




treemaker: A Python tool for constructing a Newick formatted tree from a set of classifications.

Simon J Greenhill^{1, 2}

1 Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany. **2** ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, Australia.

DOI: [10.21105/joss.01040](https://doi.org/10.21105/joss.01040)

Software

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

Submitted: 27 September 2018

Published: 08 November 2018

License

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

Summary

treemaker is a Python library to convert a text-based classification schema into a Newick file for use in phylogenetic and bioinformatic programs.

Research in linguistics or cultural evolution often produces or uses tree taxonomies or classifications. However, these are usually not in a format readily available for use in programs that can understand and manipulate trees. For example, the global taxonomy of languages published by the [Ethnologue](#) (Simons & Fennig, 2009) classifies languages into families and subgroups using a taxonomy string e.g. the language [Kalam](#) is classified as “Trans-New Guinea, Madang, Kalam-Kobon”, while [Mauwake](#) is classified as “Trans-New Guinea, Madang, Croisilles, Pihom”, and [Kare](#) is “Trans-New Guinea, Madang, Croisilles, Kare”. This classification indicates that while all these languages are part of the Madang subgroup of the Trans-New Guinea language family, Kare and Mauwake are more closely related (as they belong to the Croisilles subgroup).

Other publications use a tabular indented format to demarcate relationships, such as the example in Figure 1 from Stephen Wurm’s classification of his proposed Yele-Solomons language phylum (Wurm, 1975).

Both the taxonomy string and tabular format however are hard to load into software packages that can analyse, compare, visualise and manipulate trees. **treemaker** aims to make this easy by converting taxonomic data into [Newick](#) and Nexus (D. R. Maddison, Swofford, & Maddison, 1997) formats commonly used by phylogenetic manipulation programs.

Converting a Taxonomy to a Tree:

treemaker can convert a text file with a taxonomy to a tree. These taxonomies can easily be obtained from Ethnologue or manually entered, such as this example from Wurm’s (outdated) classification of Yele-Solomons in Figure 1:

Bilua	Yele-Solomons, Central Solomon
Baniata	Yele-Solomons, Central Solomon
Lavukaleve	Yele-Solomons, Central Solomon
Savosavo	Yele-Solomons, Central Solomon
Kazukuru	Yele-Solomons, Kazukuru
Guliguli	Yele-Solomons, Kazukuru
Dororo	Yele-Solomons, Kazukuru
Yele	Yele-Solomons

This gives the following picture of the composition of the Yele-Solomons Stock (9350¹):

- 1) The Central Solomon Family 6850
 - Bilua 4300
 - Baniata 900
 - Lavukaleve 700
 - Savosavo 950²
- 2) The Kazukuru Family }
 - Kazukuru }
 - Guliguli }
 - Dororo }
- 3) The Yele family-level Isolate 2500

Figure 1: Example of a language taxonomy in indented format from Wurm (1975).

`treemaker` can then generate a Newick representation:

```
((Baniata,Bilua,Lavukaleve,Savosavo),(Dororo,Guliguli,Kazukuru),Yele);
```

...which can then be loaded into phylogenetic programs to visualise or manipulate as in Figure 2.

`treemaker` has been used to enable the analyses in (Bromham, Hua, Cardillo, Schneemann, & Greenhill, 2018), and a number of forthcoming articles.

References

- Bromham, L., Hua, X., Cardillo, M., Schneemann, H., & Greenhill, S. J. (2018). Parasites and politics: Why cross-cultural studies must control for relatedness, proximity and covariation. *Open Science*, 5(8). doi:[10.1098/rsos.181100](https://doi.org/10.1098/rsos.181100)
- Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). Nexus: An extensible file format for systematic information. *Systematic Biology*, 46(4), 590–621. doi:[10.1093/sysbio/46.4.590](https://doi.org/10.1093/sysbio/46.4.590)
- Simons, G. F., & Fennig, C. D. (Eds.). (2009). *Ethnologue: Languages of the World* (21st ed.). Dallas, Texas: SIL International. Retrieved from <http://ethnologue.com/>
- Wurm, S. A. (1975). The East Papuan Phylum in General. In S. A. Wurm (Ed.), *New Guinea Area Languages and Language Study: Papuan languages and the New Guinea linguistic scene*. Canberra: Pacific Linguistics. doi:<http://dx.doi.org/10.15144/PL-C38>

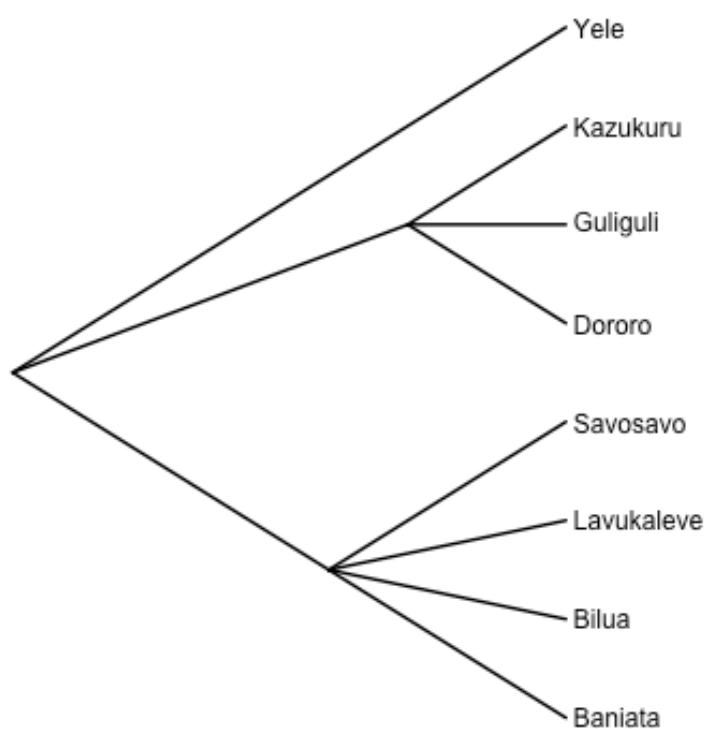


Figure 2: Tree visualisation of the relationships between the putative Yele-Solomons languages.