Factoring lexical and phonetic phylogenetic characters from word lists

Gerhard Jäger
Institute of Linguistics
University of Tübingen
Wilhelmstr. 19, 72074 Tübingen, Germany
Email: gerhard.jaeger@uni-tuebingen.de

Johann-Mattis List
CRLAO / Team AIRE
EHESS / UPMC
2 Rue de Lille, 75007 Paris, France
Email: mattis.list@lingpy.org

***Abstract*—Computational historical linguistics is a young and new field. Among it's major challenge is the collection and preparation of suitable data resources. Here we present an approach that takes lexical data taken from a large collection of publicly available wordlists as input and infers automatic assessments regarding the cognacy of words and sounds. We illustrate the workflow and test it by comparing the results obtained from the computation of Maximum Likelihood trees with those provided by experts. The results show that our workflow still lags behind simpler approaches which analyze the data within a distance-based framework. However, since distance-based analyses bear a *blackbox* character, not allowing for a rigorous check of the individual decisions which lead to a certain classification proposal, we think that our experiments are an important contribution towards the establishment of more transparent methods in quantitative historical linguistics.**

I. Introduction

Computational historical linguistics is a still very young but thriving new field. One of the major challenges it currently faces is the collection and preparation of suitable data resources. There is a plethora of sophisticated methods and techniques — often adapted from computational biology — allowing very detailed and fine-grained inferences about language change. Bayesian phylogenetic inference is a prime example. These methods, however, require data to be organized in *character matrices*, i.e. languages have to be categorized according to a collection of (historically informative) discrete features. High quality data of this type are currently only available for a small number of language families, such as Indo-European and Austronesian.

An alternative approach, currently pursued mostly in the connection with the *Automated Similarity Judgment Program* (ASJP) `http://asjp.clld.org/`, deploys pairwise sequence alignment and distance-based phylogenetic inference methods. These techniques are comparatively shallow and provide little information about the actual processes underlying the observed linguistic diversity. On the other hand, they can be used with fairly raw, un-processed data, such as the collection of over 6,000 phonetically transcribed 40-item Swadesh lists collected by the ASJP community. Therefore this approach has a much wider scope as suitable data are available for all extant language families.

In this paper we propose a workflow for bringing ASJP data into the character-matrix format, and we present results from a pertinent pilot study. For this we use the flat techniques for processing ASJP data being proposed in [1] as stepping stone to infer lexical and phonetic characters for a substantial portion of ASJP. The quality of the results is assessed by using the expert classification from Glottolog (`http://glottolog.org/`) as gold standard.

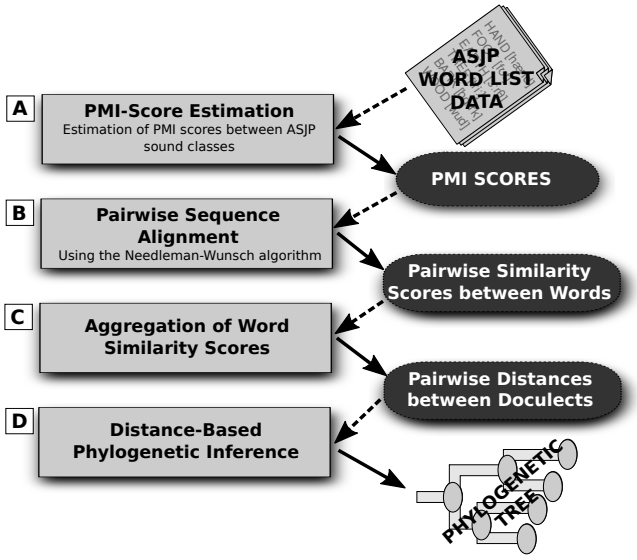

Fig. 1. Workflow of the PMI Analysis.

II. PMI Analysis of ASJP Data

In [1], [2] a collection of method for pairwise string alignment and the calculation of pairwise dissimilarities between ASJP word lists is proposed. These will be used as baseline in this paper. As the notion of *Pointwise Mutual Information* between sound occurrences play a central role in this approach, we will call the entire suite of methods "PMI analysis".

The workflow of the PMI analysis (as spelled out in [1]) can be described as shown in Figure 1. We deviate from [1] in three minor points: (a) we used the entire ASJP database for training PMI scores (while [1] only uses half of it to separate training and test data), (b) we transformed the distances produced in step C according to the function $f(x) = -\log(1-x)$ (monotonically mapping values in $(0, 1)$ to values in $(0, \infty)$), and (c) for step D we used an improved variant of the Neighbor Joining algorithm as implemented in the *fastme* software [3] (first computing the Neighbor Joining tree and

then performing a heuristic search minimizing the ordinary least square criterion via Nearest Neighbor Interchange).

This collection of methods is highly successful in identifying language families. For some language families, the family-internal structure inferred this way is also in very good agreement with the expert classification, while for other families, the fit is mediocre or even poor. (Uralic would be an example for an "easy" family whose internal structure is recovered almost perfectly by the PMI method, while Austronesian or Algic are "hard" families. Experience shows that hard families are also hard for other ASJP-based phylogenetic techniques, so the large variance between families is arguably due to the varying informativeness of the ASJP word lists rather than the algorithmic techniques, but this point still requires further investigation). In any event, to our knowledge the PMI methods suite is among the most successful approaches to infer phylogenies from ASJP data currently on the market. It essentially rests on the intuition that the closer two languages are related, the more similar are, on average, the wordforms these languages use for a given concept. The wordforms w_1, w_2 from languages L_1, L_2 (for a given concept) can be dissimilar for two reasons: (a) they are etymologically unrelated or (b) they are cognate but underwent sound changes. So the PMI approach implicitly captures information both regarding phonetic and lexical change. One of the objectives of the current study is to factorize these two sources of phylogenetic signals.

As will be spelled out below, both the pairwise sequence alignments from step B and the final tree computed in step D will be used for scaffolding the character-based methods to be developed.

III. Data

We selected the 6,080 doculects from ASJP that (a) are recent or went extinct after 1750, (b) represent neiter pidgins nor creoles nor artificial/fake languages, and (c) have entries for at least 28 of the 40 ASJP concepts. Those doculects were split into language families according to the Glottolog classification. Only language families containing between 10 and 70 doculects were considered, as (a) very small families provide too little phylogenetic information and (b) large families couldn't be analyzed adequately with the available hardware and software within a reasonable amount of time. The families *Ainu* (actually a language isolate that is, however, represented with 20 dialects in ASJP) and *Eastern Trans-Fly* are not internally structured according to the Glottolog information provided in the ASJP meta-data. As there is thus no expert phylogenetic information about those two families, they were excluded from analysis as well. This left us with 1,217 doculects from 48 families.

IV. Methods

Starting from ASJP word lists and the PMI analysis [A], we will perform automatic cognate detection [B], multiple sequence alignment within automatically detected cognate classes [C] and filter out those lexical and phonetic characters exhibiting a large amount of homoplasy [D].[1] This results in a character matrix for each language family considered here. This workflow is graphically depicted in Figure 2. The quality of these character matrices are evaluated by computing a phylogenetic tree via Maximum-Likelihood inference and comparing it to the Glottolog classification.

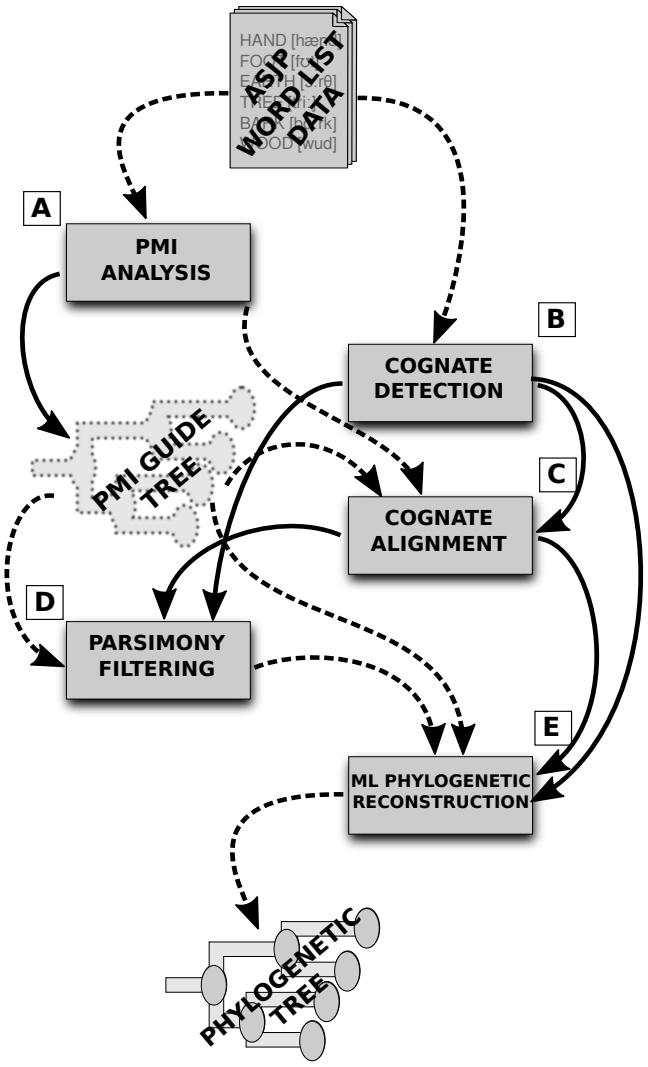

Fig. 2. Workflow for the Factoring of Lexical Data.

A. Cognate detection

The entries for the concept slots in the word lists for each language family were automatically clustered into *automatically detected sets of homologous words* (henceforth called ADH). This was done with help of the LexStat program [4], [5] provided as part of the LingPy software package [6] (`http://lingpy.org`). LexStat offers different algorithms for automatic cognate detection which vary in complexity. Since our datasets are small regarding the number of words per language, we chose the "SCA" method for cognate detection over the more accurate but also more time-consuming "LexStat" method which is more suitable to be applied to datasets containing more words per language. This method identifies potentially homologous words in two stages (see Fig. 3 AB): [A] *Distance Calculation*: In an initial stage, all

[1] The scripts and the data we used to run these analyses along with a detailed description of how to replicate the workflow are provided in the supplementary material accompanying this paper.

words in a given concept slot are compared using the Sound-Class Based Phonetic Alignment algorithm (SCA, [5], [8]). From the alignments, distance scores are retrieved using the formula proposed in [9]. [B] *Flat Clustering*: The distance scores are then used to partition the words into ADHs, using a flat version of the average linkage algorithm (UPGMA [10]) which terminates when clusters exceed the user-defined threshold of average distances. The result is a word list in which all words are assigned to a specific ADH (Fig. 3 [C]). LexStat as implemented in LingPy offers a certain range of alternative partitioning (flat cluster) algorithms, including *Markov Clustering* [16] which is very common in biology and seems to outperform alternative approaches, such as *k*-means [17] and *affinity propagation* [18], [19]. However, since the alternative cluster algorithms have not been intensively tested so far, we decided to stick to the defaults that LingPy provides, also for the sake of making the replication of our workflow easier for scholars who might be interested in taking them as a starting point for further analyses.

In all our experiments, we used a threshold of 0.45 which turned out to work fairly well in distinguishing cognate words from unrelated ones. In order to allow for further computation, the ADHs for each language family were then transformed into a *character matrix* (Fig. 3 [D]). In such a matrix, each row represents a doculect from a given language family and each column a *parsimony-informative* ADH, i.e., an ADH that includes or excludes at least one doculect. The cells indicate the presence or absence of all ADHs for a given doculect. Starting from these character matrices, we calculated the Maximum Likelihood phylogenetic tree for each language family. (We assumed constant rates and a molecular clock.)

B. Homoplasy detection

A substantial number of the ADHs display a high amount of homoplasy. This may be due to a variety of reasons. Lexical change sometimes does exhibit genuine parallel developments.[2] Additionally, there are also multiple sources of spurious homoplasies, i.e. homoplastic characters not corresponding to independent but parallel historical processes. Possible sources include (a) faulty cognate detection due to chance resemblances between non-cognate word forms, (b) borrowings, and (c) incomplete sampling in the compilation of the ASJP word lists. As such characters weaken the phylogenetic signal, we deployed a heuristics to detect and remove heavily homoplastic characters based on the *Maximum Parsimony* principle. Given a phylogenetic tree $\mathcal{T}$ and a character C with a known state at each leaf, the parsimony score $\mathrm{pars}_{\mathcal{T}}(C)$ of that character is the minimal number of mutations that has to be assumed this distribution of values if the the character evolved according to $\mathcal{T}$. It can efficiently be computed [12].

The maximal number of mutations for a given character would be achieved for a star-shaped tree where each leaf is an immediate daughter of the root. Then the most parsimonious reconstruction would reconstruct the most frequent state for the root and assume one mutation for each leaf not being in this state. Conversely, the minimal number of mutations for C would be achieved for a tree where each character state occurs only within a contiguous sub-region of the tree; it equals the number of different states minus 1.

The *Retention Index* (RI; cf. [13]) for a tree $\mathcal{T}$ and a character C is defined as

$$RI(C,\mathcal{T}) = \frac{\max_{\mathcal{T}'} \mathrm{pars}_{\mathcal{T}'}(C) - \mathrm{pars}_{\mathcal{T}}(C)}{\max_{\mathcal{T}'} \mathrm{pars}_{\mathcal{T}'}(C) - \min_{\mathcal{T}'} \mathrm{pars}_{\mathcal{T}'}(C)}.$$

(Recall that $\mathrm{pars}_{\mathcal{T}}(C)$ is character C's parsimony score relative to tree $\mathcal{T}$.) It measures how well $\mathcal{T}$ explains the distribution of states of C at the leafs. An *RI* of 0 is obtained for a tree requiring the maximal amount of homoplasy for C, while an *RI* of 1 means that $\mathcal{T}$ requires no homoplasy for C.

Using the PMI tree for each language family as reference tree, we calculated the *RI* for each character. Characters with an $RI < 0.4$ relative to the guide tree where excluded from further analysis.

An example of such a highly homoplastic character would be the following ADH (for the concept *path*):

Punjabi Majhi	`sarak`
Romanian 2	`cale`
Yiddish Eastern	`dErEX`

As these three doculects belong to different parts of the PMI tree, each instance of that class has to be explained by a separate mutation. Therefore the *RI* for this character is 0 and it is being filtered out (which is linguistically correct as those three words are in fact etymologically unrelated.[3]

C. Phonetic characters

Word lists contain information about (at least) two aspects of language change: (a) semantic change, especially the process where a language replaces a word form w_1 by a non-cognate word form w_2 to express a certain meaning, and (b) sound change, i.e. an individual segment within a word form is added, deleted or replaced by a different segment. Cognacy data only tap on the first type of information. To also utilize phonetic information for phylogenetic inference, we computed multiple sound alignments for each ADH that were not excluded during the previous step.

Each occurrence of a sound class in each column of these alignment blocks were treated as binary characters, and those were arranged in a binary character matrix. As with cognacy characters, we only included those characters that (a) included at least two doculects, (b) excluded at least two doculects, and (c) have a Retention Index > 0.4 relative to the PMI tree.

Here is an example for high homoplasy of phonetic characters. Consider the following multiple alignment (obtained using the T-Coffee algorithm; see below) for an ADH for *leaf* from various Indo-European languages:

[2]This is discussed at length in [11]; it is pointed out there that it is fairly common for cognate words in different lineages to independently undergo identical semantic shifts. One example mentioned there is a meaning change from *foot* to *leg*, which applied to descendants of Proto-Indo-European '*pod-' both in Modern Greek and in modern Indic and Iranian languages.

[3]The Majhi and Romanian words belong to different cognate classes (`http://ielex.mpi.nl/`), and the Yiddish word is a direct borrowing from Hebrew (Susan Rothstein, p.c.).

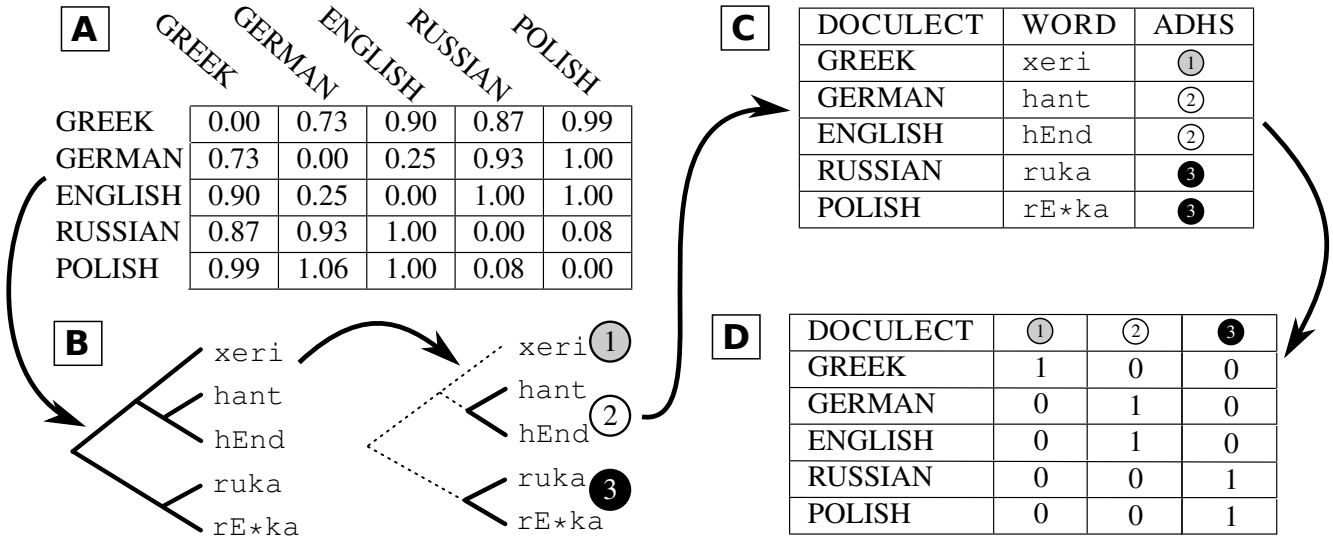

Fig. 3. Automatic Cognate Detection and Character Matrix Creation.

Irish Gaelic	di-L-og
Scottish Gaelic	tu-L-ak
Manx	do-lyak
Welsh	d3il-En
Breton	de-lien
Western Armenian	de-r-ev
Eastern Armenian	te-r-ev

This ADH is partially faulty as the Celtic words are not cognates of the Armenian words.[4] As the five Celtic doculects are monophyletic in the PMI tree, as are the two Armenian doculects, the parsimony score for this character is 2, so its Retention Index is 0.833 and it is not filtered out. However, the first column requires a parallel mutation from d to t both in Scottish Gaelic and Eastern Armenian. This leads to a Retention Index of 0 both for the binary characters corresponding to the d and the t in that column, so both characters are disregarded as being too homoplastic.

We compared two algorithms for automatic multiple sequence alignment: (a) Sound Class Based Phonetic Alignment (SCA, [5]), as provided by the LingPy software package [6], and a specific implementation of the T-Coffee algorithm which was designed in such a way that it can directly build on the inferences produced by the PMI analysis. In contrast to the common heuristic strategies for multiple sequence alignment (see [14]), the T-Coffee algorithm [15] uses a specific workflow to build individual *libraries* for each set of sequences in order to maximize the signal and their internal consistency [5].

T-Coffee alignment The problem of finding the optimal multiple alignment for k sequences, each of which is of length $\leq n$, is polynomial in n but exponential in k. So in the general case it is not possible to find the globally optimal solution. A simple but appealing polynomial-time approximation is *progressive alignment*. It requires a *guide tree*, i.e. a binary tree over the sequences in question which

[4]The two Armenian words belong to one cognate class and the five Celtic words to another (http://ielex.mpi.nl).

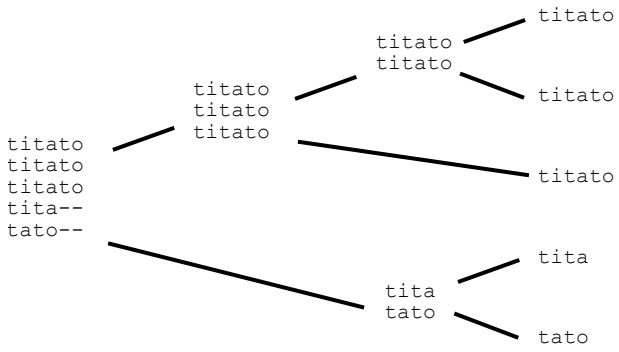

Fig. 4. Progressive Alignment: Multiple alignment is performed block-wise recursively from the tips to the root of a guide tree.

ideally captures the phylogenetic relationship between those sequences. Each tip is augmented with a sequence. The tree is then traversed tips-to-root, and at each non-terminal node the optimal alignment between the words or alignment blocks at its daughter nodes is computed. This can be achieved via a straightforward generalization of the Needleman-Wunsch algorithm [21] for pairwise alignments. At the root of the tree, all sequences are aligned in one block. The entire algorithm is cubic in the number of sequences, i.e. it is computationally tractable.

One drawback of progressive alignment is the fact that at each node, only information which is available at its daughters can be used. The point is illustrated with a schematic example in Figure 4. Suppose a proto-form titato is fully preserved in three taxa, while one taxon has elided the initial ti and another taxon the final to. If the latter two taxa are phylogenetic sisters, the optimal alignment at their mother node will come out as etymologically incorrect. This decision cannot be undone further up in the tree, no matter what additional information becomes available later on.

The T-Coffee algorithm uses heuristics to avoid this locality trap. In a first step, a library of all pairwise alignments is collected, and each pairwise alignment receives a weight

according to its quality (i.e. the number of matches and mismatches in it).[5] In a second step, an extended library of all compositions of pairwise alignments is collected. The weight of a composite alignment is the sum of the weights of its components. From this a score for each pair of symbol occurrences is derived, which is the sum of the weights of all composite alignments where those to occurrences are indirectly aligned. These scores are then used for performing progressive alignment. The entire workflow is schematically displayed in Figure 5. In the example, the highlighted composite alignment correctly connects the second `t` in `tita` with the first `t` in `tato`. If its weight is sufficiently high, it will enforce the correct alignment of `tita` and `tato` in during progressive alignment, leading the correct multiple alignment of all sequences.

SCA also allows for the use of consistency-based scoring in a T-Coffee-like framework [5, 108-114]. But while the multiple alignments produced by SCA are created from scratch, i. e., independently from further informations available about the datasets, the T-Coffee implementation was neatly integrated into the output of the PMI analysis, and both the pairwise alignments (which are fed to the T-Coffee algorithm to create an initial library) and the guide tree (which was used to successively add more and more sequences to a multiple sequence alignment) were based on the PMI analysis.

Using those character matrices, we used the software *Paup4* [22] to compute the Maximum Likelihood trees, (a) using only the phonetic characters (using constant rates and the molecular clock assumption) and (b) using both cognacy and phonetic characters simultaneously. In the latter case, mutation rates were assumed to be equal among the cognacy characters and among the phonetic characters, but possibly different for both classes. The separate mutation rates were estimated via Maximum Likelihood for the PMI tree and then kept constant during tree search.

V. Results

The resulting trees were evaluated by computing the *Generalized Quartet Distance (GQD)* [23] to the Glottolog tree for each family.[6] Directly comparing these numbers across families is problematic though, since — as mentioned above — the difficulty of the task of recovering the Glottolog classification from ASJP word lists varies heavily between families. We therefore normalized GQD values by subtracting the GQD value of the corresponding PMI tree (x in Table I).

Table I shows the results for different workflows (following the major steps as described in Figure 2).

As we can see from the table, the accuracy of our results increases along with the complexity of our workflow. At the lower end are the Maximum-Likelihood trees based on unfiltered ADH characters with an averaged normalized GQD of 3.82% and the unfiltered phonetic characters produced by SCA alignment, with an average GQD of 20.26%. At the upper end are the results obtained for the combination of filtered ADH characters and filtered phonetic characters, where phonetic alignments produced by the T-Coffee algorithm (1.65%) outperformed phonetic alignments produced by the SCA analysis (1.79%). Comparing the best results of our workflow (1.65%) with those obtained for the PMI analysis (x) further shows that our new workflow does not outperform the distance-based method, even though the remaining difference is small.

Workflow	SCA	T-Coffee
`[A]`	$15.28\%(=x)$	
`[B]->[E]`	$x+3.82\%$	
`[B]->[D]->[E]`	$x+3.42\%$	
`[C]->[E]`	$x+20.26\%$	$x+14.87\%$
`[C]->[D]->[E]`	$x+19.66\%$	$x+19.93\%$
`[B]->[C]->[E]`	$x+3.00\%$	$x+2.66\%$
`[B]->[C]->[D]->[E]`	$x+1.79\%$	$x+1.65\%$

TABLE I. Quantitative evaluation

VI. Conclusion

We have implemented a flexible workflow to factor lexical and phonetic phylogenetic characters from word list data. As our results show, this workflow does not significantly improve the quality of the trees which were obtained using simpler distance-based methods applied to the ASJP data, and on average, the agreement between expert classifications and the trees inferred with help of our workflow is even slightly worse than the agreement between expert classification and the PMI analysis. In contrast to the blackbox character of distance-based analyses, however, our workflow is transparent and allows to track and trace every single decision that led to the classificatory outcome. Furthermore, we can build on the inferences produced in the several steps of our workflow and use them as starting point for interesting and valuable further investigations, be it the comparison of automatically achieved results with those achieved with help of the traditional comparative method, or the calculation of tendencies and rates of lexical and phonological change.

Supplementary Material

The scripts and the data we used to run these analyses along with a detailed description of how to replicate the workflow can be downloaded from `https://zenodo.org/record/31987`. In order to run the analyses, quite a few software packages need to be installed. In case you run into troubles, please don't hesitate to contact us.

Acknowledgments

This research was supported by the ERC grant AdG 324246 (`http://www.evolaemp.uni-tuebingen.de/`) and the DFG grant 261553824 (`http://gepris.dfg.de/gepris/projekt/261553824`). We thank the two anonymous reviewers and the ASJP team for their efforts to collect and prepare all the data.

References

[1] G. Jäger, "Phylogenetic inference from word lists using weighted alignment with empirically determined weights," *Language Dynamics and Change*, vol. 3, no. 2, pp. 245–291, 2013.

[5] In [15] both global and local pairwise alignments are collected in the library. Our implementation of PMI-basd T-Coffee only uses global alignments.

[6] The GQD of an automatically generated tree to an expert tree is the percentage of resolved quartets in the expert tree that have the same topology in the automatically generated tree. It can be interpreted as "100% - recall". It is not possible to assess the precision analogously because for many quartets of doculects, the expert tree does not provide a resolved topology.

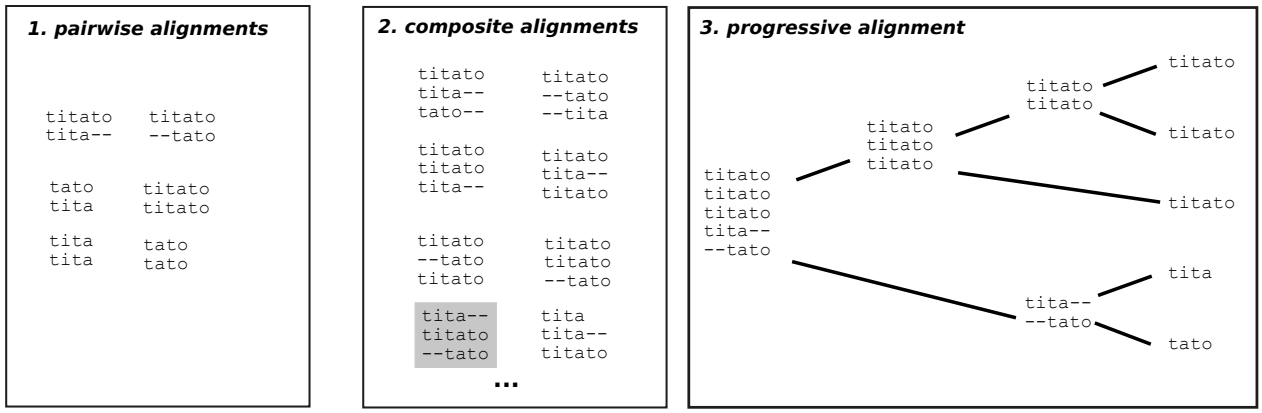

Fig. 5. The T-Coffee algorithm: Progressive alignment operates on scores derived from all composite pairwise alignments.

[2] ——, "Support for linguistic macrofamilies from weighted sequence alignment," *Proccedings of the National Academy of Sciences*, vol. 112, no. 41, 2015, doi: 10.1073/pnas.1500331112.

[3] R. Desper and O. Gascuel, "Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle," *Journal of computational biology*, vol. 9, no. 5, pp. 687–705, 2002.

[4] J.-M. List, "LexStat. Automatic detection of cognates in multilingual wordlists," *Proceedings of the EACL 2012 Joint Workshop of Visualization of Linguistic Patterns and Uncovering Language History from Multilingual Resources.* "LINGVIS & UNCLH 2012" (Avignon, 04/2304/24/2012), pp. 117–125, 2012.

[5] J.-M. List, "Sequence comparison in historical linguistics," Düsseldorf: Düsseldorf University Press, 2014.

[6] J.-M. List, S. Moran, "An open source toolkit for quantitative historical linguistics," *Proceedings of the ACL 2013 System Demonstrations*, "ACL 51" (Sofia, 08/04-08/09/2013) Association for Computational Linguistics, pp. 13-18, 2013.

[7] J.-M. List, "Multiple sequence alignment in historical linguistics. A sound class based approach," *Proceedings of ConSOLE XIX* "The 19th Conference of the Student Organization of Linguistics in Europe" (Groningen, 01/0501/08/2011). ed. by E. Boone, K. Linke, and M. Schulpen, pp. 241–260, 2012.

[8] J.-M. List, "SCA. Phonetic alignment based on sound classes," *New directions in logic, language, and computation*, ed. by M. Slavkovik and D. Lassiter. Berlin and Heidelberg: Springer, pp. 3251, 2012.

[9] S. S. Downey, B. Hallmark, M. P. Cox, P. Norquest, and S. Lansing, "Computational feature-sensitive reconstruction of language relationships: developing the ALINE distance for comparative historical linguistic reconstruction". *Journal of Quantitative Linguistics*, vol. 15. no. 4, pp. 340369, 2008.

[10] R. R. Sokal, C. D. Michener, "A statistical method for evaluating systematic relationships". *University of Kansas Scientific Bulletin*, vol. 28, pp. 1409–1438, 1958.

[11] W. Chang, C. Cathcart, D. Hall, and A. Garrett, "Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis," *Language*, vol. 91, no. 1, pp. 194–244, 2015.

[12] D. Sankoff, "Minimal mutation trees of sequences," *SIAM Journal on Applied Mathematics*, vol. 28, no. 1, pp. 35–42, 1975.

[13] J. S. Farris, "The retention index and homoplasy excess," *Systematic Biology*, vol. 38, no. 4, pp. 406–407, 1989.

[14] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchinson, "Biological sequence analysis. Probabilistic models of proteins and nucleic acids," 7th ed., Cambridge: Cambridge University Press, 2002[1998].

[15] C. Notredame, D. G. Higgins, J. Heringa, "T-Coffee. A novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, pp. 205217, 2000.

[16] S. M. van Dongen, "Graph clustering by flow simulation", PhD thesis, University of Utrecht, 2000.

[17] J. MacQueen, "Some methods for classification and analysis of multivariate observations", *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, pp. 281–297, 1967.

[18] B. J. Frey, D. Dueck, "Clustering by passing messages between data points", *Science*, vol. 315, pp. 973–976, 2007.

[19] J. Vlasblom, S. J. Wodak, "Markov clustering versus affinity propagation for the partitioning of protein interaction graphcs", *BMC Bioinformatics*, vol. 10, no. 99, 2009.

[20] J. D. Thompson, D. G. Higgins, T. J. Gibson, "CLUSTAL W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 46734680, 1994.

[21] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, p. 443453, 1970.

[22] D. Swofford, *Phylogenetic analysis using parsimony (* and other methods)*. Sunderland, MA: Sinauer Associates, 2002.

[23] S. Pompei, V. Loreto, and F. Tria, "On the accuracy of language trees," *PLoS One*, vol. 6, no. 6, p. e20109, 2011.