

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

Joey McCollum<sup>1</sup> and Robert Turnbull<sup>2</sup>

<sup>1</sup> Institute for Religion and Critical Inquiry, Australian Catholic University, Australia <sup>2</sup> Melbourne Data Analytics Platform, University of Melbourne, Australia

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: [Open Journals](#)

Reviewers:

- [@openjournals](#)

Submitted: 01 January 1970

Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

## 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 format so that it 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 digital encoding textual information for the humanities ([Ide & Sperberg-McQueen, 1995](#)). The TEI guidelines describe an XML format for encoding a critical apparatus ([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 what a source read as accurately as possible, allowing for degrees of uncertainty and multiple choices for disambiguations if necessary.

Such 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](#)).

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

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

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

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

## 67 Design

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

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

83 For text-critical purposes, differences in granularity typically concern which types of variant  
84 readings we consider important for analysis. At the lowest level, readings with uncertain  
85 or reconstructed portions are almost always considered identical with their reconstructions  
86 (provided these reconstructions can be made unambiguously) for the purpose of analysis.  
87 Defective forms that are obvious misspellings of a more substantive reading are often treated  
88 the same way. Even orthographic subvariants that reflect equally “correct” regional spelling  
89 practices may be considered too common and of too trivial a nature to be of value for  
90 analysis. Other readings that do not fall under these rubrics but are nevertheless considered

91 manifestly secondary (due to late and/or isolated attestation, for instance), may also be considered  
92 uninformative “noise” that is better left filtered out.

## 93 Use Case

94 Due to the availability of extensive collation data for the Greek New Testament, and because  
95 this project was originally developed for use with such data, we tested this library on a  
96 collation of the book of Ephesians in over 200 textual witnesses (including manuscripts,  
97 correctors’ hands, translations to other languages, and quotations from church fathers).  
98 The manuscript transcriptions used for this collation were those produced by the University  
99 of Birmingham’s Institute for Textual Scholarship and Electronic Editing (ITSEE) for the  
100 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  
101 and conciseness, we restricted the collation to a set of forty-two variation units in Ephesians  
102 corresponding to variation units in the United Bible Societies Greek New Testament (Aland et  
103 al., 2014), which highlights variation units that affect substantive matters of translation. As a  
104 result, this collation is by no means complete, and some witnesses are lacunose for the entirety  
105 of the collation. Still, it is complete enough to serve as a sufficient example of the types of  
106 details outlined in previous sections.  
107

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

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

127 The [TEI XML file](#) for this example is available in the examples directory in the GitHub  
128 repository.

## 129 Availability

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

## 133 Acknowledgements and Funding

134 The authors wish to thank Stephen C. Carlson for his feedback on this project. This work was  
135 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. ISBN: 978-3-16-153323-5
- Consortium, T. (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>.
- 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>
- 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.
- 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>
- Hyttiäinen, 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.
- 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.
- Robinson, P., & O'Hara, R. J. (1992). Report on the Textual Criticism Challenge 1991. *Bryn Mawr Classical Review*, 3(4), 331–337.
- Salemans, B. J. P. (2000). *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian, Way: The Case of Fourteen Text Versions of Lanceloet 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>

- 180 Thorpe, J. C. (2002). Multivariate Statistical Analysis for Manuscript Classification. *TC: A*  
181 *Journal of Biblical Textual Criticism*, 7. <http://jbtc.org/v07/Thorpe2002.html>  
182 Turnbull, R. (2020). *The Textual History of Codex Sinaiticus Arabicus and Its Family*. Ridley  
183 College.

DRAFT