28 Managing Historical Linguistic Data for Computational Phylogenetics and Computer-Assisted Language Comparison

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

1 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 & List 2019), such a new approach can complement traditional techniques for sub-grouping based on shared innovations (Ross & Durie 1996).

The popularization of computer-based methods has led to a greater awareness of issues resulting from limited data sustainability and proper data management (see, in particular, Mattern, chapter 5, this volume, for general discussion and Daniels & Daniels, chapter 26, this volume, for discussion of historical linguistic data). 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 number of attempts to prepare the original data in ways amenable to qualitative inspection and quantitative investigations. However, because 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 coding in which connections to the original sources have been lost, 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 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 a 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 project (cf. Berez-Kroeker et al. 2018), but also that specific attention to "fairness" during all intermediate stages of preparing, curating, and transforming the data is needed. Instead of enumerating the many possibilities of coding and using linguistic data to conduct phylogenetic analysis, we illustrate our suggestions for phylogenetic data management in a workflow based on a concrete study. We illustrate these suggestions with the help of a published data set, exploring the information, file formats, processes, and software involved, and explaining/demonstrating how to collect and store crosslinguistic information, how to guarantee that data sets are crosslinguistically 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 & Round 2015; Greenhill et al. 2017; Ringe, Warnow, & Taylor 2002; Longobardi et al. 2015).

Our analysis uses the data set of Lieberherr and Bodt (2017), which the authors made publicly available, consisting of lexical entries for a hundred concepts, derived from the concept lists of Haspelmath and Tadmor (2009) and Swadesh (1971), and translated into twenty-two "highly divergent, endangered, and poorly described" languages of the Kho-Bwa sub-group of the Sino-Tibetan language family. We then selected twenty varieties, which were all based on the authors' field notes and reflect a

unified source. The study is accompanied by a tutorial that conveniently mirrors the sections and tasks presented, allowing readers to experiment with the data set—or their own data—by following our instructions step-by-step.

2 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 (see stage 1 in section 2.1). 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 *semiautomatically*, by checking automatically inferred cognates (List 2016, see stage 2 in section 2.3). Once these data are available, we carry out the actual phylogenetic analysis. The investigation starts with exploratory data analysis (Morrison 2014, see stage 3 in section 2.4) to visualize the 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 1998), or calculating various summary statistics that quantify the signal and noise in the data set, such as consistency and retention indexes (Farris 1989), δ-scores, and Q-residuals (Holland et al. 2002; Gray et al. 2010). This also ensures that there are enough common data points among the languages (List, Walworth, et al. 2018). 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 (see stage 4 in section 2.5). 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 (see stage 5 in section 2.6).

2.1 Data collection (stage 1)

Before we can make phylogenetic analyses, the data have to be assembled, which can be done in multiple ways, including original fieldwork; 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 them and store them 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.

2.1.1 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 (2017) originally provided their data in this form. Table 28.1 provides a small sample of these data in multilingual word list forms.

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 data sets? Or should one get rid of synonyms, either following Swadesh's practice of selecting the most common form (mostly decided in terms of perceived frequency of usage; see Swadesh 1955:125–126) or Gudschinsky's (1956:179) advice of "flipping a coin"? Likewise, there is no consensus on how to annotate specific entries to

Table 28.1
Sample word list from the Kho-Bwa data set, showing words glossed as "big," "bird," and "blood" for different language varieties, in the traditional word list form

Concept	Dikhyang	Wangho	Rawa
"Big"	əpõ:	ebo^{u}	arai
"Bird"	fuə	fua	pədo:
"Blood"	əfuɛ	efua	fui

include information such as cognacy. The most common solution is to add an extra column storing information on cognacy to the right of the one devoted to each language variety, as in the STARLING software package (Starostin 2000) and as in the data provided by the authors of our data set, which is illustrated in table 28.2.

A better strategy is to follow the insights of relational databases (Codd 1970), while adopting long-table formats (Forkel et al. 2018; List, Walworth, et al. 2018). In this data structure, we give each cell containing a word form in table 28.1 its own row. Table 28.3 provides an example corresponding to the data from table 28.2. The first column of the long table is an identifier (usually a numerical identifier), and the consecutive columns define the different aspects of the word in question, for example, language, pronunciation, concept, and also cognate identifier. Although it may look redundant at 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 also easily annotate cognates and even append arbitrary information by simply adding a new column.

Table 28.2
Sample word list from the Kho-Bwa data set, derived from table 28.1, with cognate judgments added in extra columns labeled "Cog"

Concept	Dikhyang	Cog	Wangho	Cog	Rawa	Cog
"Big"	əpõː	1	ebo^{u}	1	arai	2
"Bird"	fuə	3	fua	4	pədoː	4
"Blood"	əfuɛ	5	efua	5	fui	5

Table 28.3
Sample word list from the Kho-Bwa data set, as listed in table 28.2, in long form

ID	Language	Concept	Entry	Cogset
1	Dikhyang	BIG	əpõː	BIG-1
2	Wangho	BIG	ebo^{u}	BIG-1
3	Rawa	BIG	arai	BIG-2
4	Dikhyang	BIRD	fuə	BIRD-1
5	Wangho	BIRD	fua	BIRD-1
6	Rawa	BIRD	pədoː	BIRD-2
7	Dikhyang	BLOOD	əfuɛ	BLOOD-1
8	Wangho	BLOOD	efua	BLOOD-1
9	Rawa	BLOOD	fui	BLOOD-1

2.1.2 The Cross-Linguistic Data Formats initiative

Because long tables are computationally speaking 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 facilitate data set 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. 2021). For comparative concepts, the Concepticon initiative (https://concepticon.clld.org; List, Rzymski, et al. 2021) offers identifiers for standardized concept sets. Linking our data to these two catalogs offers useful additional information (e.g., geographic locations from Glottolog, semantic categories or frequencies of word use from Concepticon). For the handling of the form part of the linguistic sign, the Cross-Linguistic Transcription Systems initiative increases the accessibility and interoperability of phonetic transcriptions by explicitly specifying which speech sounds are represented by which symbol combinations in the data. In this way, the specification greatly facilitates 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 word lists, 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 data set maintainers to use the above-mentioned reference catalogs and also offers tools to validate the content of a CLDF data set. The formats are supported by some important software tools for computational phylogenetics, such as BEASTling (Maurits et al. 2017) and LingPy (List & Forkel 2021) and libraries for reading and writing CLDF data are available for the Python (pycldf; Forkel, Bank, Greenhill, et al. 2021) and R (rcldf, https://github.com/SimonGreenhill/rcldf) programming languages. Additionally, with CLDFBench (Forkel and List 2020), a Python package is available that helps to automatize and customize the creation of data

sets in CLDF format. 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 & List 2017; Kaiping & Klamer 2018; Sagart et al. 2019; Wu et al. 2020) or for data aggregation (Rzymski et al. 2020), 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 data set can be created (tutorial 2.1.1; for all tutorials, see Supplementary Material) and loaded with LingPy (tutorial 2.1.2), and how existing data sets can be retrieved from online repositories (tutorial 2.1.3). Lieberherr and Bodt (2019) is the CLDF version of the original data set that we use in the subsequent analyses.

2.2 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 data sets (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 28.1. 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 are preprocessed 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, e.g., the EDICTOR tool by List 2017; https://digling.org/edictor), as this may help to avoid errors when working with large data sets.

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 the help of lightweight web-based interfaces (tutorial 2.2.3).

2.3 Exploratory data analysis (stage 3)

Data prepared in CLDF are 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 examples in tables 28.1–28.3 and the entire data set, with a hundred concepts, we get the matrix of similarities.

Similarity matrices, as in table 28.4, can be converted without effort to a tree using algorithms such as Unweighted Pair Group Method with Arithmetic Mean (UPGMA) or Neighbor-Joining (Saitou & Nei 1987), which mimic lexicostatistics (figure 28.1). These algorithms are implemented, among others, in the LingPy library (List & Forkel 2021), a library used in the tutorial and in R's APE (analyses of phylogenetics and evolution) library (Paradis, Claude, & Strimmer 2004). We can also load distances into other statistical inference procedures such as 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 & Moulton 2003; Huson 1998). This visualization (see figure 28.2) constructs branches proportional to the amount of change between languages where conflicting signals are represented by box-like structures. These networks provide a useful way of visualizing overlapping and conflicting signals, such as that caused by borrowing or dialect-chain processes (Heggarty, Maguire, & McMahon 2010; Gray et al. 2010). These networks are constructed in the SplitsTree package (Huson 1998), and we can easily convert the CLDF data set into a format suitable for SplitsTree. Other exploratory approaches that can be used to quantify the signal and noise in a data set are analyses through

Table 28.4
Similarity matrix of a subset of Kho-Bwa languages

	Dikhyang	Wangho	Rawa
Dikhyang	0.00	0.07	0.54
Wangho	0.07	0.00	0.52
Rawa	0.54	0.52	0.00

Language pairs with scores closer to 0.0 are more similar; scores closer to 1.0 are more dissimilar.

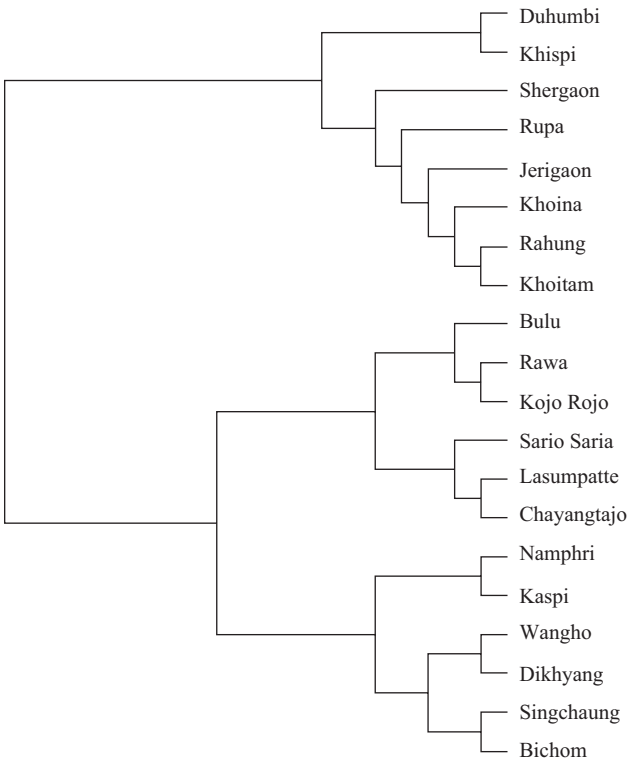


Figure 28.1
Phylogenetic visualization of the Kho-Bwa data set, with an UPGMA tree mimicking lexicostatistics.

consistency and retention indexes (Farris 1989), δ-scores, and Q-residuals (Holland et al. 2002; Gray et al. 2010). Our accompanying tutorial illustrates how to perform these tasks (tutorial 2.3).

2.4 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. Furthermore, 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 such as BEAST (Bayesian Evolutionary Analysis Sampling Trees; Bouckaert et al. 2014) tend to require data in a specific format called NEXUS (Maddison, Swofford, & Maddison 1997) that can be generated from word list or CLDF data sets with tools such as LingPy.

Here we analyze the Kho-Bwa data set using a Bayesian phylogenetic approach implemented in BEAST2 (Bouckaert et al. 2014, version 2.5.1). 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) that 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 figures 28.3 and 28.4. 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: (1) the Western Kho-Bwa (Duhumbi, Khispi, Shergaon, Rupa, Jerigaon, Khoina, Rahung, Khoitam), (2) Bugun (Bichom, Singchung, Dikhyang, Wangho, Kaspi, Namphri), and (3) 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).

The availability of a data set 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. More specifically, 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 also anyone is able to apply other quantitative methods or combine these data with different data sets for new research questions (for example, Sino-Tibetan collections offering additional data points in CLDF, as presented, e.g.,

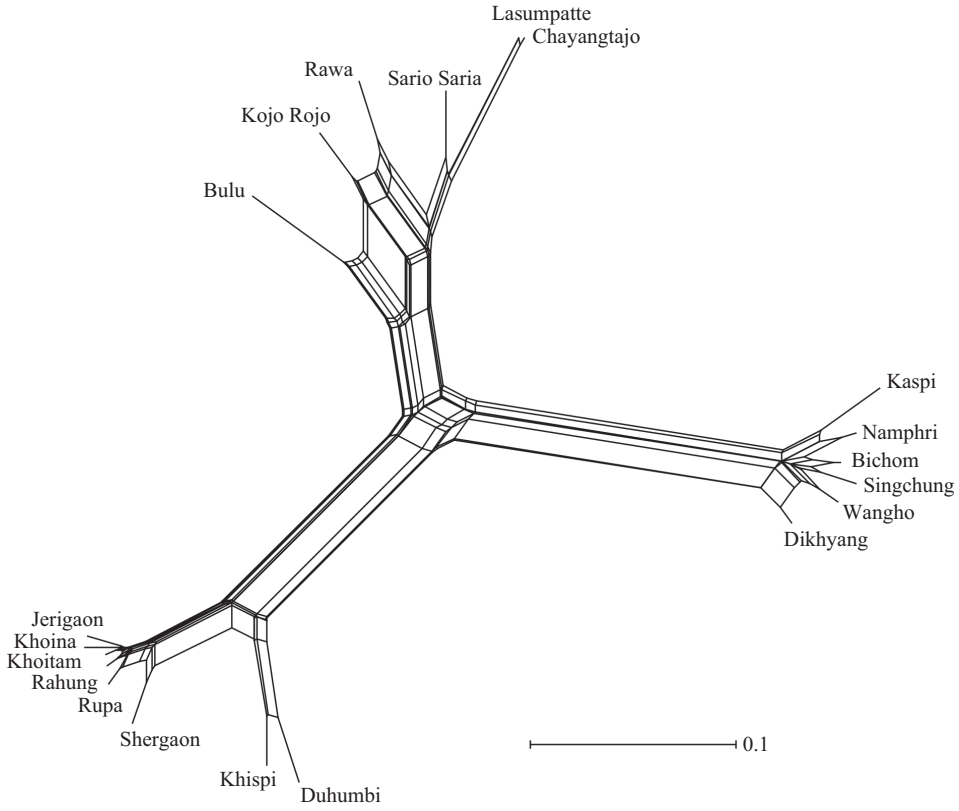


Figure 28.2
Phylogenetic visualization of the Kho-Bwa data set, with a Neighbor-Net network visualization.

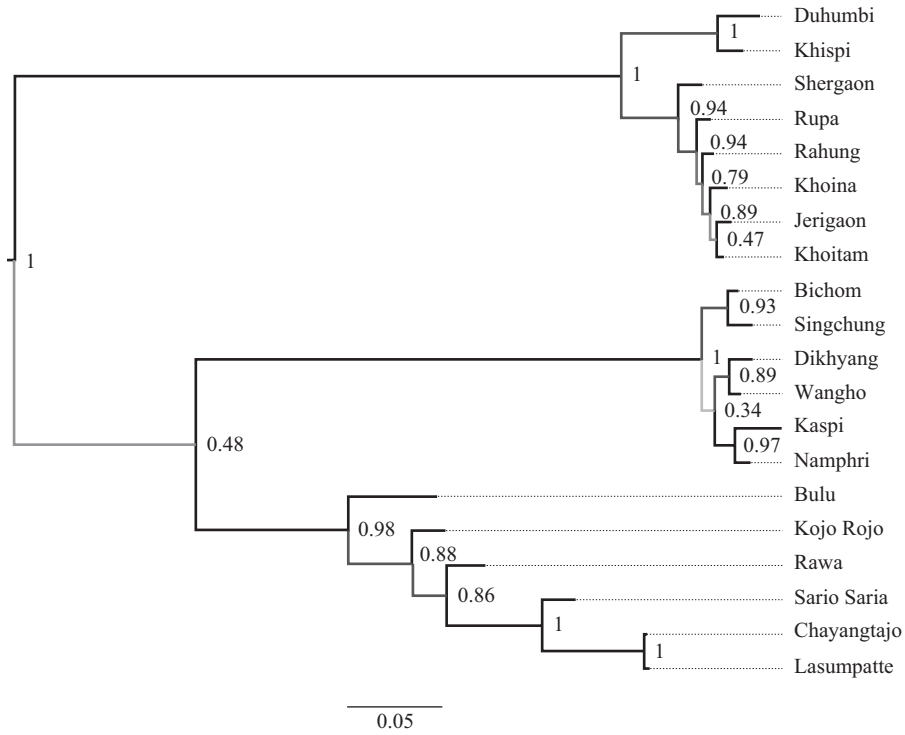


Figure 28.3
Phylogenetic visualization of the Kho-Bwa data set, with a maximum clade credibility tree of the posterior probability distribution from a Bayesian phylogenetic analysis.

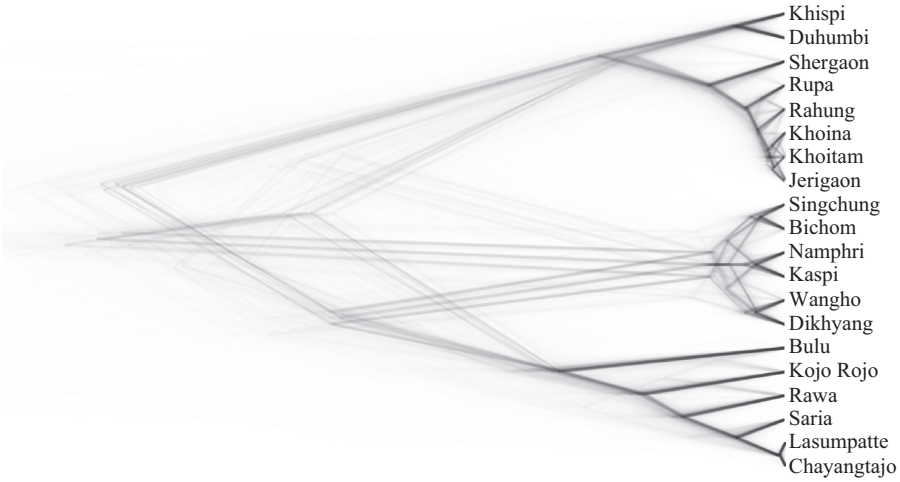


Figure 28.4
Phylogenetic visualization of the Kho-Bwa data set by means of a DensiTree (Bouckaert 2010) showing only consensus trees from the Bayesian analysis, highlighting the uncertainties in some splits and the confidence in terms of the three main groups.

in Sagart et al. 2019). In all cases, this results in desirable prospects for language groups that are still poorly understood from a historical linguistic perspective.

2.5 Data sharing and deployment (stage 5)

We encourage and practice data sharing, creating, and maintaining reusable 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 data sets were designed for convenience in sharing and deployment. While plain data sets can be shared with little effort on platforms such as GitHub and Zenodo, the related Cross-Linguistic Linked Data project (Forkel, Bank, & Rzymski 2019) allows users to deploy data into browsable web applications, as showcased in a study on colexification patterns (Database of Cross-Linguistic Colexifications 3; Rzymski et al. 2020), the typological survey of the World Atlas of Language Structures (https://wals.info; Dryer & Haspelmath 2013), a study on horizontal lexical transfer in the World Loanword Database (https://wold.clld.org; Haspelmath & Tadmor 2009), the retro-standardized version of the Tableaux phonétiques des patois suisses romands (Geisler, Forkel, & List 2020), and a collection of French and Franco-Provençal dialects (Gauchat, Jeanjacquet, & Tappolet 1925), among others. Our tutorial discusses how CLDF data sets can be shared and deployed (tutorial 2.5).

3 Conclusion

Our plan with this use 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 benefit from good practices in the representation and management of its data. Methods, questions, and solutions come and go; interdisciplinarity will evolve from its current shape; concept lists will routinely be expanded and reduced; cognate sets as basic characters of analysis might be supplemented or replaced by other data; and Bayesian phylogenetic inference might lose its momentum and be replaced by new quantitative or symbolic models, and so on, but 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 future manipulations that might be required.

Supplementary material

The supplementary material can be downloaded from https://doi.org/10.5281/zenodo.4311308 (Tresoldi et al. 2020). It contains the accompanying tutorial 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 and 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. J.-M. List and T. Tresoldi were funded by the European Research Council Starting Grant 715618 Computer-Assisted Language Comparison (https://digling.org/calc/). S. J. Greenhill was supported by the Australian Research Council's Discovery Projects funding scheme (project number DE 120101954) and the Australian Research Council 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. *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 R. 2010. DensiTree: Making sense of sets of phylogenetic trees. *Bioinformatics* 26 (10): 1372–1373.

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.

Bryant, David, and Vincent Moulton. 2003. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21 (2): 255–265.

Codd, Edgar F. 1970. A relational model of data for large shared data banks. *Communications of the ACM* 13 (6): 377–87.

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, Germany: Max Planck Institute for Evolutionary Anthropology.

Farris, J. S. 1989. The retention index and the rescaled consistency index. *Cladistics* 5:417–419.

Forkel, Robert, Sebastian, Bank, Simon J. Greenhill, and Gereon Kaiping. 2021. cldf/pycldf: pycldf, Version 1.22.0. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://pypi.org/project/pycldf/1.22.0/.

Forkel, Robert, Sebastian Bank, and Christoph Rzymski. 2021. clld/clld: clld - A toolkit for cross-linguistic databases, Version 7.4.2. Leipzig, Germany: https://pypi.org/project/clld/7.4.2.

Forkel, Robert, and Johann-Mattis List. 2020. CLDFBench: Give your Cross-Linguistic data a lift. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, 6997–7004. Marseille, France: European Language Resources Association.

Forkel, Robert, Johann-Mattis List, 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* 5:180205.

Gauchat, Louis, Jules Jeanjacquet, and Ernest Tappolet. 1925. *Tableaux phonétiques des patois suisses romands*. Neuchâtel, France: Attiger.

Geisler, Hans, Robert Forkel, and Johann-Mattis List. 2020. *The Tableaux phonétiques des patois suisses romands online*. Jena, Germany: Max Planck Institute for the Science of Human History.

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): 3923–3933.

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*, ed. K. A. Adelaar and A. Pawley, 375–397. Canberra, Australia: 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* 114 (42): E8822–E8829.

Gudschinsky, Sarah C. 1956. The ABC's of lexicostatistics (glottochronology). *Word* 12 (2): 175–210.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. Glottolog 4.4. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://glottolog.org.

Haspelmath, Martin, and Uri Tadmor, eds. 2009. *WOLD*. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology.

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*. 365:3829–3843.

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.

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 (12): 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* 294:2310–2314.

Huson, Daniel H. 1998. SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics* 14 (1): 68–73.

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.

Lieberherr, Ismael, and Timotheus A. Bodt. 2017. Sub-grouping Kho-Bwa on shared core vocabulary. *Himalayan Linguistics* 16 (2): 26–63.

Lieberherr, Ismael, and Timotheus A. Bodt. 2019. CLDF data set derived from Lieberherr and Bodt's "Comparative Wordlists of Kho-Bwa" from 2017, Version 0.9. Zenodo. December 8. https://doi.org/10.5281/zenodo.4925670.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis. 2016. *Computer-assisted Language Comparison: Reconciling Computational and Classical Approaches in Historical Linguistics*. Jena, Germany: Max Planck Institute for the Science of Human History.

List, Johann-Mattis. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological data sets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, 9–12. Valencia, Spain: Association for Computational Linguistics.

List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 45 (1): 1–24.

List, Johann-Mattis, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems*, Version v2.1.0. Leipzig: Germany: Max Planck Institute for Evolutionary Anthropology. http://clts.clld.org.

List, Johann-Mattis, and Robert Forkel. 2021. *LingPy: A Python Library for Quantitative Tasks in Historical Linguistics*, Version 2.6.7. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://lingpy.org.

List, Johann-Mattis, Christoph Rzymski, Simon Greenhill, Nathanael Schweikhard, Kristina Pianykh, Annika Tjuka, Carolin Hundt, and Robert Forkel. 2021. *Concepticon: A Resource for the Linking of Concept Lists*, Version 2.5.0. Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org.

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

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. *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*, ed. Johannes Wahle, Marisa Köllner, Harald Baayen, Gerhard Jäger, and Tineke Baayen-Oudshoorn. Tübingen, Germany: University of Tübingen.

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, and Quentin D. Atkinson. 2017. BEASTling: A software tool for linguistic phylogenetics using BEAST 2. *PLoS ONE* 12 (8): e0180908.

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? *Systematic Biology* 63 (4): 628–638.

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

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): 711–723.

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. Stroudsburg, PA: Association for Computational Linguistics.

Ringe, Donald, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. *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, 3–37. New York: Oxford University Press.

Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, et al. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data* 7:13.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *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): 406–425.

Starostin, Sergej A. 2000. *The Starling Database Program*. Moscow: RGGU.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16 (4): 157–67.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society* 96 (4): 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21 (2): 121–137.

Swadesh, Morris. 1971. *The Origin and Diversification of Language*, ed. Joel Sherzer. Chicago: Aldine.

Tresoldi, Tiago, Christoph Rzymski, Robert Forkel, Simon J. Greenhill, Johann-Mattis List, and Russell D. Gray. 2020. Supplementary code tutorial and data for "Managing Historical Linguistic Data for Computational Phylogenetics and Computer-Assisted Language Comparison," Version 0.9. Zenodo. December 8. https://doi.org/10.5281/zenodo.4311308.

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 (1): 160018.

Wu, Mei-Shin, Nathanael E. Schweikhard, Timotheus Bodt, Nathan Hill, and Johann-Mattis List. 2020. Computer-assisted language comparison: State of the art. *Journal of Open Humanities Data* 6 (2): 1–14.

This is a section of doi:10.7551/mitpress/12200.001.0001

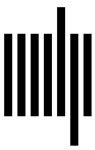
The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from the authors



The MIT Press

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363