The Journal of Open Source Software

teiphy: A Python Package for Converting TEI XML Collations to NEXUS and Other Formats

Joey McCollum [1] **and Robert Turnbull** [2]

1 Institute for Religion and Critical Inquiry, Australian Catholic University, Australia **2** Melbourne Data Analytics Platform, University of Melboune, Australia

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Textual scholars have been using phylogenetics to analyze manuscript traditions since the early 1990s (Robinson & O'Hara, 1992). Many standard phylogenetic software packages accept as input the NEXUS file format (Maddison et al., 1997). The `teiphy` program takes a collation of texts encoded using the Text Encoding Initiative (TEI) guidelines and converts it to a NEXUS file that can be used for phylogenetic analysis. The package can also convert to other formats such as Stephen C. Carlson's STEMMA format or to a NumPy array (Harris et al., 2020).

Statement of Need

The TEI aims to provide an international standard for digitally encoding textual information for the humanities (Ide & Sperberg-McQueen, 1995). The TEI guidelines describe an XML format for encoding a critical apparatus (TEI Consortium, 2022). Due to its rich and well-documented set of elements for expressing a wide range of features in manuscript transcriptions, collations, and critical editions, TEI XML has become the *de facto* format for textual data in the digital humanities (Fischer, 2020). Its expressive power has proven increasingly valuable since its release, as scholars have learned—sometimes the hard way—that digital transcriptions and collations should

1. preserve as much detail as they can from their material sources, including paratextual features;
2. reproduce the text of their sources as closely as possible, with editorial regularizations to things like orthography, accentuation, and scribal shorthand encoded alongside rather than in place of the source text; and
3. describe uncertainties about a source's contents as accurately as possible, allowing for degrees of uncertainty and multiple choices for disambiguations if necessary.

These principles have much bearing on the editing of critical texts, a task fundamental to both digital humanities and classical philology. Within the digital humanities, phylogenetic algorithms have been popular approaches to this task. Taking the most arduous part of reconstructing a textual tradition and delegating it to a computer proved to be a promising technique, and its successful demonstration with a portion of *The Canterbury Tales* was a milestone in the development of the field (Barbrook et al., 1998). Soon after this, the same methods were applied more comprehensively to the tradition of *Lanseloet van Denemerken* in a work that would formalize many practical rules for computer-assisted textual criticism (Salemans, 2000). Since then, phylogenetic methods have quickly evolved (Felsenstein, 2004), and textual critics have adapted the improvements and even added their own innovations to make the process more suitable for their purposes (Carlson, 2015; Edmondson, 2019; Hyytiäinen, 2021; Spencer et al., 2002, 2004; Turnbull, 2020).

As their name might suggest, phylogenetic methods originated in the setting of evolutionary biology. They have a natural place in textual criticism given the deep analogy between the two fields: a sequence alignment, which consists of taxa, sites or characters, and the states of taxa at those characters, corresponds almost identically to a collation, which consists of witnesses to the text, locations of textual variation (which we will call "variation units" from here on), and the variant readings attested by witnesses at those points.

Most phylogenetic software, however, expects inputs not in TEI XML format, but in NEXUS format (Maddison et al., 1997). This format was conceived with versatility in mind, and this design choice has been vindicated in its applicability with textual data, but NEXUS is not equipped or meant to express the same kinds of details that TEI XML is. Conversely, for those interested primarily in working with the collation as an alignment, TEI XML is overkill. Thus, a great chasm has been fixed between the two formats, and the only way to cross over it is by conversion.

The problem is compounded by the fact that other tools for phylogenetic and other analyses anticipate input formats other than NEXUS. A noteworthy alternative is Hennig86, which is the format of choice for the TNT phylogenetic software (Farris, 1988; Goloboff & Catalano, 2016). While this format does not allow for as much flexibility in the input as NEXUS does (e.g., it does not support ambiguities that can be disambiguated as some states and not others), TNT's remarkable performance in tree search makes support for this format a desirable option on practical grounds.

Another format of value for text-critical phylogenetics is the input format associated with the STEMMA software developed by Stephen C. Carlson for his phylogenetic analysis of the Epistle to the Galatians (Carlson, 2015). Carlson's software expands on traditional maximum parsimony-based phylogenetic algorithms with rules to account for contamination or mixture in the manuscript tradition. While it has so far only been applied to books of the New Testament, it is just as applicable to other traditions, and a way of converting TEI XML collations of other texts to a format that can be used by this software could help bridge this gap.

Other basic machine-learning approaches to textual criticism, which are frequently based on clustering and biclustering algorithms (Finney, 2018; McCollum, 2019; Thorpe, 2002), expect the collation data to be encoded as a matrix with a row for each variant reading and a column for each witness. Thus, a means of converting the essential data from TEI XML collation to a NumPy array (Harris et al., 2020) and other related formats is a need for applications like these.

Design

While the conversion process is a straightforward one for most collation data, lacunae, retroversions, and other sources of ambiguity occasionally make a one-to-one mapping of witnesses to readings impossible, and in some cases, one disambiguation may be more likely than another in a quantifiable way. Mechanisms for accommodating such situations exist in both TEI XML and NEXUS, and for likelihood-based phylogenetic methods, "soft decisions" about the states at the leaves and even the root of the tree can provide useful information to the inference process. For these reasons, we wanted to ensure that these types of judgments, as well as other rich features from TEI XML, could be respected (and, where, necessary, preserved) in the conversion process.

Collations should preserve as much detail as possible, including information on how certain types of data can be normalized and collapsed for analysis. Since one might want to conduct the same analysis at different levels of granularity, the underlying collation data should be available for use in any case, and only the output of teh conversion should reflect changes in the desired level of detail. Likewise, as noted in the previous section, uncertainty about witnesses' attestations should be encoded in the collation and preserved in the conversion of the collation.

For text-critical purposes, differences in granularity typically concern which types of variant readings we consider important for analysis. At the lowest level, readings with uncertain or reconstructed portions are almost always considered identical with their reconstructions (provided these reconstructions can be made unambiguously) for the purpose of analysis. Defective forms that are obvious misspellings of a more substantive reading are often treated the same way. Even orthographic subvariants that reflect equally "correct" regional spelling practices may be considered too common and of too trivial a nature to be of value for analysis. Other readings that do not fall under these rubrics but are nevertheless considered manifestly secondary (due to late and/or isolated attestion, for instance), may also be considered uninformative "noise" that is better left filtered out.

Use Case

Due to the availability of extensive collation data for the Greek New Testament, and because this project was originally developed for use with such data, we tested this library on a sample collation of the book of Ephesians in thirty-eight textual witnesses (including manuscripts, correctors' hands, translations to other languages, and quotations from church fathers). The manuscript transcriptions used for this collation were those produced by the University of Birmingham's Institute for Textual Scholarship and Electronic Editing (ITSEE) for the International Greek New Testament Project (IGNTP); they are freely accessible at https://itseeweb.cal.bham.ac.uk/epistulae/XML/igntp.xml. To achieve a balance between variety and conciseness, we restricted the collation to a set of forty-two variation units in Ephesians corresponding to variation units in the United Bible Societies Greek New Testament (Aland et al., 2014), which highlights variation units that affect substantive matters of translation.

In our example collation, witnesses are described in the `listWit` element under the `teiHeader`. Because most New Testament witnesses are identified by numerical Gregory-Aland identifiers, these witnesses are identified with `@n` attributes; the recommended practice is to identify such elements by `@xml:id` attributes, but this software is designed to work with either identifying attribute (preferring `@xml:id` if both are provided), and we have left things as they are to demonstrate this feature.

The `witness` elements in the example collation also contain `origDate` elements that provide dates or date ranges for the corresponding witnesses. Where a witness can be dated to a specific year, the `@when` attribute is sufficient to specify this; if it can be dated within a range of years, the `@from` and `@to` attributes or the `@notBefore` and `@notAfter` attributes should be used; the software will work with any of these options. While such dating elements are not required, our software includes them in the conversion process whenever possible. This way, phylogenetic methods that employ clock models and other chronolological constraints can benefit from this information when it is provided.

Each variation unit is encoded as an `app` element with a unique `@xml:id` attribute. Within a variation unit, a `lem` element without a `@wit` attribute presents the main text, and it is followed by `rdg` elements that describe variant readings (with the first `rdg` duplicating the `lem` reading and detailing its witnesses) and their attestations among the witnesses. (Situations where the `lem` reading is not duplicated by the first `rdg` element, but has its own `@wit` attribute, are also supported.) For conciseness, we use the `@n` attribute for each reading as a local identifier; the recommended practice for readings that will be referenced elsewhere is to use the `@xml:id` attribute, and this software will use this as the identifier if it is specified, but we have only specified `@xml:id` attributes for `rdg` elements referenced in other variation units to demonstrate the flexibility of the software. For witnesses with missing or ambiguous readings at a given variation unit, we use the `witDetail` element. For ambiguous readings, we specify their possible disambiguations with the `@target` attribute and express our degrees of certainty about these disambiguations using `certainty` elements under the `witDetail` element.

The TEI XML file for this example is available in the `example` directory of the GitHub

repository. Instructions for converting this file using `teiphy` and analyzing it with several different phylogenetic packages are provided in the documentation. Functional tests where this example file is converted and run through IQ-TREE (Minh et al., 2020), MrBayes (Ronquist et al., 2012), and STEMMA (Carlson, 2015) are part of the continuous integration (CI) pipeline. An example of the tree inferred with IQ-TREE with support values from 1000 bootstrap replicates is shown in Figure 1.

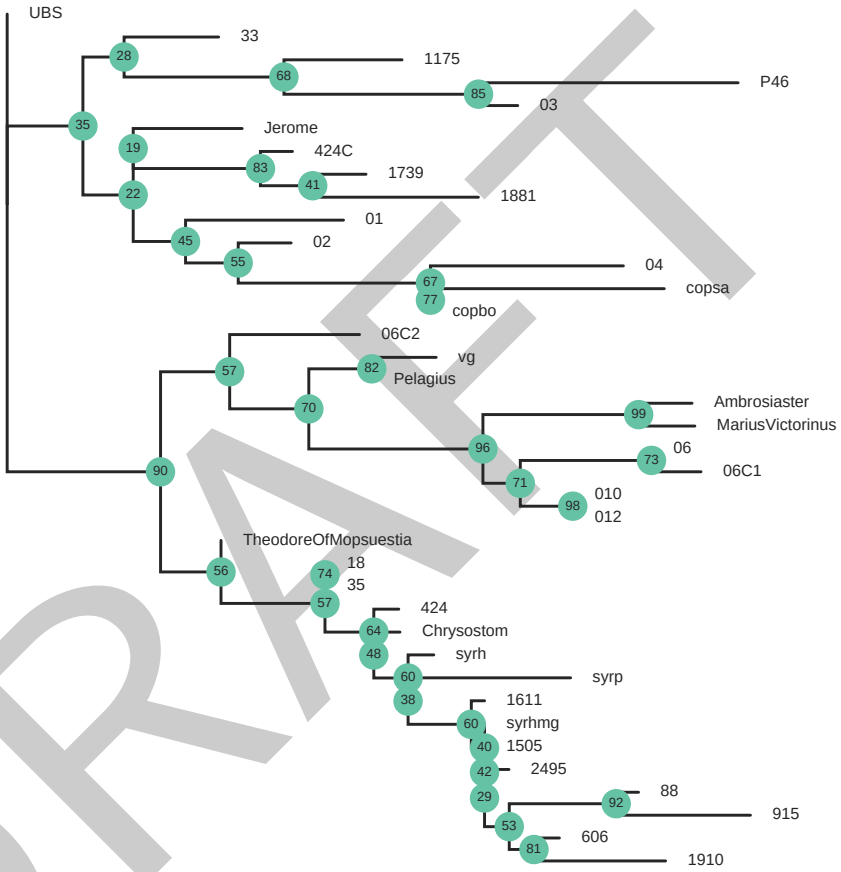

Figure 1: A phylogenetic tree inferred by IQ-TREE for the Ephesians example data. Reconstructed, defective, and orthographic sub-variants were treated as identical to their parent readings, and the text of each corrector (e.g., 06C1, 06C2) was filled in with the readings of the first hand or the previous corrector where the corrector was not active. All of these settings can be enabled or disabled in our software.

Availability

The software can be installed through the Python Package Index (PyPI), and the source code is available under the MIT license from the GitHub repository. The automated testing suite has 100% coverage.

The Journal of Open Source Software

Acknowledgements and Funding

The authors wish to thank Stephen C. Carlson for his feedback on this project. This work was supported by an Australian Government Research Training Program (RTP) Scholarship.

References

Aland, B., Aland, K., Karavidopoulos, J., Martini, C. M., & Metzger, B. M. (Eds.). (2014). *The Greek New Testament* (5th ed.). Deutsche Bibelgesellschaft.

Barbrook, A. C., Howe, C. J., Blake, N., & Robinson, P. (1998). The Phylogeny of The Canterbury Tales. *Nature*, *394*, 839. https://doi.org/10.1038/29667

Carlson, S. C. (2015). *The Text of Galatians and Its History*. Mohr Siebeck. https://doi.org/10.1628/978-3-16-153324-2

Edmondson, A. C. (2019). *An Analysis of the Coherence-Based Genealogical Method Using Phylogenetics*. University of Birmingham. http://etheses.bham.ac.uk/id/eprint/9150

Farris, J. S. (1988). *Hennig86, ver. 1.5. Program and Documentation*. James S. Farris.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates.

Finney, T. J. (2018). How to Discover Textual Groups. *Digital Studies/Le Champ Numérique*, *8*. https://doi.org/10.16995/dscn.291

Fischer, F. (2020). Representing the Critical Text. In P. Roelli (Ed.), *Handbook of stemmatology: History, methodology, digital approaches* (pp. 405–427). De Gruyter.

Goloboff, P. A., & Catalano, S. A. (2016). TNT, Version 1.5, Including a Full Implementation of Phylogenetic Morphometrics. *Cladistics*, *32*(3), 221–238. https://doi.org/10.1111/cla.12160

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hyytiäinen, P. (2021). The Changing Text of Acts: A Phylogenetic Approach. *TC: A Journal of Biblical Textual Criticism*, *26*, 1–28.

Ide, N. M., & Sperberg-McQueen, C. M. (1995). The TEI: History, Goals and Future. *Computers and the Humanities*, *29*(1), 5–15. https://doi.org/10.1007/978-94-011-0325-1_2

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). NEXUS: An Extensible File Format for Systematic Information. *Systematic Biology*, *46*(4), 590–621. https://doi.org/10.1093/sysbio/46.4.590

McCollum, J. (2019). Biclustering Readings and Manuscripts via Non-negative Matrix Factorization, with Application to the Text of Jude. *Andrews University Seminary Studies*, *57*(1), 61–89.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Haeseler, A. von, & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Robinson, P., & O'Hara, R. J. (1992). Report on the Textual Criticism Challenge 1991. *Bryn Mawr Classical Review*, *3*(4), 331–337.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space. *Systematic Biology*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

Salemans, B. J. P. (2000). *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian, Way: The Case of Fourteen Text Versions of Lanseloet van Denemerken*. Katholieke Universiteit Nijmegen. https://hdl.handle.net/2066/147058

Spencer, M., Wachtel, K., & Howe, C. J. (2002). The Greek Vorlage of the Syra Harclensis: A Comparative Study on Method in Exploring Textual Genealogy. *TC: A Journal of Biblical Textual Criticism*, *7*. http://jbtc.org/v07/SWH2002/index.html

Spencer, M., Wachtel, K., & Howe, C. J. (2004). Representing Multiple Pathways of Textual Flow in the Greek Manuscripts of the Letter of James Using Reduced Median Networks. *Computers and the Humanities*, *38*, 1–14. https://doi.org/10.1023/B:CHUM.0000009290.14571.59

TEI Consortium. (2022). *TEI P5: Guidelines for Electronic Text Encoding and Interchange: Critical Apparatus [v.4.4.0]*. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TC.html.

Thorpe, J. C. (2002). Multivariate Statistical Analysis for Manuscript Classification. *TC: A Journal of Biblical Textual Criticism*, *7*. http://jbtc.org/v07/Thorpe2002.html

Turnbull, R. (2020). *The Textual History of Codex Sinaiticus Arabicus and Its Family*. Ridley College.