




# helayo: Reconstructing Sanskrit texts from manuscript witnesses

Charles Li<sup>12</sup>

<sup>1</sup> Centre nationale de la recherche scientifique <sup>2</sup> École des hautes études en sciences sociales

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

## Software

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

Editor: [Gabriela Alessio Robles](#) 

## Reviewers:

- [@kinow](#)
- [@xiaohk](#)

Submitted: 01 December 2021

Published: unpublished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

For most ancient and medieval texts, the original text itself is no longer extant in a material form. Instead, we have manuscripts that are copies of copies of copies made over the course of hundreds or thousands of years, which accumulate errors and other changes each time they are transcribed by hand. To reconstruct the original text from these imperfect copies, scholars create a stemma — analogous to an evolutionary tree — to determine the relationships between manuscripts and trace those textual changes over time.

## Statement of need

Due to the similarities in the methods used in the fields of textual reconstruction and evolutionary biology, textual scholars have begun to employ software created for biologists to analyze texts. Specifically, textual scholars are now using sequence alignment algorithms and phylogenetic tree-building packages to help reconstruct ancient texts ([Maas, 2013](#); [Phillips-Rodriguez, 2007](#); [Salemans, 2000](#)). However, as bioinformatics becomes increasingly sophisticated, its models and algorithms have become more specific and less applicable to non-biological sequences.

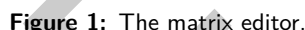
helayo has been designed from the ground up to perform multiple sequence alignment specifically for Sanskrit texts. Since Sanskrit has been written in over a dozen different scripts, each with their own orthographic peculiarities depending on their time and place, helayo performs a crucial pre-processing step in which the texts are normalized so that they can be compared meaningfully ([Li, 2017](#)). helayo can also tokenize texts either as individual characters or as *akṣaras*, since the Brahmic scripts used to write Sanskrit are abugidas, in which consonant and vowel pairs are written as a single unit.

In addition, a web-based matrix editor can be used to edit an alignment. It can also automatically reconstruct a text based on an alignment and a phylogenetic tree using the Fitch algorithm ([Fitch, 1971](#)). A full tutorial, with example files, is available at <https://chchch.github.io/sanskrit-alignment/docs>.

## Implementation

helayo is written in Haskell and implements the Center Star multiple sequence alignment algorithm ([Gusfield, 1997, pp. 347–350](#)) with an affine gap penalty model ([Li, 2021](#)). It can be run in three different tokenization modes (character, akṣara, or whitespace-delimited word) and outputs a TEI XML file which can then be edited using the matrix editor.

The matrix editor is written in Javascript and can be used either online or offline. It loads both TEI XML alignments produced by helayo as well as phylogenetic trees in NeXML format, which can be used together to reconstruct a text.



## 38

39

47

## 57

- 59

- 65 Katre, S. M., & Gode, P. K. (1941). *Introduction to Indian textual criticism*. Karnatak  
66 Publishing House.
- 67 Li, C. (2017). Critical diplomatic editing: Applying text-critical principles as algorithms. In  
68 P. Boot, A. Cappelotto, W. Dillen, F. Fischer, A. Kelly, A. Mertgens, A.-M. Sichani, E.  
69 Spadini, & D. van Hulle (Eds.), *Advances in digital scholarly editing. Papers presented at  
70 the DiXiT conferences in The Hague, Cologne, and Antwerp* (pp. 305–310). Sidestone  
71 Press.
- 72 Li, C. (2021). Align-affine: Sequence alignment with an affine gap penalty model. In *GitHub*  
73 *repository*. GitHub. <https://github.com/chchch/align-affine>
- 74 Maas, P. A. (2013). On what to do with a stemma – towards a critical edition of the  
75 Carakasaṃhitā Vimānasthāna 8.29. In D. Wujastyk, A. Cerulli, & K. Preisendanz (Eds.),  
76 *Medical texts and manuscripts in Indian cultural history*. Manohar.
- 77 Phillips-Rodriguez, W. J. (2007). *Electronic techniques of textual analysis and edition for  
78 ancient texts: An exploration of the phylogeny of the Dyūtaparvan* [PhD thesis]. University  
79 of Cambridge.
- 80 Salemans, B. J. P. (2000). *Building stemmas with the computer in a cladistic, Neo-  
81 Lachmannian, way: The case of fourteen text versions of Lanseloet van Denemerken*  
82 [PhD thesis]. Katholieke Universiteit Nijmegen.
- 83 West, M. L. (1973). *Textual criticism and editorial technique applicable to Greek and Latin*  
84 *texts*. B. G. Teubner.