6 A computational evaluation of regularly recurring sound correspondences

Frederic Blum[1], Johann-Mattis List[1,2]

[1] Dep. of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology Leipzig
[2] Chair of Multilingual Computational Linguistics, University of Passau

Regularly recurring sound correspondences are the main tools of the comparative method (Anttila 1972; Lass 1997). The cognate judgements which are based on these correspondences are also used in the phylogenetic approaches to historical linguistics that have received widespread attention in recent years (Greenhill et al. 2020). However, regularity is often more an intuitive notion than a quantified evaluation, and irregularity is argued to be more common than expected from the Neogrammarian hypothesis (Durie & Ross 1996; Labov 1981). Given the recent devel- opment of computational methods in historical linguistics and the availability of cross-linguistic comparative formats (Forkel et al. 2018; List 2019), we are now able to improve our workflows in this regard.

We provide a computational machinery that can be used as a means to improve the anno- tation of cognates in a standardized data set. For this, we focus on a quantitative measure for assessing the regularity of sound correspondences across cognates. This can, for example, be used to compare the results of different automated methods of cognate judgements and align- ments, or to identify possible errors in expert cognate annotations. Our workflow proceeds in four stages. In the first stage, we carry out a phonetic alignment analysis (List et al. 2018) of all cognate sets in a standardized wordlist. In the second stage, we preprocess the phonetic alignments by excluding spurious alignment sites (columns in a multiple phonetic alignment). In the third stage, we search for recurring correspondences across our aligned cognate sets and determine potentially regular correspondence patterns. In a fourth stage, we score the overall regularity of the individual cognate sets in our data by counting how many sites in the align- ments can be represented by recurring (regular) correspondence patterns, and how many are unique.

In the talk, we showcase the functionality of this workflow using data from the Pano-Tacanan language family. We will focus on two key issues: the automated detection of potential false positive cognate judgements, as well as the detection of potential false negatives. Potential false positives are identified as words in a cognate set with very low regularity in the correspondence patterns across the data set. For the detection of potential false negatives, we compare two different sets of cognate annotations of the same data. If no second expert annotation is available, the first annotation can be compared to an automated judgement of cognacy (List 2019). We identify all cognate words above a custom regularity threshold that are assigned different cognacy in the first set of annotations, but are part of the same cognate set in the second annotation. We show how different thresholds influence the results and discuss possible further applications and developments of this workflow.

Anttila, Raimo. 1972. An Introduction to Historical and Comparative Linguistics. New York: The Macmillan Company.
Durie, Mark & Malcolm Ross. 1996. The Comparative Method Reviewed: Regularity and Irregularity in Language Change. New York, Oxford: Oxford University Press.
Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping & Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. Scientific Data 5(1). https://doi.org/10.1038/sdata.2018.205.
Greenhill, Simon J., Paul Heggarty & Russell D. Gray. 2020. Bayesian phylolinguistics. In Barbara S. Vance Richard D. Janda Brian D. Joseph (ed.), The Handbook of Historical Linguistics, chap. 11, 226–253. Wiley. https://doi.org/10.1002/9781118732168.ch11.
Labov, William. 1981. Resolving the Neogrammarian Controversy. Language 57(2). 267–308. https://doi.org/10.2307/413692.
Lass, Roger. 1997. Historical linguistics and language change. Cambridge: Cambridge University Press.
List, Johann-Mattis. 2019. Automatic inference of sound correspondence patterns across multiple languages. Computational Linguistics 45(1). 137–161. https://doi.org/10.1162/coli_a_ 00344.
List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. Journal of Language Evolution 3(2). 130–144. https://doi.org/10.1093/jole/lzy006.