1309-1310][g-swis1247],' 'Urdu [i-urd][w-urd][a-2671][g-urdu1245],' 'Romani Sepecides [i-][w-rse][a-][g-],' and 'Germanic [i-][w-][a-][g-].' UULIDs and the mapping between the four coding schemes and linguistic entity names are freely available in the GitHub repository.

2.3 *Representing language classifications as Newick trees*

The *de facto* standard *Newick tree format*[5] is widely used in evolutionary biology, is almost universally imported and exported by software applications and libraries, and is very flexible, being able to accommodate rooted or unrooted trees, with our without leaf and internal node names, and with or without branch length information. In this format, subtrees are enclosed within parentheses "()" with the branch length optionally given as a number preceded by ":"

3 However, systems such as BCP47 (https://tools.ietf.org/html/bcp47) coupled with IANA (https://www.iana.org/assignments/lang-subtags-templates/lang-subtags-templates.xhtml) might provide a solution; thanks to Michael Cysouw for bringing this to my attention.

4 Because some ASCII characters have a special meaning in the Newick format, I have substituted them with others, as follows: "," → ".", "'" → "`", "(" → "{", ")" → "}", TAB → SPACE, ":" → "|", ";" → "|"; and characters lost their diacritics (e.g., "á" → "a" and "ã" → "a").

5 Described in http://evolution.genetics.washington.edu/phylip/newicktree.html and http://en.wikipedia.org/wiki/Newick_format.

immediately after the branch, and the description is terminated by a semicolon ";". For example, the leftmost tree in Fig. 2 (where languages are the leaves or terminal nodes, the proto-languages or groups are the internal nodes, and for simplicity all branches are taken to have the same length of 1) can be represented as:

- just the *topology* (structure): `((,),);`
- also showing leaf (terminal nodes) names: `((A,B),C);`
- also showing internal node (proto-languages, group) names (e.g., *P1* is the name of the last common ancestor of *A* and *B* and immediately follows the "()" enclosing its descendants): `((A,B)P1,C)P2;`
- leaf nodes and branch length (a number separated by ":" from the node name; e.g. the branch from the last common ancestor of *A* and *B* to *A* has length 1): `((A:1,B:1),C:1);`
- finally, everything (leaf and internal node names, and branch length): `((A:1,B:1)P1:1,C:2)P2;`

Here, the Newick trees representing language classifications use UUIDs (see Section 2.2) as the leaf and internal node names.

2.4 *Extracting the topologies of the language classifications*

The collection of the language classification from each of the four databases is met by specific challenges due to the particular representation of the genetic relationships between languages, and the actual format(s) in which these are available. I provide a set of `R` (R Core Team, 2014) classes and functions that extend the standard representation of phylogenetic trees as objects of class `phylo` (package `ape`; Paradis et al., 2004) and standardize the extraction of language family tree topologies, their conversion to the common Newick format, and its export to and import from file.

The extraction of standardized tree topologies from these diverse formats is based on the maintenance of a forest of partially built language family trees, which are updated by adding new full paths from the proto-languages to their daughter languages. More precisely, given a path between an internal node and a leaf (e.g., "Indo-European → Germanic → North-West Germanic → English"), the method attempts to identify an already existing partial tree that contains the origin of the path (here it would match an existing partial Indo-European tree) and to add the path to the tree, building the whole forest of all language family trees simultaneously from the ground up (see Fig. 2 for an example).

As a side note, a frequent issue that arises when using language classification data with `R`'s `ape` package is the mishandling of so-called "single nodes"

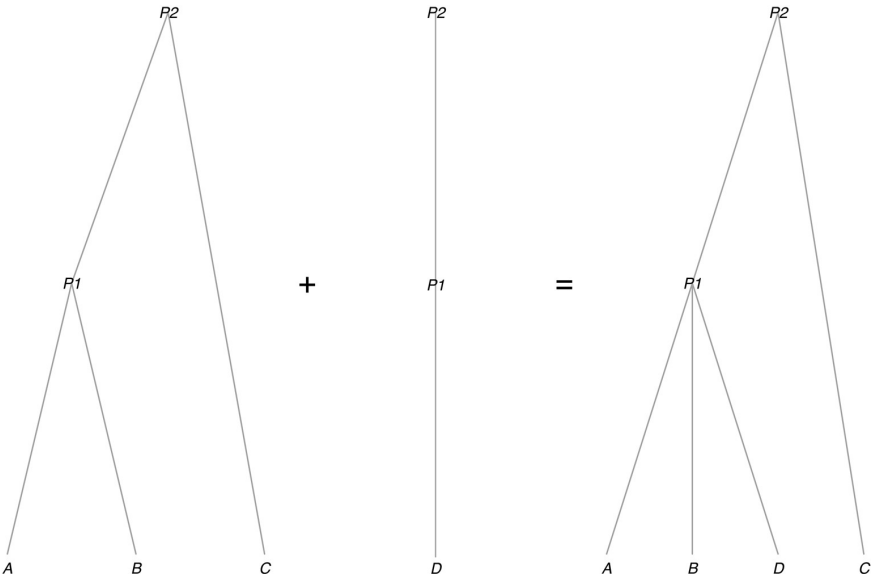

FIGURE 2 *Building the family trees by adding new paths, here adding the path $P2 \rightarrow P1 \rightarrow D$ to the left-hand tree, results in the right-hand tree because the algorithm recognizes that the path $P1 \rightarrow P2$ is already present in the partial left-hand tree.*

(i.e., internal nodes that have a single descendant in the tree, such as node $P1$ in the mid (degenerated) tree in Fig. 2). To address it, I have re-implemented `ape`'s `collapse.nodes()` function in such a way that it can correctly handle these cases.

2.5 *Adding branch length to language classifications*

In general, branch length represents the amount of evolutionary change that took place on a particular branch in the tree and must be interpreted in relative terms, unless the tree has been dated using absolute calibration points derived from other sources of data (Felsenstein, 2004; Bouckaert et al., 2012). I have added branch length information to the given tree topology T of a language classification using six methods falling into three broad classes:

1. methods that use the *tree topology* (and possibly a constant $k > 0$) to generate branch lengths: (1) constant, (2) proportional and (3) grafen,
2. a method that uses a distance matrix to generate the tree topology with branch lengths: (4) nj, and
3. methods that map a distance matrix onto the topology: (5) nnls and (6) ga.

Method (1) computes branch lengths such that for every path from the root to the leaves, the branch lengths add up to a given constant k (i.e., the same amount of evolution k has happened on all paths from the root to the terminal nodes) by defining the minimum branch length $brlen_{min} = k/$(number of levels in the tree) and allocating to each branch, starting from the root, a length of $brlen_{min}$, making sure that the terminal nodes have a total path length of k (thus, "telescoping" them if necessary to "accommodate" any remaining path length longer than $brlen_{min}$). For example, the leftmost tree in Fig. 2 with $k = 1.0$ becomes `((A:0.667,B:0.667)P1:0.333,C:1)P2;`.[6]

Method (2) simply forces each branch to the same length k (i.e., the amount of evolution is proportional to the number of splits on the path), resulting in `((A:1,B:1)P1:1,C:1)P2;`.

(3) reimplements Grafen's (1989) method, where each internal node is first given a "height" defined as the number of leaves in its subtree minus 1 (leaves get a "height" of 0), and the difference between the heights of the lower and the upper nodes defines the branch length: `((A:1,B:1)P1:1,C:2)P2;`.

Method (4) is the classic "Neighbor-Joining" (or NJ) method (Saitou and Nei, 1987), which iteratively joins taxa into higer groupings. Given a family tree T and a distance matrix D between (not necessarily all of) the languages in T, NJ (implemented by R's function `njs()` in package `ape`; Paradis et al., 2004) constructs the corresponding phylogenetic tree with branch lengths (thus the actual topology in T is discarded, as only its set of languages is used). For example, for the languages in Fig. 2, consider the distance matrix:

$$D = \begin{array}{r} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 0 & 2.1 & 3.9 \\ 2.1 & 0 & 4.2 \\ 3.9 & 4.2 & 0 \end{pmatrix} \end{array}$$

which approximates the distances between the three languages in the rightmost tree (assuming method (1) with $k = 2.0$), we obtain the NJ tree `(C:3,B:1.2,A:0.9);`—it is clear that NJ does not care about the structure (topology) of the original tree and might very well produce a very different topology.

6 Of course, any tree where `(A,P1)=(B,P1)` and `(A,P2)=(B,P2)=`k, such as `((A:0.5,B:0.5)P1:0.5,C:1)P2;`, would be equally good, but the particular implementation used here considers an extra (hidden) branch leading to the root when computing the number of levels in the tree (useful for other branch length methods and irrelevant here).

Methods (5) and (6) make use of both the given tree topology T and the distance matrix D by estimating branch lengths on T that approximate as closely as possible the original distances in D, in the sense that the distance matrix between the languages obtained by recording the minimum path lengths separating any two languages in the tree, D', is very similar to D. While these two methods have similar goals and produce very similar results, method (5) may be less robust than method (6) especially when the tree topology is complex, but method (6) is much slower, especially for very large trees, and might produce non-unique (but similar) solutions.

Specifically, method (5) estimates the branch lengths using the non-negative least squares (NNLS) approach implemented by R's function `nnls.tree()` (package `phangorn`; Schliep, 2011). The fundamental idea is that, for a given distance matrix D and a tree topology T with n branches, the method estimates the set of n branch lengths for T, b_1, b_2, ... b_n, such that the resulting *patristic distance matrix*[7] D' best approximates the original distances D in the sense of minimizing the sum of squared errors (SSE), using least squares.[8] Here it produces the tree `((A:1.05, B:1.05)P1:0.975, C:2.02)P2;`.

Method (6) is an original proposal that uses a standard genetic algorithm (R's function `ga()` in package `GA`; Scrucca, 2013) to estimate the branch length. With the notations above, I defined the "genome" as being composed of n real-valued "genes" $G = (g_1, g_2, \ldots, g_n)$ representing branch lengths, and the "fitness function" given by the SSE (sum of squared errors) between the original distances D and the current distances D', computed on topology T with the branch lengths $g_1, g_2, \ldots, g_n$: $f(T, D, G) = \frac{1}{n^2} \sum_{i,j=1}^{n} (d_{ij} - g_{ij})^2$. The genetic algorithm searches for the best solution $G^* = (g_1^*, g_2^*, \ldots, g_n^*)$ that minimizes the fitness function $f(T, D, G)$, using a population of 100 individuals for a maximum of 10,000 iterations (the search can stop earlier if the fitness didn't change for 100 consecutive iterations). A set of possible "best" trees could be:[9] `((A:0.9, B:1.2)P1:1.41, C:1.59)P2;`, `((A:0.9, B:1.2)P1:1.75, C:1.25)P2;`, or `((A:0.9, B:1.2)P1:1.73, C:1.27)P2;`.

7 The matrix of all pairwise distances between the terminal nodes in the tree, where the distance between a pair of nodes is the sum of the lengths of the branches connecting the two nodes in the tree.

8 See for example the blogpost http://blog.phytools.org/2011/03/for-fun-least-squares-phylogeny.html, where an initial version of the `nnls.tree()` function is introduced in one of the comments by Klaus Schliep.

9 Due to the randomness inherent to the genetic algorithm search process, coupled with the possibility of multiple optima, the "best" solution will most probably vary slightly between runs.

An important question concerns the *robustness* of these branch-length inference methods against violations of the conditions on D for being a true distance matrix:

a the diagonal is zero: $d_{ii} = 0$ for all $1 < i < n$,
b the off-diagonal is positive: $d_{ij} \geq 0$ for all $1 < i \neq j < n$,
c the matrix is symmetric: $d_{ij} = d_{ji}$ for all $1 < i, j < n$, and
d the triangle inequality is satisfied: $d_{ij} \leq d_{ik} + d_{kj}$ for all $1 < i, j, k < n$.

In principle, NJ and NNLS require a true distance matrix, but a closer examination of their implementation suggests that they might be robust against violations; by contrast, GA has no such requirements. To test this, I generated a set of four matrices (based on the test distance matrix used here, D) that violate each of the conditions in turn, as well as one matrix that violates all of them simultaneously. I then ran the methods NJ, NNLS and GA on them, observing that none of them crashed, nor did they fail to produce an output tree with branch length, and these trees do make sense given the violations.

Another question concerns the values of GA's parameters (population size N, maximum number of generations N_g, and number of generations without fitness improvement for premature stopping N_g^{const}), given that they affect the probability of finding the optimal solution(s), especially for complex trees, but also the computational costs of this search. Therefore, I compared the "standard" values $N = 10000$, $N_g = 100$, and $N_g^{const} = 100$ with a "thorough" set $N = 50000$, $N_g = 150$, and $N_g^{const} = 200$. When run on a compute cluster using a dedicated CPU per classification, the first required about 10 "wall clock" days to complete, while the second had to be stopped after 52 days (when all trees except one had converged). Despite this difference in computational costs, the computed branch lengths are very similar (across classifications and families: median Pearson's $r = 1.00$, median Euclidean distance $d = 0.02$), as are the optimal fitness values ($r = 0.93$, $p < 2.2 \cdot 10^{-16}$, paired t-test $t(3182) = 1.52$, $p = 0.13$). The number of generations required to stop is highly correlated ($r = 0.89$, $p < 2.2 \cdot 10^{-16}$) but significantly higher for the "thorough" condition (mean difference 1174.6, paired t-test $t(3182) = 20.14$, $p = 5.2 \cdot 10^{-85}$). Thus, I can conclude that the "normal" GA settings used here strike a good balance between computational efficiency and probability of converging to the optimal solution(s).

TABLE 1 *Summaries of the successfully harvested and exported language family tree topologies (i.e., no branch length information) per database; while the first two rows refer to the number of trees and leaf nodes in the whole database, the last five rows refer to the leaf nodes and levels per family tree.*

Summary	Ethnologue	WALS	AUTOTYP	Glottolog
Number of trees	147	214	403	435
Number of leaf nodes	7492	2607	2926	15772
Mean no. leaf nodes	51.0	12.2	7.3	36.3
Maximum no. leaf nodes	1545	371	340	3254
Minimum no. levels	3	4	3	3
Mean no. levels	4.8	4.0	3.4	4.5
Maximum no. levels	16	4	7	20

3 Results

Table 1 gives some summaries concerning the successfully harvested and exported language family trees available in the GitHub repository.

As detailed in the methods, for each of the four databases (Ethnologue, WALS, AUTOTYP and Glottolog) I applied each of the six methods of branch length estimation (constant, proportional, grafen, nj, nnls, ga). For the last three, there was a choice of 11 distance matrices (asjp16, great circle, wals (gower), wals (gower with imputation), wals (euclidean), wals (euclidean with imputation), autotyp (gower), Maurits and Griffiths (2014)'s "genetic method" mg2015 (on wals), mg2015 (on ethnologue), mg2015 (on glottolog), and mg2015 (autotyp); the last four being applied only to the corresponding database, i.e., there is no mg2015 (wals) applied to the Ethnologue trees), and each of these trees with branch length information was saved in the Newick format as part of Nexus (Maddison et al., 1997) files. Table 2 gives various summaries about these trees with branch length; please note that the number of trees differs between databases and that, within a database, the number of languages might differ by method due to the inherent missing data in the method's parameters and/or the incomplete overlap between the data in the method's parameters and the languages in the classification.

TABLE 2 *Summaries of the language family tree topologies (as listed in Table 1) with branch length information successfully added; an asterisk (*) indicates that missing data mode imputation was used.*

Classification	Method	Param./dist. mat.	No. trees	No. languages
Ethnologue	constant	$k = 1.0$	147	749
Ethnologue	proportional	k=1.0	147	749
Ethnologue	grafen		147	749
Ethnologue	nj	asjp16	147	3810
Ethnologue	nj	geo	147	7124
Ethnologue	nj	wals(gower)	147	1611
Ethnologue	nj	wals(euclid)	147	1611
Ethnologue	nj	wals(gower*)	147	2231
Ethnologue	nj	wals(euclid*)	147	2231
Ethnologue	nj	autotyp	147	365
Ethnologue	nj	mg2015(ethn)	147	7419
Ethnologue	nnls	asjp16	147	3846
Ethnologue	nnls	geo	147	7184
Ethnologue	nnls	wals(gower)	147	1017
Ethnologue	nnls	wals(euclid)	147	1017
Ethnologue	nnls	wals(gower*)	147	2273
Ethnologue	nnls	wals(euclidean*)	147	2273
Ethnologue	nnls	autotyp	147	858
Ethnologue	nnls	mg2015(ethn)	147	7479
Ethnologue	ga	asjp16	147	3846
Ethnologue	ga	geo	147	7184
Ethnologue	ga	wals(gower)	147	998
Ethnologue	ga	wals(euclid)	147	998
Ethnologue	ga	wals(gower*)	147	2273
Ethnologue	ga	wals(euclid*)	147	2273
Ethnologue	ga	autotyp	147	835
Ethnologue	ga	mg2015(ethn)	147	7479
WALS	constant	$k = 1.0$	214	2607
WALS	proportional	k=1.0	214	2607
WALS	grafen		214	2607
WALS	nj	asjp16	214	1973
WALS	nj	geo	214	2425
WALS	nj	wals(gower)	214	1807

Classification	Method	Param./dist. mat.	No. trees	No. languages
WALS	nj	wals(euclid)	214	1807
WALS	nj	wals(gower*)	214	2442
WALS	nj	wals(euclid*)	214	2442
WALS	nj	autotyp	214	462
WALS	nj	mg2015(wals)	214	2442
WALS	nnls	asjp16	214	2015
WALS	nnls	geo	214	2483
WALS	nnls	wals(gower)	214	1329
WALS	nnls	wals(euclid)	214	1329
WALS	nnls	wals(gower*)	214	2502
WALS	nnls	wals(euclid*)	214	2502
WALS	nnls	autotyp	214	884
WALS	nnls	mg2015(wals)	214	2502
WALS	ga	asjp16	214	1999
WALS	ga	geo	214	2481
WALS	ga	wals(gower)	214	1290
WALS	ga	wals(euclid)	214	1290
WALS	ga	wals(gower*)	214	2502
WALS	ga	wals(euclid*)	214	2502
WALS	ga	autotyp	214	832
WALS	ga	mg2015(wals)	214	2502
AUTOTYP	constant	$k = 1.0$	403	2926
AUTOTYP	proportional	k=1.0	403	2926
AUTOTYP	grafen		403	2926
AUTOTYP	nj	asjp16	403	2035
AUTOTYP	nj	geo	403	2547
AUTOTYP	nj	wals(gower)	403	1577
AUTOTYP	nj	wals(euclid)	403	1577
AUTOTYP	nj	wals(gower*)	403	2229
AUTOTYP	nj	wals(euclid*)	403	2229
AUTOTYP	nj	autotyp	403	559
AUTOTYP	nj	mg2015(autotyp)	403	2605
AUTOTYP	nnls	asjp16	403	2107
AUTOTYP	nnls	geo	403	2635
AUTOTYP	nnls	wals(gower)	403	1130
AUTOTYP	nnls	wals(euclid)	403	1130

TABLE 2 *Summaries of the language family tree topologies* (cont.)

Classification	Method	Param./dist. mat.	No. trees	No. languages
AUTOTYP	nnls	wals(gower*)	403	2319
AUTOTYP	nnls	wals(euclid*)	403	2319
AUTOTYP	nnls	autotyp	403	703
AUTOTYP	nnls	mg2015(autotyp)	403	2697
AUTOTYP	ga	asjp16	403	2091
AUTOTYP	ga	geo	403	2619
AUTOTYP	ga	wals(gower)	403	1065
AUTOTYP	ga	wals(euclid)	403	1065
AUTOTYP	ga	wals(gower*)	403	2299
AUTOTYP	ga	wals(euclid*)	403	2299
AUTOTYP	ga	autotyp	403	646
AUTOTYP	ga	mg2015(autotyp)	403	2697
Glottolog	constant	$k = 1.0$	435	15772
Glottolog	proportional	k=1.0	435	15772
Glottolog	grafen		435	15772
Glottolog	nj	asjp16	435	1926
Glottolog	nj	geo	435	4501
Glottolog	nj	wals(gower)	435	691
Glottolog	nj	wals(euclid)	435	676
Glottolog	nj	wals(gower*)	435	945
Glottolog	nj	wals(euclid*)	435	945
Glottolog	nj	autotyp	435	211
Glottolog	nj	mg2015(gott)	435	15507
Glottolog	nnls	asjp16	435	2000
Glottolog	nnls	geo	435	4605
Glottolog	nnls	wals(gower)	435	486
Glottolog	nnls	wals(euclid)	435	486
Glottolog	nnls	wals(gower*)	435	1019
Glottolog	nnls	wals(euclid*)	435	1019
Glottolog	nnls	autotyp	435	452
Glottolog	nnls	mg2015(glott)	435	15611
Glottolog	ga	asjp16	435	2000
Glottolog	ga	geo	435	4605
Glottolog	ga	wals(gower)	435	447
Glottolog	ga	wals(euclid)	435	447

Classification	Method	Param./dist. mat.	No. trees	No. languages
Glottolog	ga	wals(gower*)	435	1012
Glottolog	ga	wals(euclid*)	435	1012
Glottolog	ga	autotyp	435	412
Glottolog	ga	mg2015(glott)	435	15611

4 Discussion and conclusions

While I personally trust the Glottolog (Hammarström et al., 2014) classifications and use them primarily in my own latest work, the existence of alternative (and obviously not fully independent) classifications must be taken into account when dealing with "Galton's problem" (Mace and Pagel, 1994) by at least ensuring that the results are robust across classifications (e.g., Dediu, 2011; Dediu and Levinson, 2012). However, a prerequisite is the wide availability, free of charge, of such classifications in an as-standard-as-possible, machine-readable format that minimizes the amount of pre-processing prior to the actual computational and statistical analyses. An extra requirement, emerging from the recent explosion in the application of advanced phylogenetic methods to language, is that, besides the structure (topology) of the language family trees, one also needs branch length information encoding estimates of the amount of evolution that has taken place on the tree (Felsenstein, 2004). While reliable information of this type is very rarely available in linguistics (with the possible exception of the Bayesian posterior trees generated from basic vocabulary data for a few large families;[10] e.g., Dunn et al., 2011; Bouckaert et al., 2012; Gray et al., 2009), there are various methods for estimating it, based on the topology of the tree itself or on external information such as a distance matrix between pairs of languages. Because such estimates are, of course, highly contentious, one possibility is to conduct a *robustness analy-*

10 And even in these cases, it is unclear if the branch lengths derived from cognacy judgments on the basic vocabulary also encode evolutionary change transferable to other aspects of language, such as the wider lexicon or various classes of typological structures, or even beyond language to cultural features such as post-marital residence patterns (Jordan et al., 2009).

sis to test whether the results remain similar enough across language classifications (tree topologies) and estimates of the amount of evolution (branch lengths).

For many languages, data is simply not available or very restricted, often resulting in distance matrices with a very high proportion of missing data. While in some cases one may arguably use some form of data imputation, in others this is not warranted, as there are no good models available and/or the missing data patterns are non-random. However, even in such cases, a subset of families containing only a subset of languages may allow better inferences than the use of only a few large families of unknown representativeness.

The work presented in this paper is intended as an answer to these desiderata. It provides a collection of language families as phylogenetic trees in the *de facto* standard *Newick* format, free of charge and directly importable into the majority of modern phylogenetic software. Additional optional features are branch length information derived from a multitude of sources including the tree's own topology and inter-language distance matrices derived from various typological databases. Moreover, I also provide, under a liberal open source license, the actual R code (R Core Team, 2014) that loads, adds branch length information and exports the trees. This allows thus the user to consider new sources of information on the amount of evolution (e.g., from human genetic data, actual road distance, or linguistic typological databases) or new ways to map such external sources of information onto language family trees.

I hope that the method described here, the associated computer code, and the resulting language family trees will help promote quantitative approaches to problems in linguistic typology, language history and evolution, and even in the wider field of cultural evolution.

5 Supplementary material

The complete R code for extracting the language family tree topologies from these four databases, converting them to the Newick/Nexus format using the cross-database Universally Unique Language Identifiers (UULIDs), and for exporting and importing this format from file, as well as for computing the distance matrices described here, is freely available under a GPLv2 license (http://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html) in the GitHub repository https://github.com/ddediu/lgfam-newick, also containing the resulting language family trees with the various branch length estimates. This repository contains more information about the data, the various license terms, as well

as details about the results. Their use is encouraged, and bug reports and suggestions are welcome using the GitHub repository's ticketing mechanism or by directly e-mailing the author.

References

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(1). 10.1515/LITY.2009.009.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097): 957–960.

Bowern, Claire and Bethwyn Evans. 2014. *The Routledge Handbook of Historical Linguistics*. London: Routledge.

Campbell, Lyle and William J. Poser. 2008. *Language Classification: History and Method*. Cambridge: Cambridge University Press.

Dediu, Dan. 2011. A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone. *Proceedings of the Royal Society B* 278: 474–479. 10.1098/rspb.2010.1595.

Dediu, Dan and Stephen C. Levinson. 2012. Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS ONE* 7(9): e45,198. 10.1371/journal.pone.0045198.

Dryer, Matthew S. and Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at http://wals.info (accessed February 8, 2018).

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473: 79–82. 10.1038/nature09923.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates.

Gower, John C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27(4): 857–871. 10.2307/2528823.

Grafen, Alan. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society B* 326(1233): 119–157. 10.1098/rstb.1989.0106.

Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. *Science* 323(5913): 479–483. 10.1126/science.1166858.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Nordhoff

(eds.). 2014. *Glottolog 2.3*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available at http://glottolog.org (accessed February 8, 2018).

Hijmans, Robert J. 2014. geosphere: Spherical trigonometry. R package version 1.3–11. Downloadable at http://CRAN.R-project.org/package=geosphere.

Jordan, Fiona M., Russell D. Gray, Simon J. Greenhill, and Ruth Mace. 2009. Matrilocal residence is ancestral in Austronesian societies. *Proceedings of the Royal Society B* 276(1664): 1957–1964. 10.1098/rspb.2009.0088.

Ladd, D. Robert, Seán G. Roberts, and Dan Dediu. 2015. Correlational studies in typological and historical linguistics. *Annual Review of Linguistics* 1(1): 221–241. 10.1146/annurev-linguist-030514-124819.

Lewis, M. Paul, Gary F. Simons, and Charles D. Fennig (eds.). 2014. *Ethnologue: Languages of the World*. Dallas, TX: SIL International, 17th ed. Available at http://www.ethnologue.com (accessed February 8, 2018).

Mace, Ruth and Mark Pagel. 1994. The comparative method in anthropology. *Current Anthropology* 35: 549–564.

Maddison, David R., David L. Swofford, and Wayne P. Maddison. 1997. Nexus: An extensible file format for systematic information. *Systematic Biology* 46(4): 590–621. 10.1093/sysbio/46.4.590.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. 2015. cluster: Cluster analysis basics and extensions. R package version 2.0.1.

Maurits, Luke and Thomas L. Griffiths. 2014. Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences of the U.S.A.* 111(37): 13,576–13,581. 10.1073/pnas.1319042111.

Nichols, Johanna, Alena Witzlack-Makarevich, and Balthasar Bickel. 2013. The AUTOTYP genealogy and geography database: 2013 release. Accessible at http://www.autotyp.uzh.ch (accessed February 8, 2018).

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.

R Core Team. 2014. R: A language and environment for statistical computing. Downloadable at http://www.R-project.org (accessed February 8, 2018).

Roberts, Seàn and James Winters. 2013. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS ONE* 8(8): e70,902. 10.1371/journal.pone.0070902.

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4(4): 406–425.

Schliep, Klaus P. 2011. Phangorn: Phylogenetic analysis in R. *Bioinformatics* 27(4): 592–593.

Scrucca, Luca. 2013. GA: A package for genetic algorithms in R. *Journal of Statistical Software* 53(4): 1–37.

Wichmann, Søren, André Müller, Annkathrin Wett, Viveka Velupillai, Julia Bischoffberger, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Zarina Molochieva, Pamela Brown, Harald Hammarström, Oleg Belyaev, Johann-Mattis List, Dik Bakker, Dmitry Egorov, Matthias Urban, Robert Mailhammer, Agustina Carrizo, Matthew S. Dryer, Evgenia Korovina, David Beck, Helen Geyer, Pattie Epps, Anthony Grant, and Pilar Valenzuela. 2013. The ASJP database (version 16) Accessible at http://asjp.clld.org/ (accessed February 8, 2018).