On a measure of centricity in Language-family structuring: A non-parametric approach to language hierarchy structuring

Abstract

Historical classification and structuring of languages employs the concept of hierarchy among the languages, and tries to model and cluster the languages, into subfamilies on the same principle. The problem with many of these classifications is the inability to define a central language, from which the other languages might have originated. Multidimensional scaling ,in which information about the pairwise distances; among a set of objects or individuals; is translated into a configuration of points mapped into an abstract Cartesian space, is performed on an optimum number of dimensions, on the languages, w.r.t the aforementioned distance obtained between the words corresponding to the languages considered. The properties of the various depth-based medians are studied on the context of structure of the families, and a new definition of "median" in the context of Historical Linguistics is defined. The class of Languages corresponding to the Indo-European family of languages have been considered for the application of the proposed classification and structuring methods.

1 Introduction

1.1 Data-depth and Linguistics

Data depth is an important concept of non parametric approach to multivariate data analysis. It provides one possible way of ordering the multivariate data. We call this ordering a central-outward ordering.Basically, any function which provides a "reasonable" central-outward ordering of points in multidimensional space can be considered as a depth function

1.1.1 Definition-Data depth

According to *Zuo et al* (2000), Statistical depth is a function possessing:

- affine transformation invariance
- maximality at the center of symmetry of the distribution for the class of symmetric distributions
- monotonicity relative to the point with the highest depth
- vanishing at infinity

We obtain a function recognizing "typical" and "outlier" observations, a generalization of quantiles for multivariate data.

Data-depth based approach to Historical Linguistics Non parametric properties and tools, which can be utilised to study the structure of the various languages in the "language space", which is particularly helpful since we do not know any distributional properties of the languages in aforementioned space.Non parametric methods,as the name suggests,does not require parametrization of the data,and are thus quite useful in cases like ours, with no distributional structure available. This generality with respect to the non-requirement of distributional assumptions helps us in developing a general measure of centricity with a language family.

2 Approach

We follow the following setup:-

- Create sets of the most commonly used words in the languages we are interested in corresponding to given meanings.
- We then find the distance matrix corresponding to the distance between any pair of languages,w.r.t a distance metric called the Levenshtein distance.

- Dimension Scaling is done using MDS(multidimensional scaling),and embed the points corresponding to various languages in an abstract Cartesian system(with appropriate scaling measures).

- Finally, we evaluate the median based properties of this particular structure obtained using appropriate non-parametric measures. We also define a central language which can be found under appropriate circumstance to be the parent language of the family with a high probability.

3 Data Collection

The primary step is the collection of data,which has been done from the famous list of *Swadesh*(1955) and also from *Wichmann*(2011) and from the database created by *Serva et al*.

Number	Word	Late Classical Latin	Megleno Romanian	Istro Romanian
1	all	omnis	tot	tot
2	ashes	cinis	tsanusa	ceruse
3	bark	cortex	coaja	cora
4	belly	venter	foali	tarbuh
5	big	grandis	mari	mare
6	bird	avis	pul	pul
7	bite	mordere	mutscu	mucca
8	black	ater	negru	negru
9	blood	sanguis	sonzi	sanze
10	bone	os	uos	os
11	breast	pectus	chiept	clept

Figure 1: A glimpse of the data set

4 Romance Family

The Romance family is quite versatile,and the language family hierarchy is given below:- Ro-

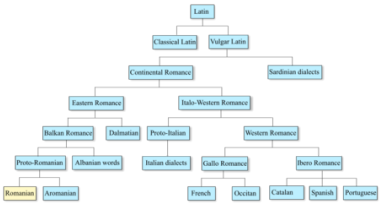

Figure 2: The romance language family

mance languages, group of related languages all derived from Vulgar Latin within historical times and forming a subgroup of the Italic branch of the Indo-European language family. The major languages of the family include French, Italian, Spanish, Portuguese, and Romanian, all national languages.

5 Distance matrix

The Levenshtein distance, a type of edit distance, is a string metric for measuring the difference between two sequences, and the Levenshtein distance between two words is defined to be the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

It is formally defined as- The Levenshtein distance between two strings p and q (of length $|p|$ and $|q|$ respectively) is given by levenshtein (p, q) where

$$\text{levenshtein.dist}(p,q) = \begin{cases} |p| & \text{if } |q| = 0, \\ |q| & \text{if } |p| = 0 \\ \text{levenshtein.dist}(\text{tail}(p), \text{tail}(q)) & \text{if head}(p) = \text{head}(q) \\ 1 + min \begin{cases} \text{levenshtein.dist}(\text{tail}(p), q) \\ \text{levenshtein.dist}(p, \text{tail}(q)) \\ \text{levenshtein.dist}(p, q) \end{cases} & \text{otherwise} \end{cases}$$

Figure 3: Levenshtein distance

Here every word has a head, which is nothing but the first letter of the word, while the tail denotes the remainder of the word after removing the head from it. This distance forms the pivotal entity which helps form the distance matrix corresponding to the languages.

For a given word-meaning, we compute the levenshtein distance between all pairs of languages,and get a matrix corresponding to the word meaning.

Using a single word meaning to create the distance matrix can lead to wrong results, particularly due to the fact that the similarity structure of words corresponding to a single meaning can show disproportionate similarity and dissimilarity among languages far and near from each other respectively,particularly due to the dominance of chance causes of similarity(or dissimilarity).

This leads us to averaging over the distance matrices obtained from the various word-meaning employed, which in turn gives us a sensible and robust distance matrix,which captures the similarity structure among the languages.

The final obtained distance matrix is shown below:-

6 Interpreting the distance matrix

The validity and meaningfulness of the obtained distance matrix is checked by performing a complete hierarchical clustering, the results of which are shown below:-

	Late.Classical.Latin	Megleno.Romanian	Istro.Romanian	Aromanian	Romanian	Dalmatian	Friulian	Gardenese.Ladin
Late.Classical.Latin	0.000000	7.122807	7.192982	7.087719	7.017544	7.245614	7.087719	6.859649
Megleno.Romanian	7.122807	0.000000	4.701754	3.403509	4.070175	6.368421	6.280702	6.754386
Istro.Romanian	7.192982	4.701754	0.000000	4.614035	3.912281	7.035088	5.824561	6.543860
Aromanian	7.087719	3.403509	4.614035	0.000000	3.403509	6.561404	6.315789	6.824561
Romanian	7.017544	4.070175	3.912281	3.403509	0.000000	6.350877	5.859649	6.333333
Dalmatian	7.245614	6.368421	7.035088	6.561404	6.350877	0.000000	5.824561	6.561404
Friulian	7.087719	6.280702	5.824561	6.315789	5.859649	5.824561	0.000000	4.473684
Gardenese.Ladin	6.859649	6.754386	6.543860	6.824561	6.333333	6.561404	4.473684	0.000000

Figure 4: A principal-submatrix corresponding to the obtained distance matrix

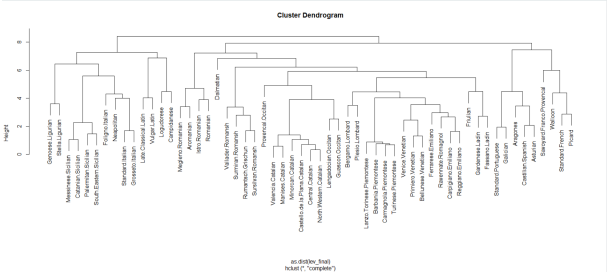

Figure 5: Hierarchical clustering(we note the lack of any measure of centricity(which can be useful in an Historical Linguistic point of view)

7 Cartesian Embedding

Now, given the distance matrix of the languages we must try to not only visualise the data set ,specifically the positioning of the languages, but also embed them in a Cartesian plane.Thus we employ multidimensional scaling with the appropriate dimensions(obtained by scaling). R provides the necessary framework for utilisation of both classical and non-parametric MDS.

The 2 dimensional MDS for the set of languages is shown-

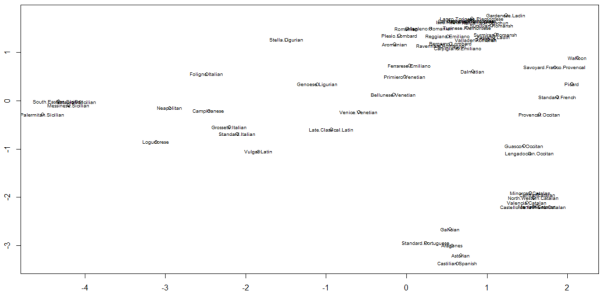

Figure 6: embedding in a cartesian plane by using MDS

8 Centricity

One of the main benefits of using the depth based approach is the ability to define a quantification of the degree of centricity of each point,which in turn, helps us get an idea about a ”central” element in some sense.

We know that the depth of a point in a data space is a quantification of how ”deep” a point is located in a language space, which is something which can be expected to be quite high for the parent language of the given language family.

In the preceding figure 6 , the concept can be understood by looking at the location of the points corresponding to the Languages-

- *Vulgar Latin* and
- *Classical Latin*

, which can be considered as parent language of the Romance sub-group of Indo-Aryan family of Languages.

A look at the previous figure 6, gives us an idea about the plausibility of the aforementioned discussion.

Hence, we define the **Median** of a Language space,as the point corresponding to that having the greatest depth measure, corresponding to some appropriate measure of depth.

However when applying methods corresponding to finding a centre element, care must be taken in clustering the sub-sub families to the best of ability, or use a weighted median approach, to be able to generate results similar to those which occurred naturally during the linguistic evolutionary process.The basic reason is we do not want to give the daughter language-dialects the same weightage as the parent languages, as daughter-language dialects can bias the results towards the particular language.Hence, we club all the dialects of a given language into a same entity,and then proceed with the non-parametric techniques.

The package *”depth: Nonparametric Depth Functions for Multivariate Analysis”* and *”ddalpha”* were used for applying the various depth structures to the language family.

For comparing among the various medians such as spatial,simplicial, zonoid,etc ,we computed the correlation between the actual rank of the languages in the hierarchy(i.e, we assign higher ranks to parent langauges,and lower rank to daughter languages;for example,we assigned 10 to Vulgar latin and 1 to french), along with the depth rank of a particular language among all the languages considered, and the results obtained were as follows-

Comparison of depth structures	
Depth type	Correlation with known hierarchical ranking
Half-Space depth	0.4752
L2-depth	0.3313
Mahalanobis depth	0.5290
Projection depth	0.4679
Qhpeeling depth	0.0531
Simplicial depth	0.2488
Zonoid depth	0.2599
Potential Depth*	0.1426
Spatial depth	0.4987

*Potential depth is not a strictly non-parametric method,but has been included in the analysis.

In this particular case,we notice that Mahalanobis depth outperforms the other depth structures,with Spatial and projection depths being close.Hence we shall usually use Mahalanobis distance as our primary definition of the median of the given language family

RESULT-After suitable sub-clustering (grouping the dialects into a single entity), we obtained *Classical Latin* as the Median of the given data set.The depth structure of the space of languages is shown here-

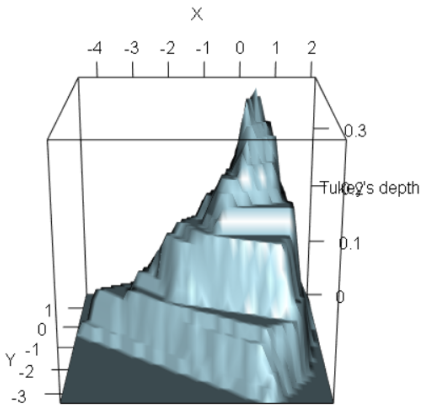

Figure 7: A 3-D perspective of the depth structure of the Languages

9 Conclusion

- Non-parametric methods helps us understand the families of languages(in our case the romance languages)
- We were able to quantify a measure of centricity of the language family such as which is found to be close to the parent language of the family, under appropriate weighing or grouping(of dialects)
- The results obtained shows a method for defining an hierarchy among languages,which can be used to understand the historical structure of the family of languages.

References

Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.

Morris Swadesh. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137, 1955.

Søren Wichmann, Taraka Rama, and Eric W Holman. Phonological diversity, word length, and population sizes across languages: The asjp evidence. 2011.

(0)(0)(0)