Please cite this draft/pre-print work as:

Tresoldi, Tiago; Rzymski, Christoph; Forkel, Robert; Greenhill, Simon J.; List, Johann-Mattis; and Gray, Russell D. (2019) “Managing historical linguistic data for computational phylogenetics and computer-assisted language comparison (PRE-PRINT)”. Jena: Max-Planck-Institute for the Science of Human History.

Managing historical linguistic data for computational phylogenetics and computer-assisted language comparison

Tresoldi, Tiago
Rzymski, Christoph
Forkel, Robert
Greenhill, Simon J.
List, Johann-Mattis
Gray, Russell D.

August 15, 2019

Table of Contents

Introduction

Computational phylogenetics is a relatively recent branch of historical linguistics that uses quantitative techniques to investigate the history of related languages. As the classical comparative method is less explicit on the techniques for constructing phylogenies of language families (see discussion in Jacques and List 2019), such a new approach can complement traditional techniques for sub-grouping based on shared innovations (Ross and Durie 1996).

The popularisation of computer-based methods has led to a greater awareness of issues resulting from limited data sustainability and proper data management (see, in particular, Chapter 5 and the User-Case on data management for historical linguistics). As linguistic data compiled for purposes other than phylogenetic reconstruction might be difficult to adapt to the needs of such analyses, we find an increasing amount of attempts to prepare the original data in ways amenable to qualitative inspection and quantitative investigations. However, since the practice of data

preparation has not been standardized so far, scholars employ a variety of custom formats as the backbone of their phylogenetic analyses. Such formats range from inadequate codings in which connections to the original sources have been lost, up to very detailed and complex formats that can only be processed by specific programs, which may at times not be publicly available. As a result, it is very difficult for newcomers to find good instructions on data handling and conversion. Additionally, data reuse is hampered because crucial information on the sources, the languages under investigation, or questionnaires used as basis for word comparisons are usually not supplied in standardized form.

Ideally, all linguistic data should be "FAIR" in the sense of Wilkinson et al. (2016): Findable, Accessible, Interoperable, and Reusable. FAIR not only implies that studies should be maximally reproducible, starting from the initial design of a projects (c.f. Berez-Kroeker et al. 2018), but also that a specific attention to "fairness" during all intermediate stages for preparing, curating, and transforming the data is needed. Instead of enumerating the many possibilities to code and use linguistic data to conduct a phylogenetic analysis, we decided to illustrate our suggestions for phylogenetic data management in a workflow based on a concrete analysis. We illustrate how we suggest data should be managed with the help of a published dataset, exploring the information, file formats, processes, and software involved, explaining and showing how to collect and store cross-linguistic information, how to guarantee that datasets are cross-linguistically comparable, how to store intermediate and final results of the analyses, and how to share data in a reusable form. While phylogenetic methods are not restricted to lexical data, the use of cognate sets (i.e., sets of related words identified by the comparative method or computer-assisted approaches) has become a quasi-standard in the discipline and will be the only method explored here (for alternative proposals using various types of structural features, see Macklin-Cordes and Round 2015; Greenhill et al. 2017, Ringe et al. 2002, Longobardi et al. 2015).

Our analysis uses the dataset of Lieberherr and Bodt (2017), which the authors made publicly available, consisting of lexical entries for 100 concepts, derived from the concept lists of Haspelmath and Tadmor (2009) and Swadesh (1971), and translated into 22 "highly divergent, endangered, and poorly described" languages of the Kho-Bwa subgroup of the Sino-Tibetan language family, of which we selected 20 varieties, which were all based on the authors' field notes and reflect a unified source. The study is accompanied by a tutorial which conveniently mirrors the sections and tasks presented, allowing readers to experiment with the dataset — or their own data — by following our instructions step-by-step.

Phylogenetic Data Life-Cycle

The initial stage of a computational phylogenetic study requires acquiring and converting digital sources to machine-readable format, which is in most cases a tabular word list (Stage 1). The second stage involves adding cognate judgments to the word list, which can be done *manually*, relying on experts or on information from the literature, *automatically*, by relying on software for automated cognate detection, or *semi-automatically*, by checking automatically inferred cognates (List 2016, Stage 2). Once these data are available, we can carry out the actual phylogenetic analysis. The investigation should start with exploratory data analysis (Morrison 2014, Stage 3) to visualise signal in the data by, for example, producing a Neighbor-Net or splits graph (a network convenient for inspecting the major patterns in the data, Bryant & Moulton 2003, Huson and Bryant 2006), or calculating various summary statistics that quantify signal and noise in the dataset, such as Consistency and Retention Indexes (Farris 1989), delta-scores and Q-residuals (Holland et al. 2002, Gray et al. 2010), also making sure that there are enough common data points among the languages (List et al. 2018b). Following this step, a detailed phylogenetic analysis using a range of different methods can be performed. Currently, the best-performing methods are based on Bayesian models that can provide a dated and rooted phylogeny (Stage 4). Independent of the stage of the analysis,

we recommend that scholars publish their data in a FAIR form, allowing colleagues to review and reuse them (Stage 5).

Data collection (Stage 1)

Before we can make phylogenetic analyses, the data has to be assembled, which can be done in multiple ways, including original field work, corpus analyses of texts (both modern and ancient), or consulting dictionaries, word lists, or glossaries. Once we have identified the sources that can deliver the data, we need to extract it and store it in a format convenient to access with software. In the following section, we will introduce the very general abstract data model we recommend to authors and give concrete recommendations on data storing and curation.

General remarks on data management

The data model that many linguists still use was popularized by Morris Swadesh, the pioneer in the large-scale collection of word lists in form of tabular data for quantitative analyses (Swadesh 1952). The crucial aspect of this data model is the semantic alignment of information, starting from a list of non-cultural concepts, at times expanded and modified, which was successively translated into the target languages of various studies. Linguists often think of the multilingual word lists produced by this procedure as a simple table, in which the rows refer to the concept labels (or elicitation glosses) and the columns capture the lexical entries in the sampled languages. This format has many plain advantages for non-computational usage. It is simple, easy to inspect, and easy to produce, and tables can be edited with common text processing or spreadsheet software. In fact, Lieberherr and Bodt originally provided their data in this form. Table 1 provides a small sample of these data in multilingual word list form.

Concept	Dikhyang	Wangho	Bulu	Rawa	Saria
"big"	əpõː	ebo^{u}	araː	arai	toʔrɨ̈ɨ̈
"bird"	fuə	fua	pəduː	pədoː	pədoː
"blood"	əfuɛ	efua	ahui	fui	hue

Table 1: Sample word list from the Kho-Bwa dataset, showing words glossed as "big", "bird" and "blood" for different language varieties, in the traditional wordlist form.

The simplicity of multilingual word list data provided in this form, however, is apparent and restricted to lexicographic entries, creating multiple complications once scholars include other information besides the translations for elicitation glosses across languages. What should one do, for example, if unable to decide for one of several alternatives to translate a concept? Should one list the synonyms separated by a comma, a slash, a dash, or even a vertical pipe ("|"), as in many existing datasets? Or should one get rid of synonyms, either following Swadesh practice of selecting the most common form (mostly decided in terms of perceived frequency of usage, see Swadesh 1955:4.5) or Gudschinsky's (1956: 179) advice of "flipping a coin"? Likewise, there is no consensus on how to annotate specific entries to include information such as cognacy. The most common solution is to add an extra cognacy column to the right of the one devoted to each language variety, as in the STARLING software package (Starostin 2000) and as in the data first provided by the authors of our dataset (as illustrated in Table 2).

Concept	Dikhyang	Cog	Wangho	Cog	Bulu	Cog	Rawa	Cog	Saria	Cog
"big"	əpõː	1	ebo^{u}	2	araː	2	arai	2	toʔrɨ̈ɨ̈	3
"bird"	fuə	3	fua	5	pəduː	5	pədoː	5	pədoː	5
"blood"	əfuɛ	6	efua	6	ahui	6	fui	6	hue	6

Table 2: Sample word list from the Kho-Bwa dataset, derived from Table 1, with cognate judgments added in extra columns labeled “Cog”.

A better strategy is to follow the insights of relational databases (Codd 1970), while adopting long-table formats (Forkel et al. 2018, List et al. 2018b). In this data structure, we give each cell containing a word form in Table 1 a row for itself. Table 3 provides an example corresponding to the data from Table 2. The first column of the long table is an identifier (usually numerical), and the consecutive columns define the different aspects of the word under question, e.g., its language, its pronunciation, its concept, and also its cognate identifier. Although it may look redundant on first sight, this format has many advantages. We can display synonyms without separating the content in a cell (by adding an alternative entry for a given concept as an extra row of our table), we can easily annotate cognates, and we could even append arbitrary information by simply adding a new column.

ID	Language	Concept	Entry	Cogset
1	Dikhyang	BIG	əpõː	BIG-1
2	Wangho	BIG	ebo^{u}	BIG-1
3	Bulu	BIG	araː	BIG-2
4	Rawa	BIG	arai	BIG-2
5	Saria	BIG	toʔrɨ̈ɨ̈	BIG-3
6	Dikhyang	BIRD	fuə	BIRD-1
7	Wangho	BIRD	fua	BIRD-1
8	Bulu	BIRD	pəduː	BIRD-2
9	Rawa	BIRD	pədoː	BIRD-2
10	Saria	BIRD	pədoː	BIRD-2
11	Dikhyang	BLOOD	əfuɛ	BLOOD-1
12	Wangho	BLOOD	efua	BLOOD-1
13	Bulu	BLOOD	ahui	BLOOD-1
14	Rawa	BLOOD	fui	BLOOD-1
15	Saria	BLOOD	hue	BLOOD-1

Table 3: Sample word list from the Kho-Bwa dataset, as listed in Table 2, in long-form.

The Cross-Linguistic Data Formats Initiative

Since long tables are nothing more than tables, we can store them in the same format in which we would store “traditional” word list tables. To increase data comparability and FAIRness, however, it is worth using additional tables for adding other information about the entities in our data, especially in terms of reference catalogs that enormously facilitate dataset aggregation. For language identification, for example, it is useful to link each variety to its corresponding code in

Glottolog (https://glottolog.org, Hammarström et al. 2019). For comparative concepts, the Concepticon initiative (https://concepticon.clld.org, List et al. 2019a) offers identifiers for standardized concept sets. Linking our data to these two catalogs gains us useful information (e.g. geographic locations from Glottolog, semantic categories or frequencies of word use from Concepticon). A recent and complementary development is a reference catalog for converting phonetic transcriptions, the Cross-Linguistic Transcription Systems initiative (CLTS, https://clts.clld.org, Anderson et al. 2018; List et al. 2018a). CLTS enhances accessibility and interoperability by explicitly specifying the phonemes in each language in the dataset, a specification that directly facilitates approaches using automated sequence comparison or enhanced interfaces for cognate annotation (see Stage 2).

To standardize the representation of data for computational phylogenetics and historical language comparison, the Cross-Linguistic Data Formats initiative (CLDF, https://cldf.clld.org, Forkel et al. 2018) offers standard formats for different data types in historical linguistics and linguistic typology, including wordlists, structural data, dictionaries, and parallel texts. To render one's data in CLDF word list format, normal spreadsheet editors can be used, but the initiative also offers software solutions that facilitate conversion from other structured formats. CLDF encourages dataset maintainers to use the above reference catalogs and also offers tools to validate the content of a CLDF dataset. The formats are supported by some important software tools for computational phylogenetics, such as BEASTLing (Maurits et al. 2018) and LingPy (List et al. 2018d) and libraries for reading and writing CLDF data are available for the Python (pycldf, Forkel et al., 2019) and R (rcldf, https://github.com/SimonGreenhill/rcldf) programming languages. Given the increasing importance of CLDF as a standard for data storing and sharing, as well as the growing amount of early adopters who have used the framework for data sharing (Hill and List 2017, Kaiping and Klamer 2018, Sagart et al. 2019) or for data aggregation (List et al. 2018c), we recommend all those who are interested in computational phylogenetics applications to code their data in the formats of the CLDF initiative. Our supporting tutorial instructs how this can be done, explaining how a CLDF dataset can be created (Tutorial 2.1.1) and loaded with LingPy (Tutorial 2.1.2), and how existing datasets can be retrieved from on-line repositories (Tutorial 2.1.3).

Cognate Identification (Stage 2)

Information on the etymological relations between words in different languages is occasionally already available in the form of classical sources, such as etymological dictionaries or lexicostatistic datasets (see e.g., McElhanon 1967). However, the annotation of cognate words for phylogenetic investigations can still be tedious, in particular when working with tabular data that follows the "classical" model shown in Table 1. If sufficient information on the history of the languages under investigation is not available, scholars will have to apply the classical workflow of the comparative method to infer regular sound correspondences crucial for identifying cognate words. Automated methods for cognate identification (List 2014, Rama et al. 2018) and sound correspondence patterns (List 2019) may come in handy, specifically in a computer-assisted framework where the data is pre-processed by the software, and then thoroughly reviewed and corrected by experts. To annotate, correct, and modify cognate sets, we recommend the use of interfaces designed for these purposes (see, for example, the EDICTOR tool by List 2017, http://edictor.digling.org), since this may help to avoid errors when working with large datasets.

Our accompanying tutorial illustrates how software for automated sequence comparison may be used to align the data automatically (Tutorial 2.2.1), how cognates can be automatically inferred with different methods and evaluated against a gold standard (Tutorial 2.2.2), and how the data can be curated with help of light-weight web-based interfaces (Tutorial 2.2.3).

Exploratory data analysis (Stage 3)

Data prepared in CLDF is easily amenable to a range of phylogenetic analyses. First, it is easy to extract distances between languages by assuming that the more similar languages are, the more related they are. This is the fundamental assumption of the classical, and problematic, approach of lexicostatistics (Swadesh 1950, 1952). Using the same languages from the example tables above and the entire dataset, with 100 concepts, we get the following matrix of similarities.

	Dikhyang	Wangho	Bulu	Rawa	Saria
Dikhyang	0.00	0.07	0.53	0.54	0.53
Wangho	0.07	0.00	0.53	0.52	0.52
Bulu	0.53	0.53	0.00	0.24	0.31
Rawa	0.54	0.52	0.24	0.00	0.20
Saria	0.53	0.52	0.31	0.20	0.00

Table 4: Similarity matrix of a subset of Kho-Bwa languages. Language pairs with scores closer to 0.0 are more similar, scores closer to 1.0 are more dissimilar.

Similarity matrices can be converted without effort to a tree using algorithms like UPGMA or Neighbor-Joining (Saitou and Nei 1987), which mimic lexicostatistics (Figure 1.a). These algorithms are implemented, among others, in the LingPy library (List et al. 2018d) library (used in the tutorial) and in R's APE library (Paradis, Claude, and Strimmer 2004). We can also load distances into other statistical inference procedures like cluster analysis, as done in Lieberherr and Bodt (2017).

One common distance-based approach to data exploration in computational historical linguistics is building a Neighbor-Net network (Bryant and Moulton 2003; Huson and Bryant 2006). This visualization (Figure 1.b) constructs branches proportional to the amount of change between languages, while conflicting signals are represented by box-like structures. These networks provide a useful way of visualizing overlapping and conflicting signal, such as that caused by borrowing or dialect-chain processes (Heggarty, Maguire, and McMahon 2010; Gray, Bryant, and Greenhill 2010). These networks are constructed in the SplitsTree package (Huson and Bryant 2006), and we can easily convert the CLDF dataset into a format suitable for SplitsTree. Other exploratory approaches that can be used to quantify the signal and noise in a dataset are analyses through Consistency and Retention Indexes (Farris 1989), delta-scores and Q-residuals (Holland et al. 2002, Gray et al. 2010). Our accompanying tutorial illustrates how to perform these tasks (Tutorial 2.3).

Phylogenetic analysis (Stage 4)

After the simpler distance-based approaches for data exploration, it is common to perform more advanced analyses. Currently, the most powerful phylogenetic approach is a set of tools known collectively as Bayesian phylogenetic methods (Huelsenbeck et al. 2001). These methods build trees in a way that mimics that of the traditional linguistic comparative method, identifying where cognate sets are innovated and retained. Further, these tools model uncertainty and error in our estimated phylogenies such that we can measure support for different sub-grouping hypotheses. Greenhill and Gray (2009) provide a more detailed overview of how Bayesian approaches work. Bayesian phylogenetic packages like BEAST (Bouckaert et al. 2014) tend to require data in a

specific format called NEXUS (Maddison, Swofford, and Maddison 1997) which can be generated from word list or CLDF datasets with tools such as LingPy.

Here we analyze the Kho-Bwa dataset using a Bayesian phylogenetic approach implemented in BEAST2 (Bouckaert et al. 2014, v2.5.1). We use a binary covarion model (Penny et al. 2001) that allows cognate sets to be gained and lost at different rates over time. We implemented a relaxed-clock model (Drummond et al. 2006), which allows each branch to change at a different rate, and this distribution of rates to be estimated from the data. The results are shown in Figure 1.c. The study indicates that all three methods show strong similarities in their overall sub-grouping and are consistent with the results presented in Lieberherr and Bodt (2017) based on hierarchical clustering. All methods split the family into three major branches: (a) the Western Kho-Bwa (Duhumbi, Khispi, Shergaon, Rupa, Jerigaon, Khoina, Rahung, Khoitam), (b) Bugun (Bichom, Singchung, Dikhyang, Wangho, Kaspi, Namphri), and (c) Puroik (Bulu, Rawa, Kojo Rojo, Sario Saria, Lasumpatte, Chayangtajo). Within these branches, the patterning is similar to that presented in Lieberherr and Bodt (2017), despite some notable differences that in most analyses are reported to the experts for investigation. Among the benefits of Bayesian approaches is the fact that we could further model variation in rate change for testing hypotheses on the evolution, which can also be reported to the experts. The discussion on Bayesian analyses goes beyond the purposes of data management of this user-case, but our tutorial shows how to prepare data for BEAST2 (Tutorial 2.4).

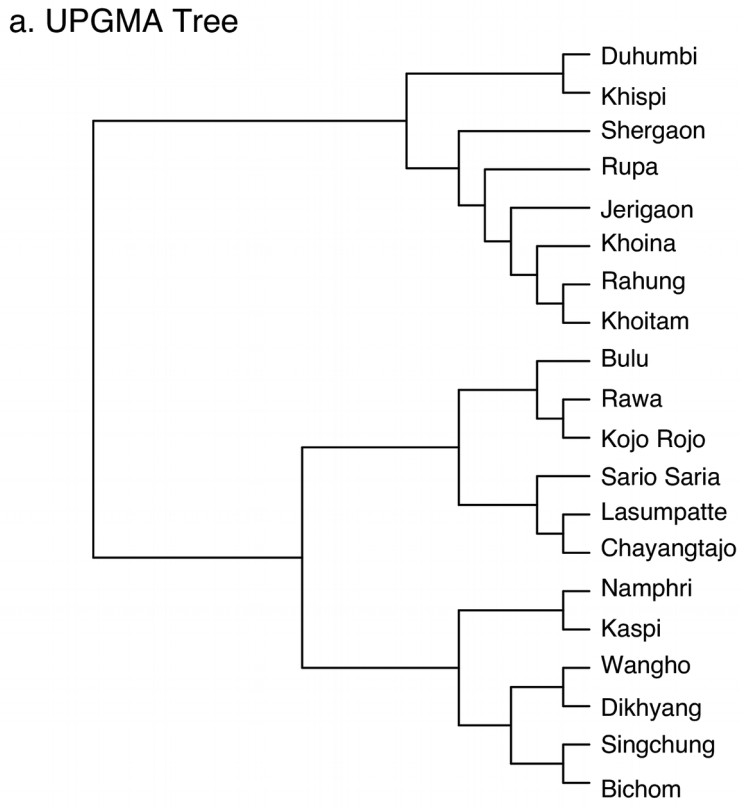

Figure 1.a: Phylogenetic visualisation of the Kho-Bwa dataset, with an UPGMA tree mimicking lexicostatistics.

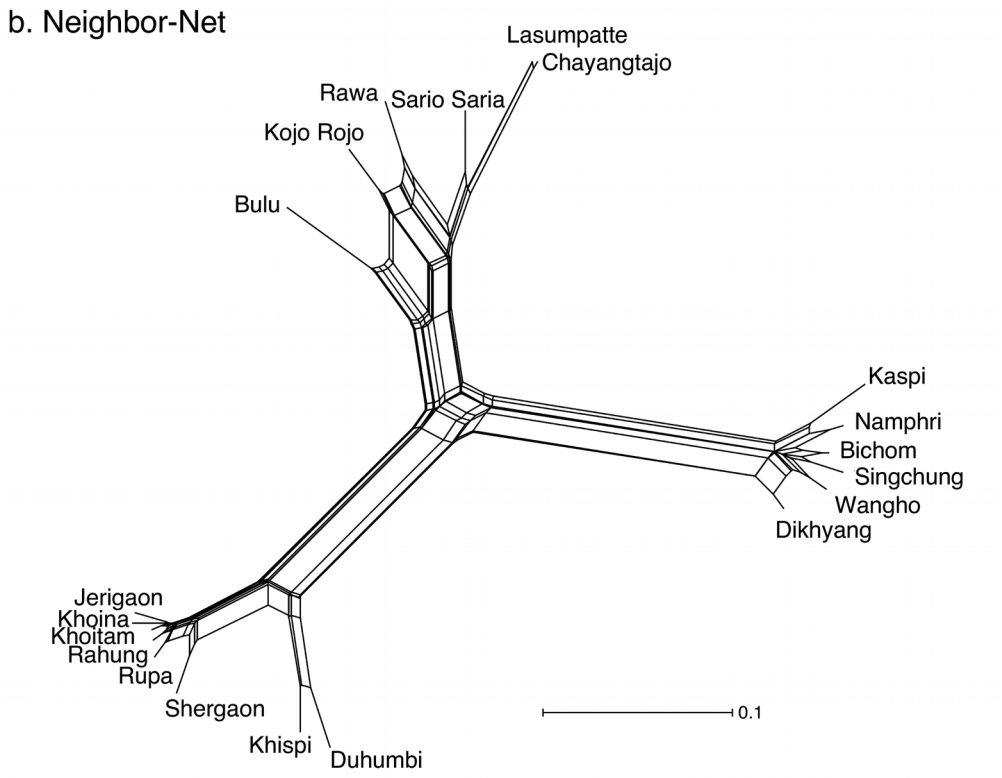

Figure 1.b: Phylogenetic visualisation of the Kho-Bwa dataset, with a Neighbor-Net network visualisation.

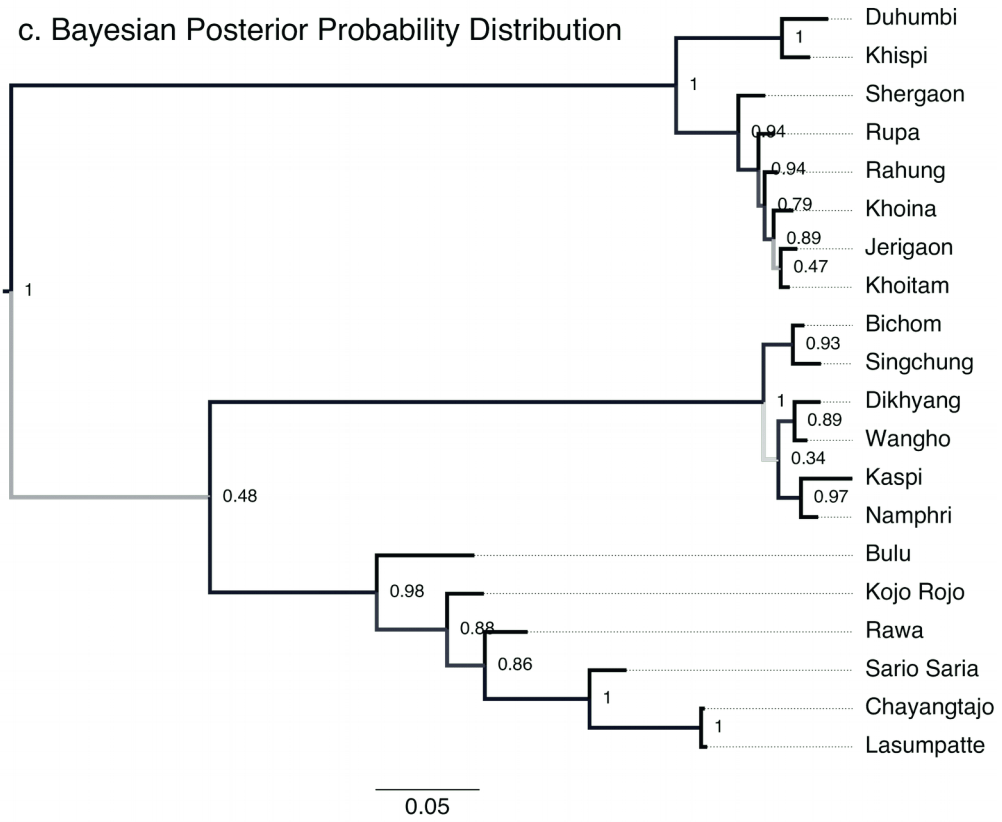

Figure 1.c: Phylogenetic visualisation of the Kho-Bwa dataset, with a maximum clade credibility tree of the posterior probability distribution from a Bayesian phylogenetic analysis.

The availability of a dataset collected and published in a long-form table, and converted to CLDF with ease, allowed us to apply different methods of investigation to support or disprove hypotheses of the original work. The analysis tried to emphasize how rewarding an adequate management of phylogenetic data can be in scientific terms. Researchers benefit from it not only by saving the time usually spent in data collection and preparation, but also because of the facilitated collaboration and the suggestions of future work offered by the results. In specific, we not only have quantitative bases on which questions should be investigated next, such as the placement of the Bugun and Puruik clades in the tree, but anyone would be able to apply other quantitative methods, or to combine these data with different datasets for new research questions (for example, Sino-Tibetan collections offering additional data points in CLDF, as presented, e.g., in Sagart et al. 2019). In all cases, very desirable prospects in terms of a language group still poorly studied from a historical linguistic perspective.

Data sharing and deployment (Stage 5)

We encourage and practice data sharing, creating and maintaining re-usable data in linguistics (Berez-Kroeker et al. (2018)). The modular architecture of CLDF allows researchers to combine and mix, more or less freely, what might best fit their individual pipelines and requirements. The main idea of this pipeline is not to enforce any theoretical constraints, but to ensure that once a research project is finished, data and results will be findable and accessible. For this reason, besides providing easily analyzable data, CLDF datasets were designed for convenience in share and deployment. While plain datasets can be shared with little effort on platforms like GitHub and Zenodo, the related CLLD project (Forkel et al. 2018b) allows to deploy data into browsable web applications, as showcased by different projects such as the study on colexification patterns CLICS2 (List et al. 2018c), the typological survey of the World Atlas of Language Structures (WALS, https://wals.clld.org, Dryer and Haspelmath 2013), and the study on horizontal lexical transfer by the World Loanword Database (https://wold.clld.org, Haspelmath and Tadmor 2009), among others. Our tutorial discusses how CLDF datasets can be shared and deployed (Tutorial 2.5).

Conclusion

Our plan with this user-case was to present principles of data management as applied to computational phylogenetics and computer-assisted language comparison, showcasing the solutions we recommend. We are confident that, no matter how it will evolve, historical linguistics will need and benefit from good practices in the representation and management of its data in order to advance. Methods, questions, and solutions come and go: interdisciplinarity will evolve from its current shape, concept lists are routinely expanded and reduced, cognate sets as basic characters of analysis might be supplemented or replaced by other data, Bayesian phylogenetic inference might lose its momentum and be replaced by new quantitative or symbolic models, and so on. The general principles of linguistic data management, and of phylogenetic data and CLDF in particular, acknowledge that such evolution is inevitable, and instruct us to prepare data for all the future manipulations that might be required.

Supplementary Material

The supplementary material can be downloaded from https://github.com/lexibank/phylogenetics-data-management-tutorial. It contains the accompanying tutorial, at https://github.com/lexibank/phylogenetics-data-management-tutorial/blob/master/Tutorial.md along with the data and the code needed to reproduce the analyses discussed in this study.

Acknowledgments

We thank Timotheus Bodt and Ismael Lieberherr, for providing help with the parsing of their data and the interpretation of their results; Gereon Kaiping for providing help with the development of interfaces between the LingPy software package and CLDF. This research would not have been possible without the generous support by many institutes and funding agencies. JML and TT were funded by the ERC Starting Grant 715618 Computer-Assisted Language Comparison (http://calc.digling.org). SJG was supported by the Australian Research Council's Discovery Projects funding scheme (project number DE 120101954) and the ARC Center of Excellence for the Dynamics of Language grant (CE140100041).

References

Anderson, Cormac, Tiago Tresoldi, Thiago C. Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. "A Cross-Linguistic Database of Phonetic Transcription Systems" in *Yearbook of the Poznań Linguistic Meeting* 4 (1): 21-53.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan S. Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. "Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field." *Linguistics* 56 (1): 1–18.

Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Suchard Marc A., Andrew Rambaut, and Alexei J. Drummond. 2014. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis." *PLoS Computational Biology* 10 (4): 1–6. https://doi.org/10.1371/journal.pcbi.1003537.

Bryant, David, and Vincent Moulton. 2003. "Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks." *Molecular Biology and Evolution* 21 (2).

Codd, Edgar F. 1970. "A Relational Model of Data for Large Shared Data Banks." *Commun. ACM* 13 (6): 377–87. https://doi.org/10.1145/362384.362685.

Drummond, Alexei J., Simon Y. W. Ho, Matthew J. Phillips, and Andrew Rambaut. 2006. "Relaxed Phylogenetics and Dating with Confidence." *PLOS Biology* 4 (5): e88.

Dryer, Matthew S., and Martin Haspelmath, eds. 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wals.info/.

Farris, J. S. 1989. "The Retention Index and the Rescaled Consistency Index". *Cladistics*, 5: 417-419. doi:10.1111/j.1096-0031.1989.tb00573.x

Forkel, Robert, Sebastian Bank, and Christoph Rzymski. 2018. "clld/clld: clld - a toolkit for cross-linguistic databases." https://doi.org/10.5281/zenodo.1436382.

Forkel, Robert, Sebastian Bank, Simon J Greenhill, Christoph Rzymski, & Gereon Kaiping. (2019, July 30). cldf/pycldf: pycldf (Version v1.6.4). Zenodo. http://doi.org/10.5281/zenodo.3355430

Forkel, Robert, List Johann-Mattis, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. "Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics." *Scientific Data*.

Gray, Russell D., David Bryant, and Simon J. Greenhill. 2010. “On the Shape and Fabric of Human History.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365 (1559): 3923–33. https://doi.org/10.1098/rstb.2010.0162.

Greenhill, Simon J., and Russell D. Gray. 2009. “Austronesian Language Phylogenies: Myths and Misconceptions About Bayesian Computational Methods.” In *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, edited by K A Adelaar and A Pawley, 375–97. Canberra: Pacific Linguistics.

Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson, and Russell D. Gray. 2017. “Evolutionary dynamics of language systems.” *Proceedings of the National Academy of Sciences*, 201700388. https://doi.org/10.1073/pnas.1700388114.

Gudschinsky, Sarah C. 1956. “The ABC's of lexicostatistics (glottochronology)”. *Word* 12. 175-210.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath. 2019. “Glottolog 4.0.” Jena: Max Planck Institute for the Science of Human History. https://glottolog.org/ accessed 2019-07-30.

Haspelmath, Martin, and Uri Tadmor, eds. 2009. *WOLD*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wold.clld.org/.

Heggarty, Paul, Warren Maguire, and April McMahon. 2010. “Splits or Waves? Trees or Webs? How Divergence Measures and Network Analysis Can Unravel Language Histories.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences.* 365: 3829–43. https://doi.org/10.1098/rstb.2010.0099.

Hill, Nathan W., and Johann-Mattis List. 2017. “Challenges of Annotation and Analysis in Computer-Assisted Language Comparison: A Case Study on Burmish Languages.” *Yearbook of the Poznań Linguistic Meeting* 3 (1): 47–76. https://www.degruyter.com/view/j/yplm.2017.3.issue-1/yplm-2017-0003/yplm-2017-0003.xml.

Holland, Barbara R., Katharina T. Huber, Andreas Dress, and Vincent Moulton. 2002. “δ-plots: A tool for analyzing phylogenetic distance data”. *Molecular Biology and Evolution* (19): 2051–2059.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. “Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology.” *Science* 294: 2310–4.

Huson, Daniel H., and David Bryant. 2006. “Application of Phylogenetic Networks in Evolutionary Studies.” *Molecular Biology and Evolution* 23 (2): 254–67.

Jacques, Guillaume and Johann-Mattis List. 2019. "Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them)". *Journal of Historical Linguistics* 9 (1): 128-166.

Kaiping, Gereon A., and Marian Klamer. 2018. “LexiRumah: An Online Lexical Database of the Lesser Sunda Islands.” *PLOS ONE* 13 (10): 1–29. https://doi.org/10.1371/journal.pone.0205250.

Lieberherr, Ismael, and Timotheus A. Bodt. 2017. “Sub-Grouping Kho-Bwa on Shared Core Vocabulary.” *Himalayan Linguistics* 16 (2).

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press. http://sequencecomparison.github.io/.

———. 2016. "Computer-Assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics". Jena: Max Planck Institute for the Science of Human History.

———. 2017. “A Web-Based Interactive Tool for Creating, Inspecting, Editing, and Publishing Etymological Datasets.” In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, 9–12. Valencia: Association for Computational Linguistics. http://edictor.digling.org.

———. 2019. “Automatic Inference of Sound Correspondence Patterns Across Multiple Languages.” *Computational Linguistics* 45 (1): 1–24. https://doi.org/10.1101/434621.

List, Johann-Mattis, Cormac Anderson, Tiago Tresoldi, Simon J. Greenhill, Christoph Rzymski, and Robert Forkel. 2018a. “Cross-Linguistic Transcription Systems (Version V1.1.1).” Jena: Max Planck Institute for the Science of Human History. https://clts.clld.org/.

List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi, and Robert Forkel. 2018b. "Sequence comparison in computational historical linguistics" in *Journal of Language Evolution* 3 (2): 130-144.

List, Johann-Mattis, Michael Cysouw, Simon J. Greenhill, and Robert Forkel. 2019. “Concepticon. A Resource for the Linking of Concept List. Version 2.1.” Jena: Max Planck Institute for the Science of Human History. http://concepticon.clld.org/.

List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel 2018c. *CLICS²*. Jena: Max Planck Institute for the Science of Human History. https://clics.clld.org/.

List, Johann-Mattis, Simon J. Greenhill, Tiago Tresoldi and Robert Forkel. 2018d. “LingPy. A Python Library for Quantitative Tasks in Historical Linguistics. Version 2.6.4” Jena: Max Planck Institute for the Science of Human History. http://lingpy.org.

Longobardi, Giuseppe, Silvia Ghirotto, Cristina Guardiano, Francesca Tassi, Andrea Benazzo, Andrea Ceolin and Guido Barbujan. 2015. "Across language families: Genome diversity mirrors linguistic variation within Europe" in *American Journal of Physical Anthropology* 157 (4): 630-640.

Macklin-Cordes, Jayden L., and Erich R. Round. 2015. “High-Definition Phonotactics Reflect Linguistic Pasts.” In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*, edited by Johannes Wahle, Marisa Köllner, Harald Baayen, Gerhard Jäger, and Tineke Baayen-Oudshoorn. Tübingen: University of Tübingen. https://doi.org/10.15496/publikation-8609.

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.

Maurits, Luke, Robert Forkel, Gereon A. Kaiping, Quentin D. Atkinson. 2017. “BEASTling: A software tool for linguistic phylogenetics using BEAST 2”. *PLoS ONE* 12 (8): e0180908. https://doi.org/10.1371/journal.pone.0180908

McElhanon, Kenneth A. 1967. “Preliminary Observations on Huon Peninsula Languages”. *Oceanic Linguistics*. 6, 1-45.

Morrison, David A. 2014. "Is the Tree of Life the best metaphor, model, or heuristic for phylogenetics?" in *Systematic Biology* 63 (4): 628-638.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. “APE: Analyses of Phylogenetics and Evolution in R Language.” *Bioinformatics* 20: 289–90.

Penny, David, Bennet J. McComish, Michael A. Charleston, and Michael D. Hendy. 2001. “Mathematical Elegance with Biochemical Realism: The Covarion Model of Molecular Evolution.” *Journal of Molecular Evolution* 53 (6): 711–23.

Rama, Taraka, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. “Are Automatic Methods for Cognate Detection Good Enough for Phylogenetic Reconstruction in Historical Linguistics?” In *Proceedings of the North American Chapter of the Association of Computational Linguistics*, 393–400. https://aclanthology.coli.uni-saarland.de/papers/N18-2063/n18-2063.

Ringe, Donald, Tandy Warnow and Ann Taylow. 2002. "Indo-European and computational cladistics" in *Transactions of the Philological Society* (100) 1: 59-129.

Ross, Malcom and Mark Durie. 1996. "Introduction" in *The Comparative Method reviewed: regularity and irregularity in sound change* (ed. Mark Durie). New York: Oxford University Press.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, Johann-Mattis List. 2019. "Dated language phylogenies shed light on the ancestry of Sino-Tibetan" in *Proceedings of the National Academy of Science of the United States of America* 166: 10317–10322.

Saitou, Naruya, and Masatoshi Nei. 1987. “The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees.” *Molecular Biology and Evolution* 4 (4): 406–25. http://www.ncbi.nlm.nih.gov/pubmed/18343690.

Starostin, Sergej A. 2000. *The Starling Database Program*. Moscow: RGGU. http://starling.rinet.ru.

Steel, Mike A., Michael D. Hendy, and David Penny. 1988. “Loss of Information in Genetic Distances.” *Nature* 36: 118.

Swadesh, Morris. 1950. “Salish Internal Relationships.” *International Journal of American Linguistics* 16 (4): 157–67.

———. 1952. “Lexico-Statistic Dating of Prehistoric Ethnic Contacts.” *Proceedings of the American Philosophical Society* 96 (4): 452–63.

———. 1955. "Towards Greater Accuracy in Lexicostatistic Dating". *International Journal of American Linguistics,* 21(2), 121–137.

———. 1971. "The origin and diversification of language". Edited post mortem by Joel Sherzer. Chicago: Aldine.

Wilkinson, Mark D., Michel Dumontier, IJsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The Fair Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3.