Interpreting and Post-Correcting the Minimum Spanning Tree

Armin Hoenen
CEDIFOR, Empirical Linguistics, Goethe University Frankfurt
hoenen@em.uni-frankfurt.de

Producing trees which exemplify the evolutionary relationships between languages is one of the fields CL is involved in. Criticism has arisen early, e.g. based on language contact. To date lexicostatistics using cognate lists of words chosen to minimize the influence of language contact as the “genome of a language” (Swadesh, 1971) is still practiced.
Bio-informatic software is being applied, see for instance Bouckaert et al. (2012); Pereltsvaig and Lewis (2015) for Indo-European. Bio-informatically generated trees do not place languages at the internal space of a tree meaning that there is no direct connection between any two languages. This is due to their biological origin. The Minimum Spanning Tree (MST) method generates topologically entirely different trees where direct relations, such as *Latin-Italian* can be present, see Figure 1, data from Bouckaert et al. (2012), a subset where the IPA was present.

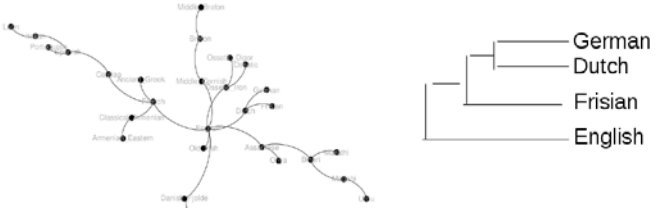

Figure 1. Subtree of self-generated MST visualized with gephi and bio-informatic NJ subtree.

Looking at the distributions of numbers of comparisons, a plateau with 51 IE languages from 8 families (Armenian, Celtic, Germanic, Greek, Indian, Iranian, Romance and Slavic) was found to have IPA and enough comparable but imbalanced data. Pairwise distances were generated using the Damerau-Levenshtein distance. We present a bio-informatic tree, an MST, a 2-d cluster plot and a method for post correcting MSTs.

References:

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the indo-european language family. Science, 337(6097):957–960.

Pereltsvaig, A. and Lewis, M. (2015). The Indo-European Controversy: Facts and Fallacies in Historical Linguistics. Cambridge University Press.

Swadesh, M. (1971). The origin and diversification of language. Aldine Transaction.