SCIENTIFIC DATA

OPEN

Data Descriptor: Global-scale phylogenetic linguistic inference from lexical resources

Gerhard Jäger

Received: 21 February 2018
Accepted: 17 July 2018
Published: 9 October 2018

Automatic phylogenetic inference plays an increasingly important role in computational historical linguistics. Most pertinent work is currently based on *expert cognate judgments*. This limits the scope of this approach to a small number of well-studied language families. We used machine learning techniques to compile data suitable for phylogenetic inference from the ASJP database, a collection of almost 7,000 phonetically transcribed word lists over 40 concepts, covering two thirds of the extant world-wide linguistic diversity. First, we estimated *Pointwise Mutual Information* scores between sound classes using weighted sequence alignment and general-purpose optimization. From this we computed a dissimilarity matrix over all ASJP word lists. This matrix is suitable for *distance-based* phylogenetic inference. Second, we applied *cognate clustering* to the ASJP data, using supervised training of an SVM classifier on expert cognacy judgments. Third, we defined two types of binary *characters*, based on automatically inferred cognate classes and on sound-class occurrences. Several tests are reported demonstrating the suitability of these characters for *character-based* phylogenetic inference.

Design Type(s)	database creation objective • natural language processing objective
Measurement Type(s)	linguistic taxon
Technology Type(s)	machine learning technique
Factor Type(s)	Language • Language Family
Sample Characteristic(s)	

Tübingen University, Institute of Linguistics, Wilhelmstr. 19, 72074 Tübingen, Germany. Correspondence and requests for materials should be addressed to G.J. (email: gerhard.jaeger@uni-tuebingen.de)

Background & Summary

The cultural transmission of natural languages with its patterns of near-faithful replication from generation to generation, and the diversification resulting from population splits, are known to display striking similarities to biological evolution[1,2]. The mathematical tools to recover evolutionary history developed in computational biology — phylogenetic inference — play an increasingly important role in the study of the diversity and history of human languages[3–14].

The main bottleneck for this research program is the presently limited availability of suitable data. Most extant studies rely on manually curated collections of expert judgments pertaining to the cognacy of core vocabulary items or the grammatical classification of languages. Collecting such data is highly labor intensive. Therefore sizeable collections currently exist only for a relatively small number of well-studied language families[8,11,15–18].

Basing phylogenetic inference on expert judgments, especially judgments regarding the cognacy between words, also raises methodological concerns. The experts making those judgments are necessarily historical linguists with some prior information about the genetic relationships between the languages involved. In fact, it is virtually impossible to pass a judgment about cognacy without forming a hypothesis about such relations. In this way, data are enriched with prior assumptions of human experts in a way that is hard to control or to precisely replicate.

Modern machine learning techniques provide a way to greatly expand the empirical base of phylogenetic linguistics while avoiding the above-mentioned methodological problem.

The *Automated Similarity Judgment Program* (ASJP; see Data Citation 55) database contains 40-item core vocabulary lists from more than 7,000 languages and dialects across the globe, covering about 75% of the extant linguistic diversity. All data are in phonetic transcription with little additional annotations. (The only expert judgments contained in the ASJP data are rather unsystematic manual identifications of loan words. This information is ignored in the present study). It is, at the current time, the most comprehensive collection of word lists available.

Phylogenetic inference techniques comes in two flavors, *distance-based* and *character-based* methods. Distance-based methods require as input a matrix of pairwise distances between taxa. Character-based methods operate on a character matrix, i.e. a classification of the taxa under consideration according to a list of discrete, finite-valued characters. Character-based methods more directly infer the evolutionary process of descent with modification, but they can be computationally expensive. Distance-based methods return a tree diagram which groups taxa according to similarity, which does not necessarily equal relatedness (shared characteristics through common descent) but is highly computationally efficient and can be adequate for some purposes or useful as a first approximation.

The literature contains proposals to extract both pairwise distance matrices and character data from phonetically transcribed word lists[19–21]. In this paper we apply those methods to the ASJP data and make both a distance matrix and a character matrix for 6,892 languages and dialects—these are all languages in ASJP v. 17 except reconstructed, artificial, pidgin and creole languages—derived this way available to the community. Also, we demonstrate the suitability of the results for phylogenetic inference. The results of phylogenetic inference, i.e., fully bifurcating phylogenetic trees including branch lengths, for 66 language families and for the entire set of 6,892 of languages and dialects are also made publicly available. While these trees still await a detailed qualitative assessment by trained comparative linguists, they are (by construction) compatible with the Glottolog (see Data Citation 2) classification. This provides a useful resource for applications of the *Phylogenetic Comparative Methods* (see, e.g.[22], for an overview) to questions in linguistic typology.

While both the raw data and the algorithmic methods used in this study are freely publicly available, the computational effort required was considerable (about ten days computing time on a 160-cores parallel server). Therefore the resulting resource is worth publishing in its own right.

Methods

Creating a distance matrix from word lists

In[19] a method is developed to estimate the dissimilarity between two ASJP word lists. The main steps will be briefly recapitulated here.

Pointwise Mutual Information. ASJP entries are transcribed in a simple phonetic alphabet consisting of 41 sound classes and diacritics. (See[6] for a detailed description of the code). In all steps described in this paper, diacritics are removed. For instance, a sequence `tʰ`, indicating an aspirated "t", is replaced by a simple `t`. This way, each word is represented as a sequence over the 41 ASJP sound classes.

The *pointwise mutual information* (PMI) between two sound classes is central for most methods used in this paper. It is defined as

$$PMI(a,b) \doteq -\log \frac{s(a,b)}{q(a)q(b)}, \quad (1)$$

where $s(a, b)$ is the probability of an occurrence of a to be cognate with b in a pair of cognate words, and $q(x)$ are the probabilities of occurrence of x in an arbitrarily chosen word.

Let "-" be the gap symbol. A pairwise alignment between two strings (x,y) is a pair of strings (x',y') over sound class symbols and gaps of equal length such that x is the result of removing all gap occurrences in x',

and likewise for y′. A licit alignment is one where a gap in one string is never followed by a gap in the other string. There are two parameters gp_1 and gp_2, the *gap penalties* for opening and extending a gap. The aggregate PMI of an alignment is

$$PMI(x_i', y_i') = \sum_i PMI(x_i', y_i'), \quad (2)$$

where $PMI(x_i', y_i')$ is the corresponding gap penalty if x_i' or y_i' is a gap.

For a given pair of ungapped strings (x, y), $PMI(x, y)$ is the maximal aggregate PMI of all possible licit alignments between x and y. It can efficiently be computed with a version of the Needleman-Wunsch algorithm[23]. In this study, we used the function pairwise2.align.globalds of the *Biopython* library[24] for performing alignments and computing PMI scores between strings.

Parameter estimation

The probabilities of occurrence $q(a)$ for sound classes a are estimated as relative frequencies of occurrence within the ASJP entries. The scores $PMI(a,b)$ for pairs of sound classes (a,b) and the gap penalties are estimated via an iterative procedure.

In a first step, pairwise distances between languages are computed via the method described in the next subsection, using $1\text{-}LDN(x, y)$ instead of $PMI(x, y)$ as measure of string similarity, where $LDN(x, y)$ is the *normalized Levenshtein distance*[25] between x and y, i.e. the edit distance between x and y divided by the length of the longest string. In the next subsection I will describe a method how distances between strings can be aggregated to yield a distance measures between languages (i.e., word lists). (For the sake of readability, I will use the term "language" to refer to languages proper and to dialects alike; "doculect" would be a more correct if cumbersome term. A *doculect* is any linguistic variety (language, dialect, sociolect etc.) that "has been described or otherwise documented in a coherent way" (http://www.glottopedia.org/index.php/Doculect, accessed on June 12, 2018.) Working with doculects allows us to remain neutral about the notoriously difficult language/dialect distinction.). All pairs of languages $(l_1 l_2)$ where this distance ≤ 0.7 are considered as *probably related*. This threshold was chosen somewhat arbitrarily but is highly conservative; 99.9% of all probably related languages belong to the same language family and about 60% to the same sub-family.

Next, for each pair of probably related languages (l_1, l_2) and each concept c, each word for c from l_1 is aligned to each word for c from l_2. The pair of words with the lowest LDN score is considered as *potentially cognate*.

All pairs of potentially cognate words are aligned using the Levenshtein algorithm, and for each pair of sound classes (a, b), $s_0(a, b)$ is estimated as the relative frequency of a being aligned to b across all such alignments. Alignments to gaps are excluded from this computation. $PMI_0(a, b)$ is then calculated according to (1). As pointed out by a reviewer, this procedure gives more weight to large families, as there are much more pairs of probably related languages from large families than from small ones. Under the quite plausible prior assumption, however, that the probability of sound changes is not lineage-dependent, this does not lead to biased estimates though.

Suppose gap penalties gp_1, gp_2 and a threshold parameter θ are given. The final PMI scores are estimated using an iterative procedure inspired by the *Expectation Maximization* algorithm[26]:

- For i in 1…10:

1. All potential cognate pairs are aligned using the PMI_{i-1}-scores.
2. $s_i(a, b)$ is estimated as the relative frequency of a aligned with b among all alignments between potential cognates x,y with $PMI_{i-1}(x, y) \geq \theta$.
3. PMI_i is calculated using formula (1).

The *target function* $f(gp_1, gp_2, \theta)$ is the average distance between all probably related languages using the PMI_{10}-scores. The values for gp_1, gp_2, θ are determined as those minimizing f, using Nelder-Mead optimization[27]. The following optimal values were found: $gp_1 \approx -2.330, gp_2 \approx -1.276, \theta \approx 4.401$.

The threshold $\theta \approx 4.401$ ensures that only highly similar word pairs are used for estimating PMI scores. For instance, between French and Italian only five word pairs have a PMI similarity $\geq \theta$ according to the final scores: *soleil* [sole] - *sole* [sole] ('sun'; PMI = 11.6), *corne* [korn] - *corno* [korno] ('horn'; PMI = 7.7), *arbre* [arbr3] - *albero* [albero] ('tree'; PMI = 7.1), *nouveau* [nuvo] - *nuovo* [nwovo] ('new'; PMI = 7.0), and *montagne* [motaj] - *montagna* [monta5a] ('mountain'; PMI = 4.9).

The final PMI scores between sound classes are visualized in Fig. 1. It is easy to discern that $PMI(a,a)$ is positive for all sound classes a, and that $PMI(a,b)$ for $a \neq b$ is negative in most cases. There are a few pairs a,b with positive score, such as b/f. Generally, sound class pairs with a similar place of articulation tend to have relatively high scores. This pattern is also visible in the cluster dendrogram. We observe a primary split between vowels and consonants. Consonants are further divided into labials, dentals, and velar/uvular sounds.

It should be noted that the pairwise PMI-scores between sound classes measure something like the propensity of these sound classes to participate in a sound correspondence; it does not say anything about

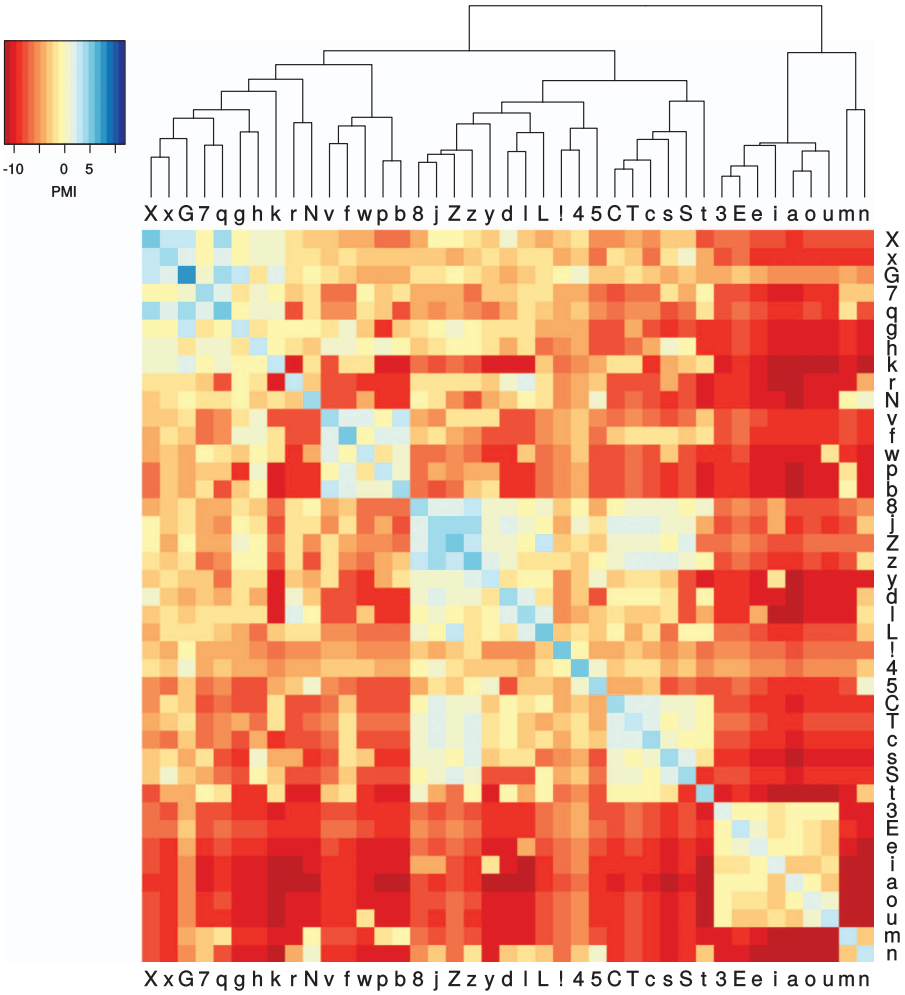

Figure 1. **PMI scores.** Heatmap and hierarchical clustering dendrogram.

sound correspondences between specific pairs of languages. The ASJP word lists, with just 40 items, are definitely too short to infer sound correspondences between individual language pairs.

Pairwise distances between languages

When aggregating PMI similarities between individual words into a distance measure between word lists, various complicating factors have to be taken into consideration:

- Entries for a certain language and a certain concept often contain several synonyms. This is a potential source of bias when averaging PMI similarities of individual word pairs. (As correctly pointed out by a reviewer, this effect is aggrevated by the observation that the number of synonyms may reflect the extent and quality of the language's documentation. Currently I see no way to control for this kind of bias).
- Cognate words tend to be more similar than non-cognate ones. However, the average similarity level between non-cognate words depends on the overall similarity between the sound inventories and phonotactic structure of the languages compared. To assess the informativeness of a certain PMI similarity score, it has to be calibrated against the overall distribution of PMI similarities between non-cognate words from the languages in question.
- Many ASJP word lists are incomplete, so the word lists are of unequal length.

To address the first problem, [19] defined the similarity score between languages l_1 and l_2 for concept c as the maximal PMI similarity between any pair of entries for c from l_1 and l_2.

The second problem is addressed by estimating, for each concept c for which both languages have an entry, the p-value for the null hypothesis that none of the words for c being compared are cognate. This is done in a parameter-free way. For each pair of concepts (c_1,c_2), the PMI similarities between the words for c_1 from l_1 and the words for c_2 from l_2 are computed. The maximum of these values is the similarity score for (c_1,c_2). Under the simplifying assumption that cognate words always share their meaning the

distribution of such similarity scores for $c_1 \neq c_2$ constitutes a sample of the overall distribution of similarity scores between non-cognates.

It should be noted that the assumption of general synomymity of cognates is evidently false when considering the entire lexicon. There is a plethora of examples, such as as English *deer* vs. German *Tier* 'animal', which are cognate (cf.[28], p. 94) without being synonyms. However, within the 40-concept core vocabulary space covered by ASJP, such cross-concept cognate pairs are arguably very rare.

Now consider the null hypothesis that the words for concept c are non-cognate. We assume *a priori* that cognate word pairs are more similar than non-cognate ones. Let the similarity score for c be x. The maximum likelihood estimate for the p-value of that null hypothesis is the relative frequency non-cognate pairs with a similarity score $\geq x$. If $PMI(c_i,c_j)$ is the similarity &score between concept c_i and c_j, we have

$$p_c = \frac{\left|\{(c,c)\} \cup \{(c_i,c_j) \middle| c_i \neq c_j \ \&\ PMI(c_i,c_j) \geq PMI(c,c)\right|}{\left|\{(c,c)\} \cup \{(c_i,c_j) \middle| c_i \neq c_j\}\right|}. \quad (3)$$

Analogously to Fisher's method[29], the p-values for all concepts are combined according to the formula

$$\sum_c -\log p_c. \quad (4)$$

If the null hypothesis is true for concept c, p_c is distributed approximately according to a continuous uniform distribution over the interval (0, 1]. Accordingly, –log p_c is distributed according to an exponential distribution with mean and variance $=1$.

Suppose there are N concepts for which both l_1 and l_2 have an entry. The sum of N independently distributed random variables, each with mean and variance $=1$, approximately follows a normal distribution with mean $=N$ and variance $=N$. This can be transformed into a Z-statistic by normalizing according to the formula

$$Z(l_1,l_2) = \frac{\sum_{i=1}^{N} -\log p_{c_i} - N}{\sqrt{N}} \quad (5)$$

This normalization step addresses the third issue mentioned above, i.e., the varying length of word lists.

$Z(l_1, l_2)$ increases with the degree of similarity between l_1 and l_2. It is transformed into a dissimilarity measure as follows: (We will talk of *distance measure* from now on for simplicity, even though it is not a metric distance).

$$d(l_1,l_2) = \frac{Z_{\max} - Z(l_1,l_2)}{Z_{\max} - Z_{\min}} \quad (6)$$

The maximal possible value Z_{mix} for Z would be achieved if both word lists have the maximal length of $N=40$, and each synonymous word pair has a higher PMI score than any non-synonymous word pair. Therefore

$$Z_{\max} = \frac{40 \times -\log \frac{1}{40^2-40+1} - 40}{\sqrt{40}} \approx 40.18$$

The minimal value $Z_{\min}$ for Z would be achieved if all p_c equal 1 and both word lists have length 40:

$$Z_{\min} = \frac{40 \times -\log 1 - 40}{\sqrt{40}} = -\sqrt{40} \approx -6.32$$

We computed $d(l_1, l_2)$ for each pair of the above-mentioned 6,892 languages from the ASJP database. This distance matrix is available at (Data Citation 3).

Automatic cognate classification

Background. In[20] a method is developed to cluster words into equivalence classes in a way that approximates manual expert classifications. In this section this approach is briefly sketched.

The authors chose a supervised learning approach. They use word lists with manual expert cognate annotations from a diverse collection of language families, taken from[15–18]; http://ielex.mpi.nl. A part of these gold standard data were used to train a *Support Vector Machine* (SVM). For each pair of words (w_1,w_2) from languages (l_1, l_2), denoting concept c, seven feature values were computed:

1. **PMI similarity.** This is the string similarity measure according to[19] as described in the previous section.
2. **Calibrated PMI distance.** p_c as defined in equation (3) above.
3. The negative logarithm thereof.
4. **Language similarity.** $Z(l_1, l_2)$, as defined in equation (5) above.
5. The logarithm thereof.
6. **Average word length** of words for concept c across all languages from the database, measured in number of symbols in ASJP transcription.

Dataset	Source	Words	Concepts	Languages	Families	Cognate classe
ABVD	[15]	2,306	34	100	Austronesian	409
Afrasian	[48]	770	39	21	Afro-Asiatic	351
Chinese	[49]	422	20	18	Sino-Tibetan	126
Huon	[50]	441	32	14	Trans-New Guinea	183
IELex	http://ielex.mpi.nl	2,089	40	52	Indo-European	318
Japanese	[51]	387	39	10	Japonic	74
Kadai	[52]	399	40	12	Tai-Kadai	102
Kamasau	[53]	270	36	8	Torricelli	59
Mayan	[6]	1,113	40	30	Mayan	241
Miao-Yao	[52]	206	36	6	Hmong-Mien	69
Mixe-Zoque	[54]	355	39	10	Mixe-Zoque	79
Mon-Khmer	[52]	579	40	16	Austroasiatic	232
ObUgrian	http://starling.rinet.ru	769	39	21	Uralic	68
total		10,106	40	318	13	2,311

Table 1. Gold standard data used for this study.

7. **Concept-language correlation.** The Pearson correlation coefficient between feature 3 and feature 4 for all word pairs expressing concept *c*.

For each such word pair, the gold standard contains an evaluation as *cognate* (1) or *not cognate* (0). An SVM was trained to predict these binary cognacy labels. Applying Platt scaling[30], the algorithm predics a *probability of cognacy* for each pair of words from different languages denoting the same concept. These probabilities were used as input for hierarchical clustering, yielding a partitioning of words into equivalence classes for each concept.

The authors divided the gold standard data into a training set and a test set. Using an SVM trained with the training set, they achieve B-cubed F-scores[31] between 66.9% and 90.9% on the data sets in their test data, with a weighted average of 71.8% when comparing automatically inferred clusters with manual cognate classifications. (The *F-score* is an aggregate measure relying both on the *precision* — the proportion of predicted cognates that actually are cognate — and the *recall* — the proportion of true cognates that are correctly identified).

In[20] it is shown that this approach leads to slightly improved results if compared with LexStat[32], which can be considered as state of the art. In[21] it is furthermore demonstrated that the SVM-based approach is especially superior when applied to short and poorly transcribed word lists, while the differences virtually level out for longer and high-quality lists.

Creating a gold standard

We adapted this approach to the task of performing automatic cognate classification on the ASJP data. Since ASJP contains data from different families and it is confined to 40 core concepts (while the data used in[20] partially cover 200-item concept lists), the method has to be modified accordingly.

We created a gold standard dataset from the data used in[21] (which is is drawn from the same sources as the data used in[20] but has been manually edited to correct annotation mistakes). Only the 40 ASJP concepts were used. Also, we selected the source data in such a way that each dataset is drawn from a different language family. Words from different families were generally classified as non-cognate in the gold standard. All transcriptions were converted into ASJP format. Table 1 summarizes the composition of the gold standard data.

Clustering

We used the *Label Propagation* algorithm[33] for clustering. For each concept, a network is constructed from the words for that concept. Two nodes are connected if and only if their predicted probability of cognacy is $\geq$0.25. This threshold was chosen somewhat arbitrarily, based on manual trial and error. *Label Propagation* detects community structures within the network, i.e., it partitions the nodes into clusters.

Model selection

To identify the set of features suitable for clustering the ASJP data, we performed *cross-validation* on the gold standard data. The data were split into a *training set*, consisting of the data from six randomly chosen language families, and a *test set*, consisting of the remaining data. We slightly deviated from[20] by replacing features 4 and 5 by *language distance* $d(l_1, l_2)$ as defined in equation (6), and $-\log(1-d(l_1, l_2))$.

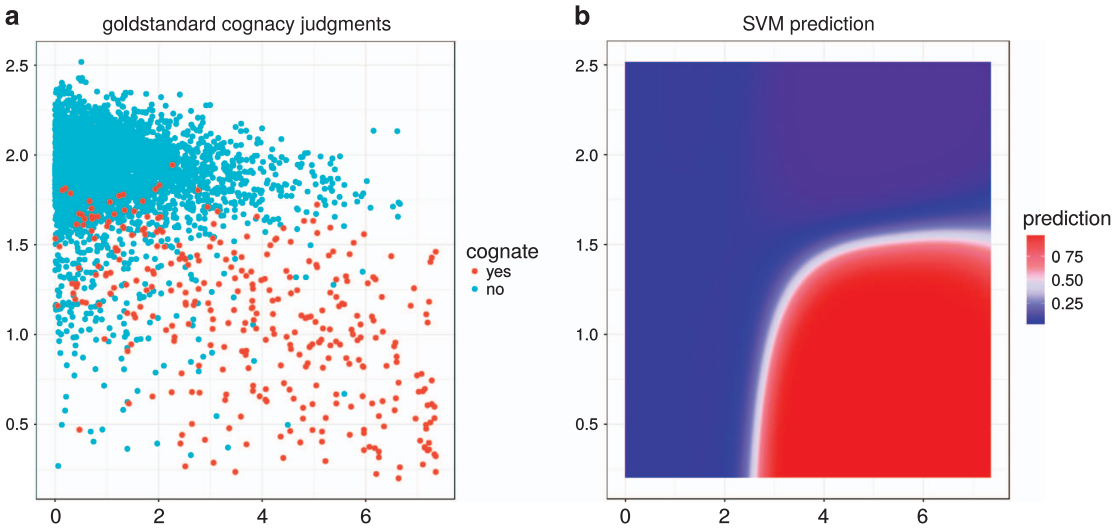

Figure 2. Goldstandard vs. automatic classification. Expert cognacy judgments (**a**) and prediction of cognacy (**b**) depending on the selected features.

Both are linear transformations of the original features and therefore do not affect the automatic classification.

For each of the 127 non-empty subsets of the seven features, an SVM with an RBF-kernel was trained with 7,000 randomly chosen synonymous word pairs from the training set. Explorative tests revealed that accuracy of prediction does not increase if more training data are being used. The trained SVM plus Platt scaling were used to predict the probability of cognacy for each synonymous word pair from the test set, and the resulting probabilities were used for Label Propagation clustering. This procedure was repeated ten times for random splits of the gold standard data into a training set and a test set.

For each feature combination, the B-cubed F-score, averaged over the ten training/test splits, was determined. The best performance (average B-cubed F-score: 0.86) was achieved using just two features:

- **Word similarity.** The negative logarithm of the calibrated PMI distance, and
- **Language log-distance.** $-\log(1-d(l_1, l_2))$, with $d(l_1, l_2)$ as defined in equation (6).

Figure 2 displays, for a sample of gold standard data, how expert cognacy judgments depend on these features and how the trained SVM + Platt scaling predicts cognacy depending on those features. Most cognate pairs are concentrated in the lower right corner of the feature space, i.e., they display both high word similarity and low language log-distance. The SVM learns this non-linear dependency between the two features.

Clustering the ASJP data

A randomly selected sample of 7,000 synonymous word pairs from the gold standard data were used to train an SVM with an RBF-kernel, using the two features obtained via model selection. Probabilities of cognacy for all pairs of synonymous pairs of ASJP entries were obtained by (a) computing word similarity and language log-distance, (b) predict their probability of cognacy using the trained SVM and Platt scaling, and (b) apply Label Propagation clustering.

Phylogenetic inference

Distance-based. The language distances according to the definition in equation (6) can be used as input for distance-based phylogenetic inference. In the experiments reported below, we used the BIONJ[34] algorithm for that purpose.

Character-based. We propose two methods to extract discrete character matrices from the ASJP data.

1. **Automatically inferred cognate classes.** We defined one character per automatically inferred (in the sense described above) cognate class *cc*. If the word list for a language *l* has a missing entry for the concept the elements of *cc* refer to, the character is undefined for this language. Otherwise *l* assumes value 1 if its word list contains an element of *cc*, and 0 otherwise.
2. **Soundclass-concept characters.** We define a character for each pair (*c*,*s*), where *c* is a concept *c* and *s* an ASJP sound class. The character for (*c*,*s*) is undefined for language *l* if *l*'s word list has a missing

entry for concept *c*. Otherwise *l*'s value is 1 if one of the words for *c* in *l* contains symbol *s* in its transcription, and 0 otherwise.

The motivation for these two types of characters is that they track two different aspects of language change. Cognacy characters contain information about lexical changes, while soundclass-concept characters also track sound changes within cognate words. Both dimensions provide information about language diversification.

Let us illustrate this with two examples.

- The Old English word for 'dog' was *hund*, i.e., `hund` in ASJP transcription. It belongs to the automatically inferred cognate class *dog_149*. The Modern English word for that concept is *dog*/`dag`, which belongs to class *dog_150*. This amounts to two mutations of cognate-class characters between Old English and Modern English, $0 \rightarrow 1$ for *dog_150* and $1 \rightarrow 0$ for *dog_149*.The same historic process is also tracked by the sound-concept characters; it corresponds to five mutations: $0 \rightarrow 1$ for *dog:a* and *dog:g*, and $1 \rightarrow 0$ for *dog:h*, *dog:u*, and *dog:n*.
- The word for 'tree' changed from Old English *treow* (`treow`) to Modern English *tree* (`tri`). Both entries belong to cognate class *tree_17*. As no lexical replacement took place for this concept, there is no mutation of cognate-class characters separating Old and Modern English here. The historical sound change processes that are reflected in these words are captured by mutations of sound-concept characters: $0 \rightarrow 1$ for *tree:i* and $1 \rightarrow 0$ for *tree:e*, *tree:o* and *tree:w*.

For a given sample of languages, we use all *variable* characters (i.e., characters that have value 1 and value 0 for at least one language in the sample) from both sets of characters. Phylogenetic inference was performed as Maximum-Likelihood estimation assuming Γ-distributed rates with 25 rate categories, and using ascertainment bias correction according to[35]. Base frequencies and variance of rate variation were estimated from the data.

In our phylogenetic experiments, the distance-based tree was used as initial tree for tree search. This method was applied to three character matrices:

- cognate class characters,
- soundclass-concept characters, and
- a partitioned analysis using both types of characters simultaneously.

Inference was performed using the software RAxML[36].

Applying more advanced methods of character-based inference, such as Bayesian inference[37–39] proved to be impractical due to hardware limitations.

Code availability

The code used to conduct this study is freely available at (Data Citation 3). The workflow processes the sub-directories in the following order: 1. `pmiPipeline`, 2. `cognateClustering`, and 3. `validation`. All further details, including software and software versions used, are described in the `README` files in the individual sub-directories and sub-sub-directories.

Data Records

All data that were produced are available at (Data Citation 3) as well.

Phylogenetic trees

Accessible from subdirectory `trees/` of (Data Citation 3).

- a family for each of the 66 Glottolog families comprising at least 10 languages in ASJP; the trees were inferred using Maximum Likelihood with combined characters (see Technical Validation for details), using the Glottolog classification as constraint tree. Rooting was performed as described in subsection **A case study: punctuated language evolution**.
- a tree over all 6,982 ASJP languages (`world.tre`), using the same inference methods, but applying midpoint rerooting.

PMI data

Accessible from subdirectory `pmiPipeline/` of (Data Citation d3).

- estimated PMI scores (`pmiScores.csv/`) and gap penalties (`gapPenalties`.csv)
- pairwise distances between languages (`pmiWorld.csv`)

Automatic cognate classification

Accessible from subdirectory `cognateClustering/` of (Data Citation 3).

- word list with automatically inferred cognate class labels (`asjp17Clustered.csv`)

Phylogenetic inference

Accessible from subdirectory `validation/` of (Data Citation 3).

- family-wise data and trees (described in Subsection *Phylogenetic Inference* within the Section *Technical Validation*) are in sub-directory `validation/families/`. For each Glottolog family `F`, there are the following files (replace `[F]` by name of the family):

1. `[F].cc.phy:` character matrix, cognate class characters, Phylip format
2. `[F].sc.phy:` character matrix, soundclass-concept characters, Phylip format
3. `[F].cc_sc.phy:` combined character matrix, cognate class and soundclass-concept characters, Phylip format
4. `[F].part.txt:` partition file
5. `[F].pmi.nex:` pairwise `PMI` distances, Nexus format
6. `[F].pmi.tre:BIONJ` tree, inferred from `PMI` distances, Newick format
7. `glot.[F].tre:` Glottolog tree, Newick format
8. `RAxML_bestTree.[F]_cc:` Maximum Likelihood tree, inferred from cognate class characters, Newick format
9. `RAxML_bestTree.[F]_sc:` Maximum Likelihood tree, inferred from soundclass-concept characters, Newick format
10. `RAxML_bestTree.[F]_cc_sc:` Maximum Likelihood tree, inferred from combined character matrix, Newick format

- global data over all 6,892 languages in the database are in the sub-directory `validation/`, and global trees in the sub-directory `validation/worldTree/:`

1. `validation/world_cc.phy:` character matrix, cognate class characters, Phylip format
2. `validation/world_sc.phy:` character matrix, soundclass-concept characters, Phylip format
3. `validation/world_sc_cc.phy:` combined character matrix, cognate class and soundclass-concept characters, Phylip format
4. `validation/world.partition.txt:` partition file
5. `validation/glottologTree.tre:` Glottolog tree, Newick format
6. `validation/worldTree/distanceTree.tre:` BIONJ tree, inferred from PMI distances, Newick format
7. `validation/worldTree/RAxML_bestTree.world_cc:` Maximum Likelihood tree, inferred from cognate class characters, Newick format
8. `validation/worldTree/RAxML_bestTree.world_sc:` Maximum Likelihood tree, inferred from soundclass-concept characters, Newick format
9. `validation/worldTree/RAxML_bestTree.world_sc_cc:` Maximum Likelihood tree, inferred from combined character matrix, Newick format
10. `validation/worldTree/RAxML_bestTree.world_sc_ccGlot:` Maximum Likelihood tree, inferred from combined character matrix using the Glottolog classifcation as constraint tree, Newick format

For language names, I generally follow the convention `[WALS family].[WALS genus].[ASJP doculect name]`, all in upper case. English, e.g., is named `IE.GERMANIC.ENGLISH`. If necessary, the corresponding iso codes and other meta-data can easily be accessed from the original ASJP database.

Technical Validation

Phylogenetic inference

To evaluate the usefulness of the distance measure and the character matrices defined above for phylogenetic inference, we performed two experiments:

- **Experiment 1**. We applied both distance-based inference and character-based inference for all language families (according to the Glottolog classification) containing at least 10 languages in ASJP.
- **Experiment 2**. We sampled 100 sets of languages with a size between 20 and 400 at random and applied all four methods of phylogenetic inference to each of them.

Method character type	Character-based			Distance-based
	Cognate classes	Soundclass-concept	Combined	
total	0.215	0.186	0.173	0.148
small families (n<20)	0.193	0.219	0.219	0.100
medium families ($20 \leq n<100$)	0.249	0.179	0.171	0.168
large families ($n \geq 100$)	0.147	0.147	0.102	0.154

Table 2. Median Generalized Quartet Distances for Glottolog families.

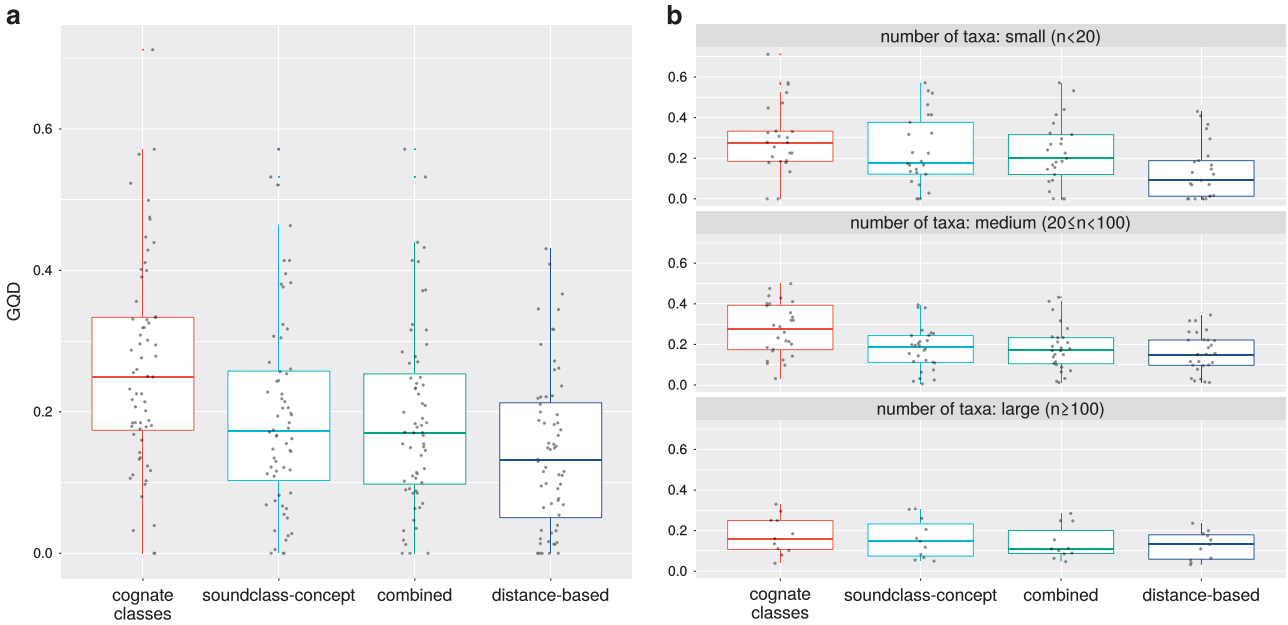

Figure 3. Experiment 1 computed Generalized Quartet Distances for Glottolog families depending on phylogenetic inference method. Aggregated over all families (**a**) and split according to family size (**b**).

In both experiments, each automatically inferred phylogeny was evaluated by computing the *Generalized Quartet Distance* (GQD)[40] to the Glottolog expert tree (restricted to the same set of languages). The GQD gives the proportion quartets of languages that are grouped differently by the expert tree and the automatically generated tree. Quartets that are not *resolved* in the expert tree (e.g., because they come from four different families) are not counted.

The results of the first experiment are summarized in Table 2 and visualized in Fig. 3. The results for the individual families are given in Table 3 (available online only).

Aggregating over all families suggests that distance-based inference produces the best fit with the expert gold standard. However, a closer inspection of the results reveals that the performance of the different phylogenetic inference methods depend on the size of the language families (measured in number of taxa available in ASJP). Combining both types of characters in a partitioned model always leads to better results than the two character types individually. While distance-based inference is superior for small language families (less than 20 taxa), character-based inference appears to be about equally good for medium-sized (20–199 taxa) and large (more than 200 taxa) language families.

This assessment is based on a small sample size since there are only 33 medium-sized and 6 large language families. The results of experiment 2 confirm these conclusions though. They are summarized in Table 4 and illustrated in Fig. 4.

All four methods improve with growing sample size, but this effect is more pronounced for character-based inference. While combined character-based inference and distance-based inference are comparable in performance for smaller samples of languages ($n \leq 100$), character-based inference outperforms distance-based inference for larger samples, and the difference grows with sample size.

Method character type	Character-based			Distance-based
	Cognate classes	Soundclass-concept	Combined	
total	0.187	0.147	0.066	0.130
$20 \leq n \leq 100$	0.286	0.210	0.099	0.174
$100 < n \leq 200$	0.226	0.135	0.065	0.115
$200 < n \leq 300$	0.157	0.136	0.063	0.132
$300 < n \leq 400$	0.131	0.132	0.061	0.114

Table 4. Median Generalized Quartet Distances to Glottolog for random samples of languages.

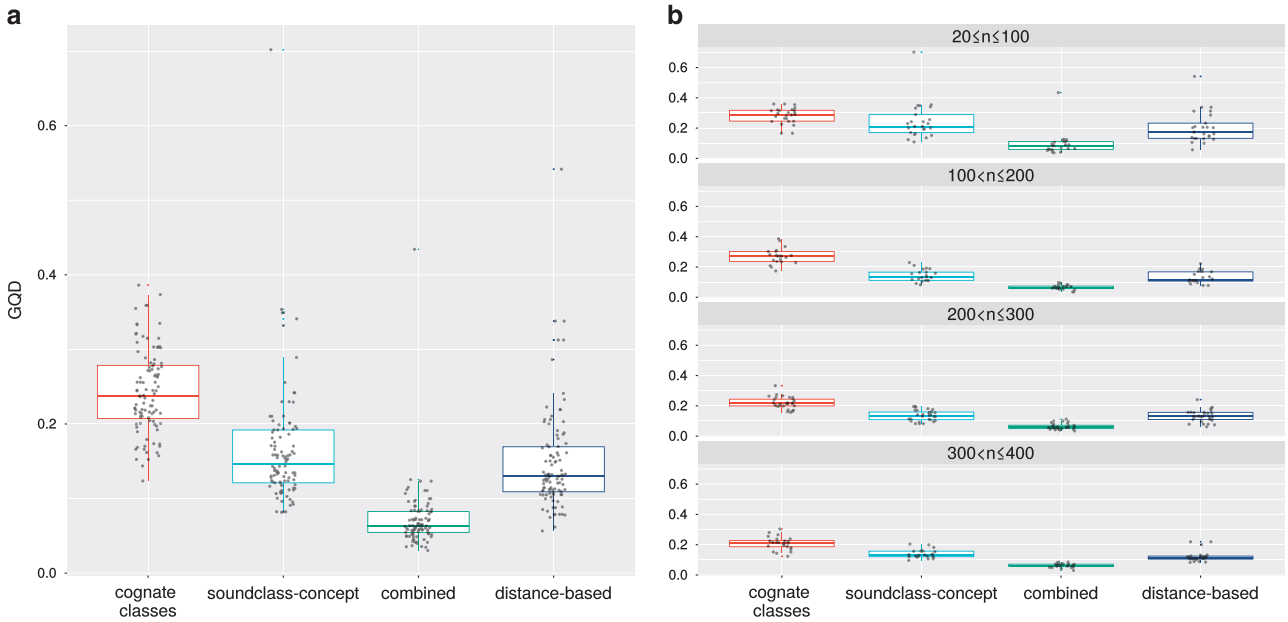

Figure 4. Experiment 2 computed Generalized Quartet Distances for random samples of languages depending on phylogenetic inference method. Aggregated over all samples (**a**) and split according to sample size (**b**).

The same pattern is found when the different versions of phylogenetic inference is applied to the full dataset of 6,892 languages. We find the following GQD values:

- distance based tree: 0.078
- cognate-class based ML tree: 0.052
- soundclass-concept based ML tree: 0.089
- ML tree from combined character data: 0.035

Relation to geography

Both the distances between languages and the two methods to represent languages as character vectors are designed to identify similarities between word lists. There are essentially three conceivable causal reasons why the word lists from two languages are similar: (1) common descent, (2) language contact and (3) universal tendencies in sound-meaning association due to sound symbolism, nursery forms etc.[41]. The third effect is arguably rather weak. The signal derived from common descent and from language contact should be correlated with geographic distance. If the methods proposed here extract a genuine signal from word lists, we thus expect to find such a correlation.

To test this hypothesis, we computed the geographic distance (great-circle distance) between all pairs from a sample of 500 randomly selected languages, using the geographic coordinates supplied with the ASJP data.

We furthermore extracted pairwise distances from character vectors by computing the cosine distance between those vectors, using only characters for which both languages have a defined value. In this way we obtained three matrices of pairwise linguistic distances for the mentioned sample of 500 languages: (1)

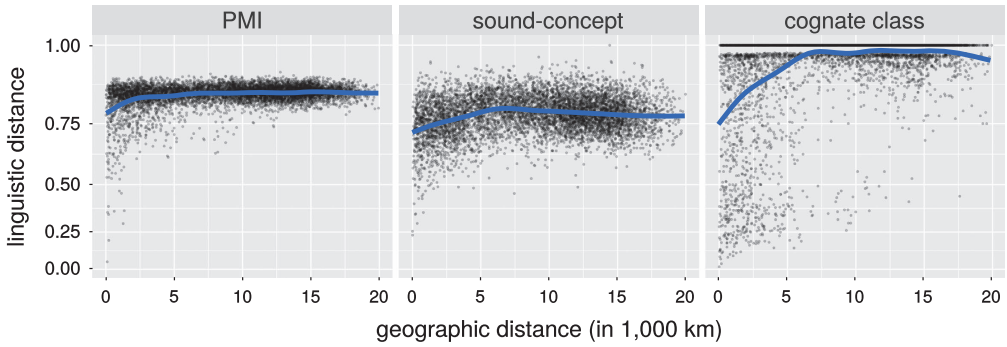

Figure 5. **Geographic vs. linguistic distances between languages.**

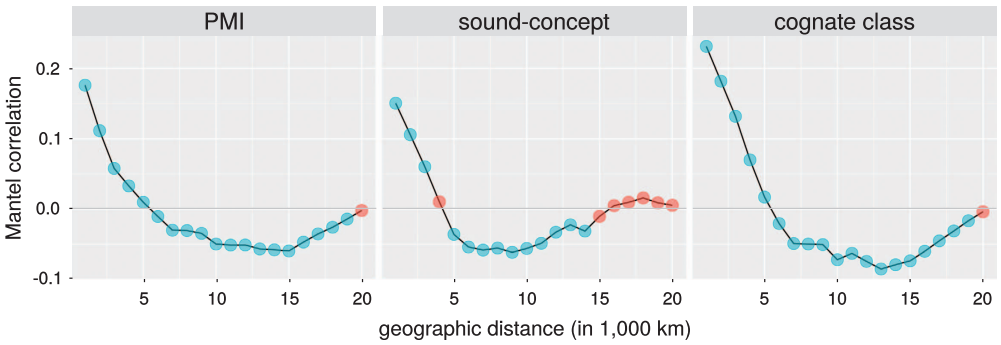

Figure 6. **Mantel correlograms. Blue: significant, red: non-significant at $p < 0.05$.**

The distance as defined in equation (6), called **PMI distance**, (2) the cosine distance between the cognate-class vectors, and (3) the cosine distance between the sound-concept vectors.

All three linguistic distance measures show a significant correlation with geographic distance. The Spearman correlation coefficients are 0.193 for PMI distances, 0.280 for cognate-class distance and 0.0888 for sound-concept distance. Figure 5 shows the corresponding scatter. The p-values according to the Mantel test are $\leq$0.0001 in all three cases. Figure 5 shows the corresponding scatter plots.

The visualization suggests that for all three linguistic distance measures, we find a signal at least up to 5,000 km. This is confirmed by the Mantel correlograms[42] shown in Fig. 6. We find a significant positive correlation with geographic distance for up to 5,000 km for PMI distance, and up to 4,000 km for cognate-class distance and sound-concept distance.

Usage Notes

Character-based inference from expert cognacy judgment data have been used in various downstream applications beyond phylogenetic inference, such as estimating the time course of prehistoric population events[3,7,9] or the identification of overarching patterns of cultural language evolution[5,43]. In this section it will be illustrated how the automatically inferred characters described above can be deployed to expand the scope of such investigations to larger collections of language families. It also demonstrates that automatically inferred branch lengths of phylogenetic trees — a kind of information for which we do not have any manually collected data — provides useful information about language history.

A case study: punctuated language evolution

A few decades ago, [44] proposed that biological evolution is not, in general, a gradual process. Rather, they propose, long periods of stasis are separated by short periods of rapid change co-occurring with branching speciation. This model goes by the name of *punctuated equilibrium*. This proposal has initiated a lively and still ongoing discussion in biology. Pagel, Venditti and Meade[45] developed a method to test a version of this hypothesis statistically. They argue that most evolutionary change occurs during speciation events. Accordingly, we expect a positive correlation between the number of speciation events a lineage underwent throughout its evolutionary history and the amount of evolutionary change that happened during that time.

Estimates of both quantities can be read off a phylogenetic tree — the number of speciation events corresponds to the number of branching nodes, and the amount of change to the total path length — provided (a) the tree is rooted and (b) branch lengths reflect evolutionary change (e.g., the expected

Family	Slope	*p*-value	Number of taxa	Significant
Atlantic-Congo	0.003	< 1E-14	1332	yes
Austronesian	0.005	< 1E-14	1259	yes
Afro-Asiatic	0.008	2E-13	356	yes
Sino-Tibetan	0.005	9E-8	279	yes
Indo-European	0.004	2E-7	367	yes
Nuclear_Trans_New_Guinea	0.003	7E-4	259	yes
Pama-Nyungan	0.005	6E-4	167	yes
Tai-Kadai	0.007	8E-3	142	no
Kiwaian	0.024	9E-3	10	no
Nakh-Daghestanian	0.011	0.01	55	no
Turkic	0.009	0.02	60	no
Quechuan	0.006	0.03	62	no
Siouan	0.004	0.04	17	no
Cariban	0.016	0.05	30	no
Eskimo-Aleut	−0.052	0.05	10	no
Central_Sudanic	0.010	0.07	58	no
Salishan	0.015	0.08	30	no
Chibchan	0.011	0.10	23	no
Ainu	0.013	0.10	22	no
Dravidian	0.008	0.10	38	no
Sko	0.038	0.10	14	no
Uralic	0.017	0.12	30	no
Ndu	0.025	0.14	10	no
Lower_Sepik-Ramu	0.070	0.15	19	no
Japonic	0.010	0.17	32	no
Gunwinyguan	0.027	0.21	14	no
Heibanic	−0.041	0.24	11	no
Khoe-Kwadi	0.015	0.39	12	no
Tungusic	0.011	0.40	25	no
Tucanoan	−0.011	0.45	32	no
Angan	0.018	0.46	17	no
Cochimi-Yuman	0.005	0.47	13	no
Chocoan	−0.027	0.51	10	no
Kadugli-Krongo	−0.006	0.71	11	no
Pano-Tacanan	0.001	0.78	33	no
Tupian	0.001	0.80	59	no
Totonacan	−0.002	0.80	14	no
Ta-Ne-Omotic	−0.003	0.83	24	no
Algic	−0.009	0.86	32	no
Lakes_Plain	−0.002	0.89	22	no
Timor-Alor-Pantar	0.005	0.92	59	no
Bosavi	−0.003	0.99	13	no

Table 5. Test for punctuated language evolution for the families without node density artifact. Significance is determined via Holm-Bonferroni correction at the significance level of 0.05.

number of mutations of a character) rather than historical time. In[45] a significant correlation is found for biomolecular data, providing evidence for punctual evolution.

In[43], the same method is applied to the study of language evolution, using expert cognacy data from three language families (Austronesian, Bantu, Indo-European). The study results in strong evidence for punctuated evolution in all three families.

We conducted a similar study for all Glottolog language families with at least 10 ASJP languages. The workflow was as follows. For each family *F*:

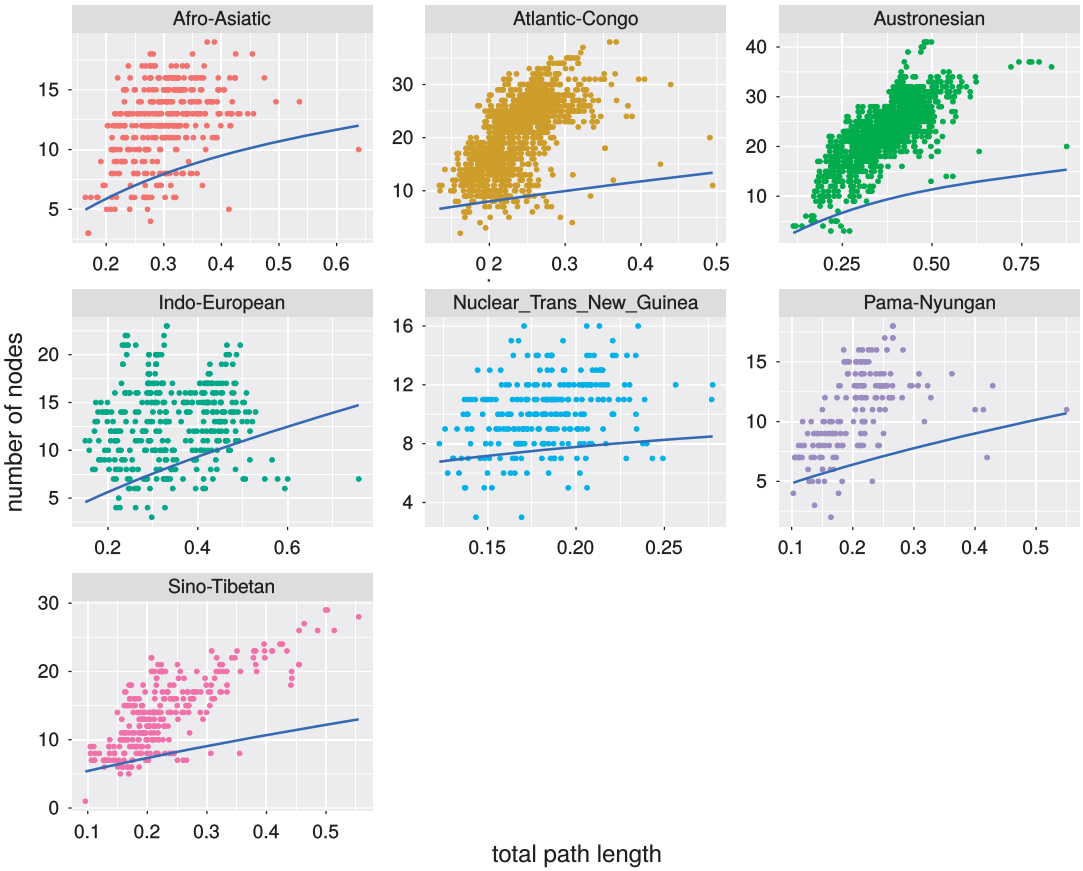

Figure 7. Dependency between total path length and the number of branching nodes for the families with a significantly positive association. Blue lines are regression lines according to phylogenetic generalized least squares, using δ-correction.

- Find the language $o \notin F$ which has the minimal average PMI distance to the languages in F. This language will be used as *outgroup*.
- Infer a Maximum-Likelihood tree over the taxa $F \cup \{o\}$ with the Glottolog classification as constraint tree, using a partitioned analysis with cognate-class characters and soundclass-concept characters.
- Use o as outgroup to root the tree; remove o from the tree.
- Apply the *δ-test*[46] to control for the *node density artifact*.
- Perform *Phylogenetic Generalized Least Square*[47] regression with root-to-tip path lengths for all taxa as independent and root-to-tip number of nodes as dependent variable.
- If the δ-test is negative and the regression results in a significantly positive slope, there is evidence for punctuated evolution in F.

Among the 66 language families considered, the δ-test was negative for 43 families. We applied Holm-Bonferroni correction for multiple testing to determine significance in the regression analysis. The numerical results are given in Table 5.

A significant positive dependency was found for the seven largest language families (Atlantic-Congo, Austronesian, Indo-European, Afro-Asiatic, Sino-Tibetan, Nuclear Trans-New Guinea, Pama-Nyungan). The relationships for these families are visualized in Fig. 7. No family showed a significant negative dependency. This strengthend the conclusion of Atkinson et al.[43] that languages evolve in punctuational bursts.

References

1. Atkinson, Q. D. & Gray, R. Curious parallels and curious connections — phylogenetic thinking in biology and historical linguistics. *Systematic Biology* **54,** 513–526 (2005).
2. Levinson, S. C., D., R. & Gray, R. Tools from evolutionary biology shed new light on the diversification of languages. *Trends in Cognitive Sciences* **16,** 167–173 (2012).
3. Gray, R. D. & Jordan, F. M. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405,** 1052–1055 (2000).
4. Dunn, M., Terrill, A., Reesink, G., Foley, R. A. & Levinson, S. C. Structural phylogenetics and the reconstruction of ancient language history. *Science* **309,** 2072–2075 (2005).

5. Pagel, M., Atkinson, Q. D. & Meade, A. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449,** 717–720 (2007).
6. Brown, C. H., Holman, E. W., Wichmann, S. & Velupillai, V. Automated classification of the world's languages: A description of the method and preliminary results. *STUF — Language Typology and Universals* **4,** 285–308 (2008).
7. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323,** 479–483 (2009).
8. Dunn, M., Greenhill, S. J., Levinson, S. & Gray, R. D. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* **473,** 79–82 (2011).
9. Bouckaert, R. *et al.* Mapping the origins and expansion of the Indo-European language family. *Science* **337,** 957–960 (2012).
10. Bowern, C. & Atkinson, Q. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* **88,** 817–845 (2012).
11. Bouchard-Côté, A., Hall, D., Griffiths, T. L. & Klein, D. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences* **36,** 141–150 (2013).
12. Pagel, M., Atkinson, Q. D., Calude, A. S. & Meade, A. Ultraconserved words point to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences* **110,** 8471–8476 (2013).
13. Hruschka, D. J. *et al.* Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* **25,** 1–9 (2015).
14. Jäger, G. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences* **112,** 12752–12757 (2015). Doi: 10.1073/pnas.1500331112.
15. Greenhill, S. J., Blust, R. & Gray, R. D. The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* **4,** 271–283 (2008).
16. Wichmann, S., Holman, E. W. Languages with longer words have more lexical change. InBorin L. & Saxena A.eds. *Approaches to Measuring Linguistic Differences* 249–284 (Mouton de Gruyter Berlin, 2013).
17. List, J.-M. *Data from: Sequence comparison in historical linguistics GitHub Repository* http://github.com/SequenceComparison/SupplementaryMaterial (2014).
18. Mennecier, P., Nerbonne, J., Heyer, E. & Manni, F. A Central Asian language survey: Collecting data, measuring relatedness and detecting loans. *Language Dynamics and Change* **6,** 57–98 (2016).
19. Jäger, G. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change* **3,** 245–291 (2013).
20. Jäger, G., Sofroniev, P. Automatic cognate classification with a Support Vector Machine. InDipper S., Neubarth F. & Zinsmeister H.eds. *Proceedings of the 13th Conference on Natural Language Processing, vol. 16 of Bochumer Linguistische Arbeitsberichte* 128–134Ruhr Universität Bochum, (2016).
21. Jäger, G., List, J.-M. & Sofroniev, P. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. *In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. (ACL, 2017).
22. Nunn, C. L. *The Comparative Approach in Evolutionary Anthropology and Biology*. The University of Chicago Press Chicago, (2011).
23. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48,** 443–453 (1970).
24. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).
25. Holman, E. W. *et al.* Advances in automated language classification. InArppe A., Sinnemäki K., Nikanne U.eds. *Quantitative Investigations in Theoretical Linguistics* 40–43 (University of Helsinki, 2008).
26. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* **29,** 1–38 (1977).
27. Nelder, J. A. & Mead, R. A simplex method for function minimization. *The computer journal* **7,** 308–313 (1965).
28. Kroonen, G. *Etymological Dictionary of Proto-Germanic*. (Brill Leiden: Boston, 2013).
29. Fisher, R. A. *Statistical methods for research workers*. (Genesis Publishing Pvt Ltd, 1925).
30. Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers* 61–74 (MIT Press, 1999).
31. Bagga, A. & Baldwin, B. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1* 79–85Association for Computational Linguistics (1998).
32. List, J.-M. Lexstat: Automatic detection of cognates in multilingual wordlists. In Butt M. & Prokić J.eds *Proceedings of LINGVIS & UNCLH, Workshop at EACL 2012* 117–125 (Avignon, 2012).
33. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E* **76,** 036106 (2007).
34. Gascuel, O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14,** 685–695 (1997).
35. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50,** 913–925 (2001).
36. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
37. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574 (2003).
38. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution* **29,** 1969–1973 (2012).
39. Pagel, M. & Meade, A. *BayesPhylogenies 2.0. software distributed by the authors* (2015).
40. Pompei, S., Loreto, V. & Tria, F. On the accuracy of language trees. *PLoS One* **6,** e20109 (2011).
41. Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F. & Christiansen, M. H. Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences* **113,** 10818–10823 (2016).
42. Legendre, P. & Legendre, L. F. J. *Numerical Ecology*. Elsevier: Amsterdam/Oxford, (2012).
43. Atkinson, Q. D., Meade, A., Venditti, C., Greenhill, S. J. & Pagel, M. Languages evolve in punctuational bursts. *Science* **319,** 588–588 (2008).
44. Gould, S. J. & Eldredge, N. Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology* **3,** 115–151 (1977).
45. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314,** 119–121 (2006).
46. Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny reconstruction. *Systematic Biology* **55,** 637–643 (2006).

47. Grafen, A. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 119–157 (1989).
48. Militarev, A. I. *Towards the chronology of Afrasian (Afroasiatic) and its daughter families*. McDonald Institute for Archaelogical Research Cambridge, (2000).
49. Běijīng Dàxué *Hànyǔ fāngyán cíhuì [Chinese dialect vocabularies]*. (Wénzì Géigé, 1964).
50. McElhanon, K. A. Preliminary observations on Huon Peninsula languages. *Oceanic Linguistics* **6,** 1–45 (1967).
51. Hattori, S. Japanese dialects. InHoenigswald H. M. & Langacre R. H.(eds) *Diachronic, areal and typological linguistics* 368–400 (Mouton The Hague and Paris, 1973).
52. Peiros, I. Comparative linguistics in Southeast Asia. *Pacific Linguistics* **142** (1998).
53. Sanders, J. & Sanders, A. G. Dialect survey of the Kamasau language. *Pacific Linguistics. Series A. Occasional Papers* **56,** 137 (1980).
54. Cysouw, M., Wichmann, S. & Kamholz, D. A critique of the separation base method for genealogical subgrouping. *Journal of Quantitative Linguistics* **13,** 225–264 (2006).

Data Citations

1. Wichmann, S., Holman, E. W. & Brown, C. H. *The ASJP Database (version 17)* http://asjp.clld.org/static/listss17.zip (2016).
2. Hammarström, H., Forkel, R. & Haspelmath, M. *Zenodo* https://doi.org/10.5281/zenodo.1321024 (2018).
3. Jäger, G. *Open Science Framework* https://doi.org/10.17605/OSF.IO/CUFV7 (2018).

Acknowledgements

This research was supported by the ERC Advanced Grant 324246 EVOLAEMP and the DFG-KFG 2237 *Words, Bones, Genes, Tools*.

Author Contributions

G.J. carried out the analyses and wrote up the paper.

Additional Information

Table 3 is only available in the online version of this paper.

Competing interests: The author declares no competing interests.

How to cite this article: Jäger, G. Global-scale phylogenetic linguistic inference from lexical resources. *Sci. Data*. 5:180189 doi: 10.1038/sdata.2018.189 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.