Applications of Tree Spaces to Language Ancestry

Garett Ordway

Department of Statistics, FSU, Tallahassee, Florida, USA

Abstract

Two sets of three languages are considered as follows: English, French, and German; and French, Italian, and Spanish. This poster will show analysis on open books with a 3-leaf tree constructed using a single linkage method for clustering based on distances. The mean is determined for samples of 3-leaf trees. Some initial results have found both non-sticky (one leaf dominates) and sticky means (no leaf dominates). For a non-sticky mean, one language may have a different ancestor. For a sticky mean, the languages may share a common ancestor. If the means between populations are significantly different, the tree structure may be different.

Introduction

The **Indo-European Tree** traces modern languages back to the Proto-Indo-European root. Swadesh's **cognates** developed a historical perspective. Cognate words should be fundamental (not borrowed) in each language. This dendrogram shows one example for languages which use the Latin alphabet.

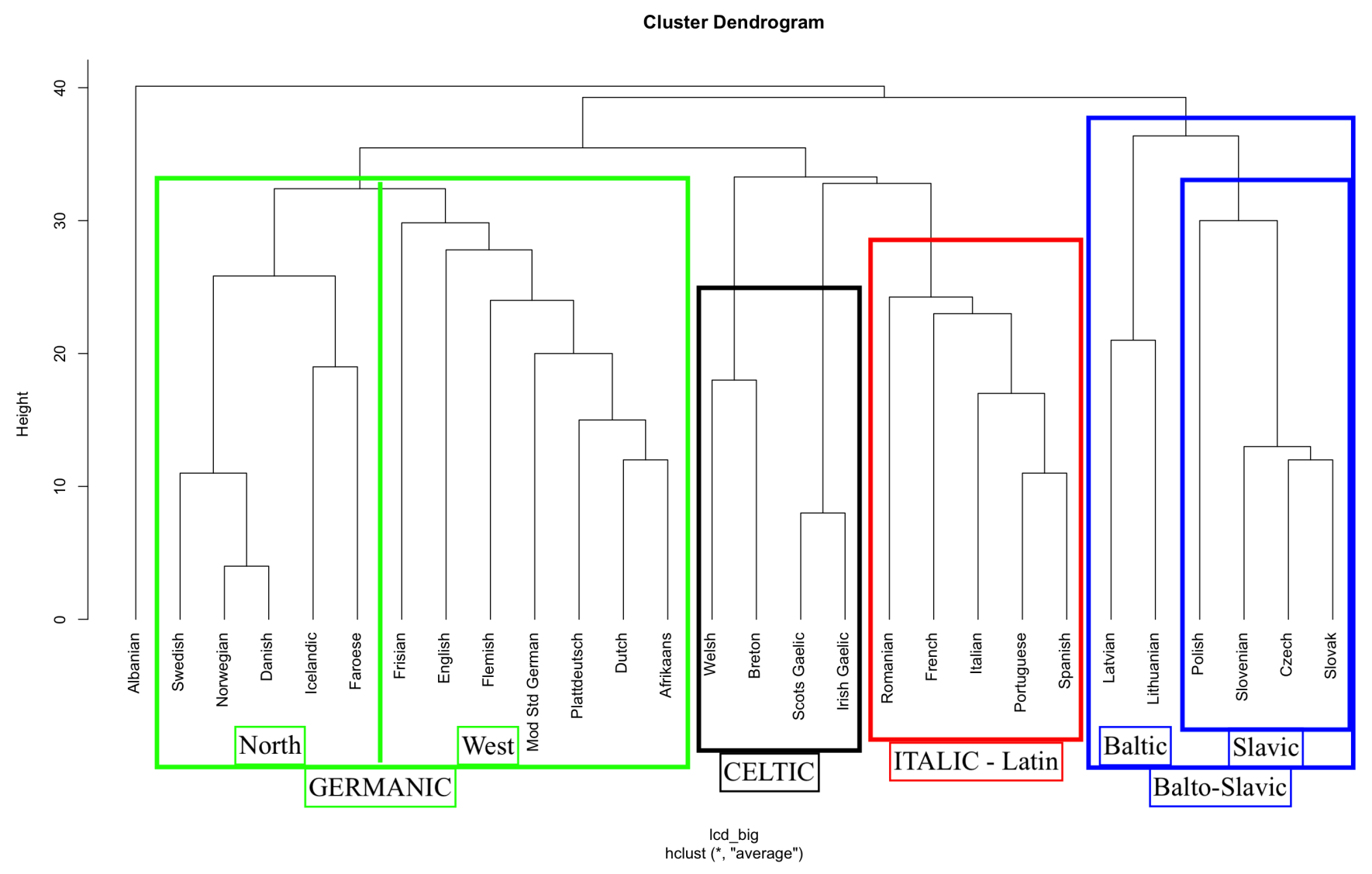


Figure: Dendrogram using the numbers one to ten

Distances between languages is a quantitative way to compare languages. Creating many small samples of language distances can allow for inferential techniques Here, only three languages are compared at a time.

Distance

A **distance measure** between two points P and Q of dimension d is denoted $d(P,Q)$ and is valid if

$$\begin{array}{ll} d(P,Q) = d(Q,P) & \\ d(P,Q) > 0 & \text{if } P \neq Q \\ d(P,Q) = 0 & \text{if } P = Q \\ d(P,Q) \leq d(P,R) + d(R,Q) & \text{(triangle inequality)} \end{array}$$

where R is some intermediate point.

For simplicity, define a distance measure having a distance zero indicating the words in two languages begin with the same letter and a distance one indicating they do not. Also of interest, *Edit distances*, including the Levenshtein distance, can be used for letter sequences.

Clustering

Single linkage is an **agglomerative hierarchical technique** (n clusters to 1 cluster) for clustering which uses the minimum distance between a point in the cluster and all other points (also known as *nearest neighbor*) to build clusters and can be pictured on a dendrogram.

In single linkage with only 3 points, four possible ways to group are considered

1. All points are equidistant necessitating only one cluster.
2. One point is furthest away and is equidistant from the other two.
3. One point is furthest away, but is closer to one point than the other.
4. Two groups of points tie for the minimum distance. For example, point E is distance c_1 from point F and point F is also distance c_1 from point G, but point E is a distance greater than c_1 from point G. In this case, randomly decide to cluster E with F or F with G.

Tree Space Properties

A tree space is an example of a stratified space which is a metric space (M, ρ) that decomposes as finite disjoint unions of manifolds (strata) in such a way that the singularities of the M are constant along each stratum. The tree space under consideration has a singularity at the point where three rays are glued together.

The mean of a tree space can be found by finding the point which minimizes the function $\Sigma_{n=1}^{N} d(p, p_n)^2$. Begin by finding the k-th folded averages $\eta_{k,N} = \frac{1}{N}\Sigma_{n=1}^{N} F_k p_n$ where the folding map $F_k p$ keeps a point positive if it is on the k-th leaf or makes negative if on a different leaf.

By theory, only one k-th folded average may be positive. The mean is then the positive $\eta_{k,N}$ indicating a dominant leaf, a *non-sticky* mean, or zero indicating no dominant leaf, a *partly sticky* (if at most one $\eta_{k,N} = 0$) or *sticky* (if no $\eta_{k,N} \geq 0$) mean.

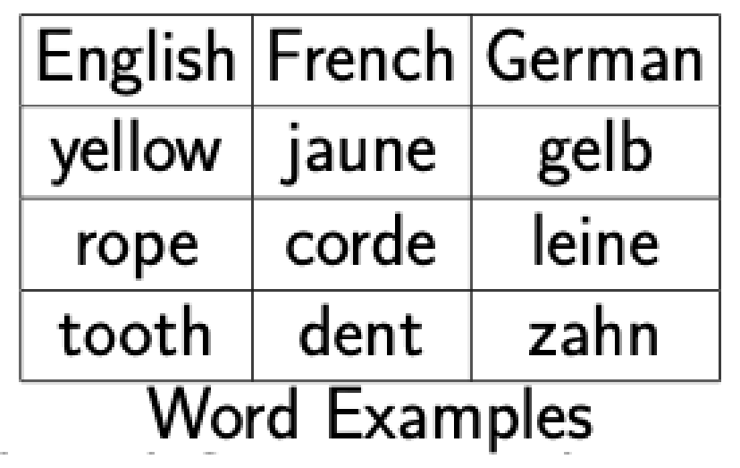
For inference, both sticky and non-sticky Central Limit Theorems have been proven (see Bhattacharya et al.).

The sample variance for a sticky mean with $d = 1$ is just $s^2 = \frac{1}{N-1}\Sigma_{n=1}^{N} p_n^2$. The sample variance for a non-sticky mean is $s^2 = \frac{1}{N-1}\Sigma_{n=1}^{N} (F_k p_n - \eta_{k,N})^2$.

Design and Examples

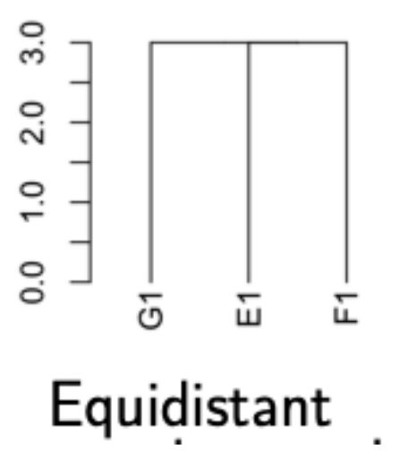
For each set of 3 languages, the Swadesh 207 list was broken into 69 random samples of size 3 and then broken into 2 random samples of size 35 and 34 simulating samples from two populations. Each sample had its distances calculated between languages and made into a dendrogram and assigned points on a 3-spider using the earlier techniques.

In this first example, all languages are distance 3 from one another, so the dendrogram connects all at the same time and the point on the 3-spider goes at the origin.

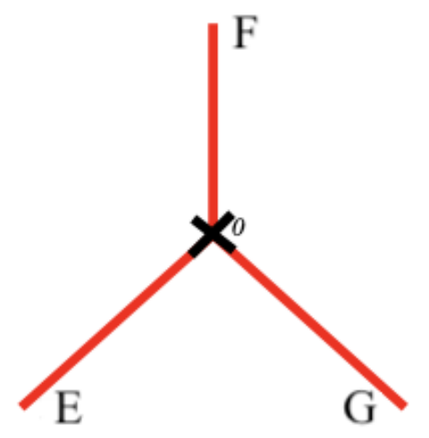


English	French	German
yellow	jaune	gelb
rope	corde	leine
tooth	dent	zahn

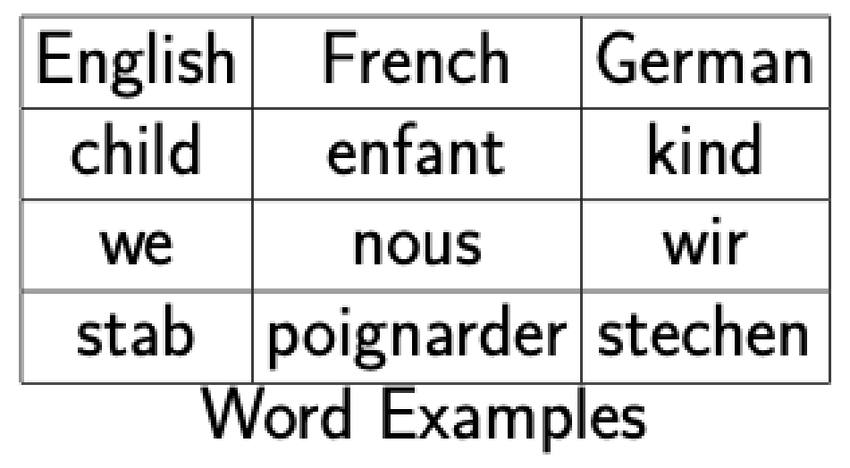
Word Examples



Equidistant

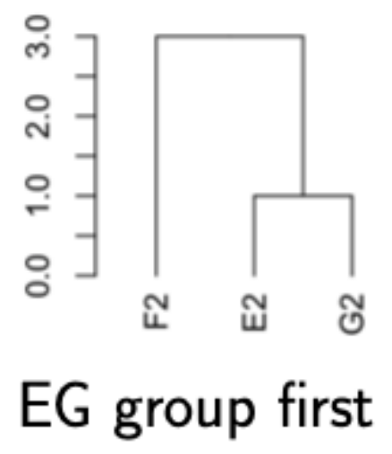


In this second example, English and German are distance 1 from each other while French is distance 3 from both English and German giving a dendrogram connecting English and German first and putting a point on the French leaf of the 3-spider.

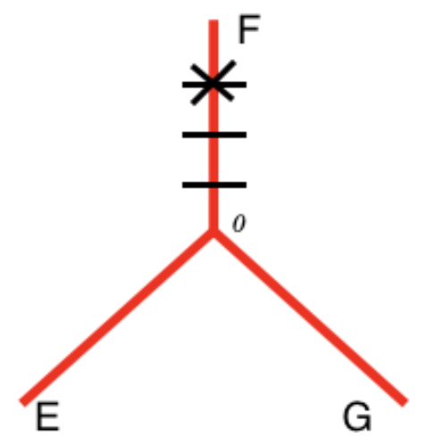


English	French	German
child	enfant	kind
we	nous	wir
stab	poignarder	stechen

Word Examples



EG group first



Results

The analysis of English, French, and German languages resulted in a non-sticky mean for both samples where French is the dominant leaf.

Language	Mean Distance 1	Mean Distance 2
English	-2.00	-2.41
German	-1.66	-1.41
French	1.49	1.41

The sample variances for each sample using French as the mean are 4.14 and 4.61, respectively. Since the current analysis was to simulate two populations, the assignment was random, and thus the t-statistic calculated of 0.16 is not surprising.

The analysis of French, Italian, and Spanish languages resulted in a sticky mean for both samples.

Language	Mean Distance 1	Mean Distance 2
French	-0.51	-0.76
Spanish	-0.74	-0.35
Italian	-0.11	-0.05

The sample variances using a mean of 0 were calculated to be 2.88 and 2.30.

Conclusion

The non-sticky mean from the English, French, and German tree indicates that French may be of different ancestry than English and German. That is the accepted position in linguistics as French is of Latin (Italic) ancestry. While the samples were simply randomly assigned, if two groups of words were different for some reason, say medical terms versus agricultural terms, then perhaps some significant difference may be found in the ancestry as a different leaf may dominate in each case.

The sticky mean from the French, Italian, and Spanish tree indicates they may all be of the same ancestry. That is the accepted position in linguistics as all are of Latin ancestry.

Further analysis of languages using tree spaces including more languages or subsets of languages or different types of distances or different alphabets (Greek, Cyrillic, Arabic) even may be of interest for further research. This elementary introduction provides expected results from several common modern languages.

References

[1] M. Swadesh (1952). Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society.* `http://www.jstor.org/stable/3143802`

[2] L. Billera and S. Holmes and K. Vogtmann (2001). Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics.* `https://doi.org/10.1006/aama.2001.0759`

[3] R.A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, 2008.

[4] T. Hotz and S. Skwerer and S. Huckemann and H. Le and J.S. Marron and J. Mattingly and E. Miller and J. Nolen and M. Owen and V. Patrangenaru (2013). Sticky central limit theorems on open books. *The Annals of Applied Probability.* `https://doi.org/10.1214/12-AAP899`

Contact Information

- Email: gao20h@fsu.edu