

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

27 September 2018

Summary

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

Often research in linguistics or cultural evolution produces tree taxonomies or classifications. However, these are often 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 and 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 (Maddison, Swofford, and Maddison 1997) formats commonly used by phylogenetic manipulation programs.

Converting a Taxonomy to a Tree:

treemaker can convert a text file with a taxonomy (easily obtained from Ethnologue or manually entered) like this:

```
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
```

Into a Newick tree 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 et al. 2018), and a number of forthcoming articles.

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).

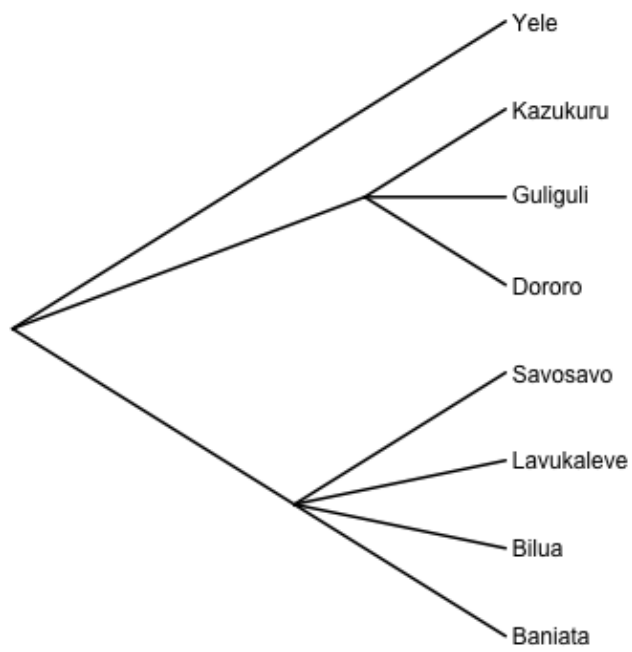


Figure 2: Tree visualisation of the relationships between these languages.

References

- Bromham, Lindell, Xia Hua, Marcel Cardillo, Hilde Schneemann, and Simon J. Greenhill. 2018. “Parasites and Politics: Why Cross-Cultural Studies Must Control for Relatedness, Proximity and Covariation.” *Open Science* 5 (8). The Royal Society. <https://doi.org/10.1098/rsos.181100>.
- Maddison, D R, D L Swofford, and Wayne P. Maddison. 1997. “Nexus: An Extensible File Format for Systematic Information.” *Systematic Biology* 46 (4): 590–621.
- Simons, Gary F., and Charles D. Fennig, eds. 2009. *Ethnologue: Languages of the World*. 21st ed. Dallas, Texas: SIL International.
- Wurm, S. A. 1975. “The East Papuan Phylum in General.” In *New Guinea Area Languages and Language Study: Papuan Languages and the New Guinea Linguistic Scene*, edited by S. A. Wurm. Canberra: Pacific Linguistics.