Write-up of the cluster analysis in the Kho-Bwa paper

January 6, 2017

In the paper "A lexico-statistical survey of Kho-Bwa" we manually identified cognates in a set of 100 concepts translated into 22 Kho-Bwa languages. We then performend a hierarchical cluster analysis in order to group languages according to percentages of shared cognates. Following the suggestions of the reviewers and the editorial board, we excluded the technical details from the main paper. Here is the excluded part:

1 Hierarchical cluster analysis

The algorithm we chose for our analysis, known as "standard agglomerative method" (e.g. in Everitt et al. 2011), is as follows:

1. Join the two languages with the shortest "distance" in the similarity matrix.
2. Calculate the "distances" of this cluster to all other languages.
3. Join the closest two languages or clusters to another cluster and calculate the distances of this new cluster to all other languages and clusters.
4. repeat 1,2,3.
5. Stop when all languages are joined to under one node.

There are different ways to measure a distance in a multidimensional space and there are two different types of distances that have to be measured in the algorithm above: 1) distance between two languages 2) distance between two clusters. For the first distance we took the percentage of items that are different between a pair of language, also known as "Hamming distance". This is nothing else but 100 - cognacy_percentage. If Khispi and Duhumbi share 91 percent of core vocabulary, then the Hamming distance would be 9.

The second distance, so called "inter group proximity measure", is less straightforward. In geometry, if one had to give the distance between two non-intersecting circles in the plain, one would probably take the distance between the two points on the two circles which are closest to each other. This is indeed a measure used in cluster analysis (single linkage or nearest neighbor technique). For our purpose, however, this is not appropriate because it would give too much weight to one single language. The proximity could be due to language contact. We need a measure which takes all languages of a cluster into consideration, of which the average of all the distances between the languages in the two clusters is the most intuitive choice. More precisely: for a cluster u and a cluster v, the distance between the two clusters d_{av} is defined as in (1)

$$d_{av}(u, v) = \sum_{ij} \frac{d(u[i], v[j])}{|u||v|} \tag{1}$$

where u[i] is a language in cluster u, v[j] a language in cluster v, d(u[i], v[j]) the Hamming distance between u[i] and v[j], |u| and |v| the cardinality of the cluster u or v respectively. The "standard agglomerative method" using the average to measure distance between clusters is also known as UPGMA (Unweighted Pair Group Method with Arithmetic Mean).

In figure 1 the first eleven steps of the algorithm applied on our data. The first two columns stand for the languages or clusters compared, the third column is the distance, the fourth is the number of items the new cluster contains[1].

Table 1: First eleven rows of the linkage matrix

Cluster1	Cluster2	Distance	Cardinality
[18.	19.	2.	2.]
[2.	4.	4.	2.]
[5.	6.	4.08163265	2.]
[7.	30.	5.55555556	3.]
[9.	11.	5.61797753	2.]
[20.	21.	6.31578947	2.]
[8.	10.	6.89655172	2.]
[31.	32.	6.93053723	5.]
[12.	13.	8.98876404	2.]
[0.	1.	9.	2.]
[33.	35.	9.67971071	4.]

[1] In the Python module composed for this paper heatmap_dendrogram `--linkage` to show the linkage matrix (e.g. `heatmap_dendrogram plot --linkage [path/to/dataset_khobwa.csv]`)

In prose this means:

1. Sanchu and Lasumpatte are joined to a cluster of 2 languages[2]. Distance is 2.
2. Rupa and Rahung are joined to a cluster of 2. Distance is 4.
3. Khoitam and Jerigaon are joined to a cluster of 2. Distance is 4.08163265.
4. Khoina and the new cluster [Rupa, Rahung] are joined to a cluster of 3. Distance Khoina-Rupa is 6.06060606, Distance Khoina-Rahung is 5.05050505, average of the two distances is 5.55555556. New cluster of three languages.
5. etc.

2 Remarks

- A hierarchical cluster analysis alone does not involve any statistics, i.e. a judgement whether the result could have been produced by chance. To judge how many clusters there are is a non-trivial problem.
- A hierarchical cluster analysis is "always" possible, and there will always be a dendrogram, even if there is no meaningful clustering in the data. Due to the nature of the algorithm, all languages always link up to one single origin node. Even a language with no cognacy at all (0%) would still link up to one single node. The last node is a default node and does not provide evidence that all languages compared belong together (e.g. that all are Tibeto-Burman).
- A dendrogram *per se* is not a phylogenetic tree. It depends on the the data whether the dendrogram can have a pylogenetic interpretation. An example is the "Votes for Republican Candidate in Presidential Elections" dataset analysed by Tal Galili[3]. States in the US, which vote similarly for democrats and republicans over the years are close together in the dendrogram. This allows to investigate average political orientation of states. The dendrogram does not mean that all states under one node, descend from one proto-state or that the political orientation of these states descend from one proto-political orientation.

[2] Note that in Python the first item in a list is item number 0, the second item is item number 1 etc.

[3] https://cran.r-project.org/web/packages/dendextend/vignettes/Cluster_Analysis.html

- There are several software implementations for hierarchic cluster analysis, which are free and work out of the box (e.g. in scientific Python libraries, packages in R). We did our computations and visualisations in Python using the modules seaborn[4], numpy[5] and pandas[6]. Our data and computations we bundled on Github[7] as an small module with a command line interface in order to allow the reader to reproduce our results.

References

Everitt, Brian S. et al. (2011). *Cluster analysis*. 5th. Wiley series in probability and statistics. London: John Wiley and Sons.

[4] https://stanford.edu/~mwaskom/software/seaborn/index.html
[5] http://www.numpy.org/
[6] http://pandas.pydata.org/
[7] https://github.com/metroxylon/kho-bwa-lexicostat