Statistical Classification of Languages: Generalising Ward's Method for Use with Manhattan Distances

T Strauss, MJ von Maltitz

Department of Mathematical Statistics and Actuarial Science, Faculty of Natural and Agricultural Sciences, University of the Free State, Bloemfontein, South Africa

Corresponding author: MJ von Maltitz

Abstract

The question arises whether it is possible to autonomously classify languages without any prior linguistic knowledge or assumptions. We perform statistical analyses on languages using methods normally applied to biology and genetics classification. We are concerned with the differences in character traits between languages and use a Statistical Language Signature based on relative di-gram frequencies to calculate a distance matrix between 32 Indo-European languages. We then use hierarchical clustering methods to classify the languages. We expand on existing theory by evaluating clustering methods and seek the most suitable method for classifying the languages. We identify the Manhattan distance as the most appropriate distance measure and Ward's linkage method as the most suitable linkage method. However, application of Ward's linkage method is limited to the Euclidean distance measure. We extend Ward's method to Manhattan distances and confirm that it is possible to autonomously classify languages without any prior linguistic knowledge or assumptions.

Keywords

Cluster analysis, Hierarchical classification, Ward's minimum variance method

1. Introduction

The question arises whether groupings of languages, similarities between languages and language traits well known in the field of linguistics can be extracted or independently observed using unsupervised machine learning techniques; that is, whether it is possible to autonomously classify languages without any prior linguistic knowledge or assumptions. In this paper, we discuss methods of numerical biological classification. We assume that languages can be classified in a similar way to natural organisms, and we are able to classify languages by means of these numerical biological classification methods.

Schleicher (1848) observed that languages change over time and follow the same trends as Darwin suggested for biological organisms in terms of evolution and change (Schleicher, 1863:13). Taub (1993:176) considers Schleicher's "classification of languages into types" as one of his most important contributions to linguistics. If we assume that languages can indeed be classified in a similar way to natural organisms, we can classify languages by means of a numerical biological classification system known as numerical taxonomy.

The concept of numerical taxonomy was introduced by Sokal and Sneath in 1963. This approach classifies items, based on their properties or character traits, by using numerical techniques. Numerical taxonomy uses multivariate techniques applied to classification problems (Sokal and Sneath, 1963:49). Sokal and Sneath distinguish two types of relationship between organisms: "relationships based on similarity and those based on descent" (Sokal and Sneath, 1963:95). The affinity, or overall similarity between organisms based on specific character traits, is referred to as a phenetic relationship (Sokal and Sneath, 1963:4). Sokal and Sneath (1963:220) quoted Cain and Harrison (1960) and defined the phylogenetic relationship as "that which aims to show the course of evolution". Phenetic classification is therefore defined as "a system of classification based on the overall similarity

of the organisms being classified" (Sokal, 1986). Phyletic or phylogenetic classification, on the other hand, takes into account the evolutionary ancestry of the organisms. In this paper we focus on phenetic classification, and use a method suggested by Boyce (1964) for this classification, namely, cluster analysis.

Many authors have noted the connection between biological and linguistic character traits. Mantegna, Buldyrev, Goldberger, Havlin, Peng, Simons and Stanley (1994) identify two features of language that are extended to DNA sequences in biology: Zipf's law and redundancy. Zipf's law states that "the frequency of a word decays as a (universal) power law of its rank" (Ferrer i Cancho and Solé, 2003:788). Redundancy refers to the fact that a written language can still be decipherable when characters or words are omitted or misspelt. This feature of language was shown and quantified by Shannon (1951) in an explanation of the concept of entropy in languages. "The entropy is a statistical parameter which measures, in a certain sense, how much information is produced on the average for each letter of a text" (Shannon, 1951:50). Supported by the work of Shannon (1951), authors like Turchi and Cristianini (2006) and Benedetto, Caglioti, and Loreto (2002) determined the distance between languages based on the relative frequency of di-grams, *i.e.* sequences of two letters, and the relative entropy between texts.

In Turchi and Cristianini (2006) we find an example of phylogenetic classification of languages. These authors propose a statistical signature based on the frequency of observing di-grams (pairs of letters) as explained by Shannon (1951) and a signature similar to the genetic signature in biology. They use this Statistical Language Signature (SLS) as a quantitative measure to analyse written text, and that the SLS remains more or less constant within languages, but differentiates between languages.

An issue of concern identified by Turchi and Cristianini (2006) is that written texts are simplified into using only the 26 letters of the Latin alphabet. They mapped every special letter to its closest counterpart in the Latin alphabet without considering the linguistic implications. The authors assumed that they could ignore this effect, because the approach is statistical in nature rather than linguistic, and this was the case for most of the languages. However, this simplification can cause deceptive results – as was seen with their misclassification of Breton (Turchi and Cristianini, 2006). The authors mentioned a possible solution for the problem of special characters – they suggested that languages be described not by texts written using the Latin alphabet, but rather using the International Phonetic Alphabet (IPA). It is, however, easier to obtain several translations in the Latin alphabet. We therefore continue using Latin alphabet translations, but incorporate special and accentuated characters in the Latin alphabet.

Using distance matrices, Turchi and Cristianini (2006) construct phylogenetic trees of 34 languages. The trees include 33 Indo-European languages and Basque, defined as a language isolate (Warnow, 1997) and clearly shown to be so in the way the classification trees are formed.

A similar language tree is constructed by Benedetto *et al.* (2002), where the relative entropy between pairs of texts constitutes the elements of the distance matrix. Benedetto *et al*. (2002) then apply the Fitch-Margoliash method that applies a weighted least squares method for clustering, to the distance matrix to obtain the language tree (Fitch and Margoliash, 1967). Benedetto *et al.* (2002) describe the tree they constructed as 'unrooted', *i.e.* not making any assumptions about evolutionary ancestry of the languages. This analysis relates to the phenetic analysis, where classification is done based on the similarities and differences between items and not on the presence or absence of a common ancestor.

In this paper we consider the use of cluster analysis for phenetic classification of languages. We expand on the approach of Turchi and Cristianini (2006), by using an extended version of the Latin alphabet and aim to find the most appropriate distance measure and linkage method for the classification of languages.

Section 2 presents an overview of the methods we use as well as a discussion on the relevance and suitability of each of these methods. In Section 3 we discuss the applications and results of the methods discussed in Section 2. Section 4 of this paper consists of our findings and a discussion of the suitability of the methods we used.

2. Methodology

Thirty-two Indo-European languages are analysed in this research, with the aim of identifying phenetic relationships between these languages. The texts used are translations of the Universal Declaration of Human Rights (United Nations General Assembly, 1948), as suggested by Turchi and Cristianini (2006) and Benedetto *et al*. (2002). Using the Universal Declaration of Human Rights provides the advantage that the different texts are more or less the same in length. The problem, however, is that borrowed words and words that have exactly the same translation in related languages could bias results in terms of assessing the proximity between languages (Turchi and Cristianini, 2006).

While all the selected languages use the Latin alphabet, there are different characters or special letters in each language representing different sounds and accents. Whereas Turchi and Cristianini (2006) mapped each of the accented characters to its closest equivalent in the Latin alphabet, ignoring the linguistic implications, we introduce an alphabet consisting of 60 characters: the 26 letters of the Latin alphabet, blank spaces between characters and 33 special characters found in the languages we analyse. Our extended alphabet is defined in Table 1.

Table 1. Table of characters used for analysis

a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
k	*l*	*m*	*n*	*o*	*p*	*q*	*r*	*s*	*t*
u	*v*	*w*	*x*	*y*	*z*	*ä*	*à*	*á*	*â*
å	*ã*	*æ*	*ç*	*ê*	*ë*	*è*	*é*	*ì*	*í*
î	*ñ*	*ö*	*ø*	*ó*	*ò*	*õ*	*ô*	*š*	*ś*
ß	*ü*	*ù*	*ú*	*û*	*ý*	*ž*	*ź*	*ð*	*_*

2.1 Statistical Language Signature (SLS)

The probability of observing a certain character in a linguistic sequence is highly dependent on the previous characters in the sequence as well as the language under consideration (Shannon, 1951). Based on this, Turchi and Cristianini (2006) suggest that for any given language, a SLS can be obtained by using di-grams (pairs of letters). We are interested in the number of times any given di-gram is observed in a text. We know that the di-gram 'th' will be observed often in the English language, while a di-gram such as 'en' will be more common in Afrikaans or German. The SLS for each language is based on the number of occurrences of each di-gram in that specific language. The SLS that we calculate is the relative frequency of the di-gram in each language. This is one of the methods suggested by Turchi and Cristianini (2006).

We let $n_{\alpha\beta}$ denote the number of times the di-gram '$\alpha\beta$' is observed in the document. The table consisting of the relative di-gram frequencies is defined as matrix RF with cells:

$$RF(\alpha, \beta) = \frac{n_{\alpha\beta}}{(n-1)}, \text{ where } n \text{ is the document length.} \quad (1.1)$$

Matrix RF is size 60×60. In order to avoid complications when performing cluster analysis, we henceforth describe our data as a set of 32 observations, where each observation is the SLS in vector form. Each observation is a vector of $p = 60 \times 60 = 3600$ elements.

Turchi and Cristianini (2006) investigate the use of the relative di-gram frequency table as a SLS. They propose that the SLS of a text depends on the language in which it is written and not on its semantic content. Another observation made by Turchi and Cristianini (2006:349) is that the SLS is unique to a language. If we assume this is true, we can continue using this quantitative measure in our analyses of languages. We can then quantify the proximity between languages by introducing a concept of distance, appropriate in $\mathbb{R}^{3600}$.

After each language is assigned an SLS, we consider different methods to determine the statistical distance between two languages. We then construct a distance or dissimilarity matrix. This matrix is squared with zeros along the diagonal and the number of rows and columns correspond to the number of languages under consideration.

2.2 Notation

To avoid confusion, we now specify the notation used in the rest of the paper:

- Uppercase letters such as A, B, C, or C_i, C_j, C_k, *etc*. will be used to denote clusters of languages.
- The number of elements in clusters A, B, C will be denoted by n_A, n_B, n_C, respectively. Similarly, the number of elements in clusters C_i, C_j, C_k will be denoted by n_i, n_j, n_k, respectively

- Lowercase bold letters *e.g.* $\mathbf{a}_i$ and $\mathbf{a}_j$ denote the SLS vectors of languages i and j in cluster A. Similarly the SLS vectors of language i in cluster B will be denoted by $\mathbf{b}_i$. The vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ have $p = 3600$ elements.
- Distance functions will be denoted by $d(x, y)$ where d represents a general distance, $e(x, y)$ where e represents the distance function defined by Székely and Rizzo (2005), or $e_M(x, y)$ where e_M denotes the distance function we define in the expansion of the approach followed by Székely and Rizzo (2005).
- Distance between clusters A and B is denoted as $d(A, B)$ and distance between clusters C_i and C_j is denoted as $d(C_i, C_j)$ and simplified to d_{ij}.

2.3 Distance Matrix

Let $\mathbf{a}_i$ and $\mathbf{a}_j$ be defined as above: the SLS vectors for languages i and j.

We consider the Minkowski distance suggested by Rencher (2002:453) defined in vector space $\mathbb{R}^p$.

$$D_{Minkowski}(\mathbf{a}_i, \mathbf{a}_j) = \left[\sum_{l=1}^{p} |a_{il} - a_{jl}|^r\right]^{1/r} \qquad (1.2)$$

where a_{il} represents the l^{th} element of the SLS vector $\mathbf{a}_i$ for language i in cluster A.

The Minkowski distance is a generalisation of the Euclidean distance (when the norm $r = 2$) and the Manhattan or City-block distance (when $r = 1$). We discuss these two methods of calculating the distance between a pair of languages. We then comment on the suitability of the distance measures.

2.4 The Euclidean Distance

The Euclidean distance between two languages i and j is calculated by obtaining the square root of the sum of the squared difference between each pair of elements when considering the signatures of the two languages:

Euclidean (2-Norm) distance:

$$D_E(\mathbf{a}_i, \mathbf{a}_j) = \sqrt{\sum_{l=1}^{p} \left| a_{il} - a_{jl} \right|^2} \qquad (1.3)$$

where a_{il} represents the l^{th} element of the SLS vector $\mathbf{a}_i$ for language i in cluster A.

2.5 The Manhattan Distance

The Manhattan distance between two languages i and j is calculated by the sum of the absolute difference between each pair of elements of the SLS vectors of the two languages.

Manhattan (1-Norm) Distance:

$$D_M(\mathbf{a}_i, \mathbf{a}_j) = \sum_{l=1}^{p} \left| a_{il} - a_{jl} \right| \qquad (1.4)$$

where a_{il} represents the l^{th} element of the SLS vector $\mathbf{a}_i$ for language i in cluster A.

2.6 Suitability of Distance Measures

Since we are working with categorical data and are not considering actual distances between points, we propose using the Manhattan distance to determine the dissimilarity between languages. Farris (1972) stated that the Manhattan metric is preferred to the Euclidean distance metric in numeric cladistic studies. Burgman and Sokal (1989) investigated factors that influence stability in phenetic classification and concluded that "Manhattan distances consistently produce relatively more stable classifications than do other coefficients

evaluated here" (Burgman and Sokal, 1989:67). We therefore suggest using the Manhattan distance to calculate the dissimilarity matrix between the SLS vectors of two languages.

The constructed distance matrix measures the quantitative difference between languages and has the structure 32 rows $\times$ 32 columns with zeroes along the diagonal. We perform cluster analysis using the distance matrix to construct tree diagrams or dendrograms.

2.7 Cluster Analysis (Hierarchical Clustering)

In cluster analysis we group observations into clusters. We find the optimal grouping where homogenous observations are grouped together as clusters, but the different clusters are separate from one another. We use an agglomerative hierarchical clustering approach as reviewed by Rencher (2002:455). We start with n clusters, where each observation is its own cluster. We then measure the similarity or distance between the observations by making use of the distance matrix. At each step of the agglomerative hierarchical clustering process the two clusters with the smallest distance between them are merged together into a new cluster. The distance between the new cluster and the rest of the cluster is determined by the linkage method.

2.8 Linkage Methods

Rencher (2002:456-471) summarises the following six linkage methods: single linkage, complete linkage, average linkage, centroid method linkage, median method linkage and Ward's linkage. The following properties are taken into consideration when considering the suitability of a specific clustering algorithm suggested by Rencher (2002:471-475):

- Lance-Williams form
- Monotonicity
- Space-distortion

2.8.1 Lance-Williams Algorithm

Lance and Williams (1966) suggested an algorithm for updating distances between clusters when new clusters have been formed. The two elements C_i and C_j in a dissimilarity matrix, with the smallest measure of dissimilarity between them, will be clustered together. To find the distance between cluster C_{ij} and the rest of the elements, Lance and Williams (1966) suggest the following formula where d_{ij}, d_{ik} and d_{jk} are the pairwise distances between clusters C_i, $C_j$and C_k. If C_i and C_j were to form a new cluster C_{ij}, the distance between cluster C_k and the new cluster C_{ij} is denoted as $d_{k(ij)}$. A clustering algorithm belongs to the Lance-Williams family if $d_{k(ij)}$ can be computed recursively by the following formula:

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}| \qquad (1.5)$$

Where α_i, α_j, β and γ are the parameters that together with the distance function d_{ij}, determine the clustering algorithm (Székely and Rizzo, 2005).

2.8.2 Monotonicity

Theodoridis and Koutroumbas (2003:461) explain that when the monotonicity property of a clustering method holds, each cluster is formed at a "higher dissimilarity level than any one of its components." Thus, the monotonicity property implies that a cluster cannot join another cluster at a distance that is less than the distance between previously joined clusters before merging. If a clustering method is not monotonic, it is possible that reversals can be encountered in the dendrograms; *i.e.* the resulting graphical interpretations of the clustering could contain crossovers. Monotonic clustering methods are also referred to as "ultrametric" (Rencher 2002:471).

Milligan (1979) provides conditions for the Lance-Williams parameters, which indicate whether the monotonicity property holds for a certain clustering algorithm:

$$\alpha_i + \alpha_j + \beta \geq 1$$
$$\min(\alpha_i, \alpha_j) \geq 0$$
$$\gamma \geq 0 \qquad (1.6)$$

Milligan (1979:344) also shows which clustering methods are ultrametric and finds that the centroid and median method linkage violate the monotonicity property under certain conditions.

2.8.3 Space Distortion

When new clusters are formed, the qualities of the distances between the original points before clustering do not always stay intact. Clustering algorithms that preserve the characteristics of the distances between the original points are referred to as space-conserving.

It is possible that the spatial relationships of these original distances may change (Lance and Williams, 1966). If a clustering algorithm brings about a change in the properties of this space, the clustering algorithm is space-distorting. A space-distorting clustering algorithm can either be space-contracting or space-dilating.

If the spatial relationship of the distance between original points becomes smaller, *i.e.* observations join existing clusters rather than form new clusters by joining with individual observations, then the system is said to 'chain' (Lance and Williams, 1966). In this case, clusters tend to move closer to each other and the clustering algorithm is space-contracting.

A space-dilating clustering algorithm is the opposite; an observation joins another individual observation rather than join an already-existing cluster. This means that the spatial relationship becomes larger as clusters form and clusters move further away from each other.

Székely and Rizzo (2005) mention that space-conserving or space-dilating methods are desirable in most applications. This is true in our cluster analysis of languages as we prefer separate clusters, and not 'chained' results.

Chen and Van Ness (1996) explain that the Lance Williams parameters of a clustering algorithm can be used to determine whether an algorithm is space-conserving, space-dilating, or space-contracting. For an algorithm to be space-conserving, the following conditions regarding the Lance-Williams parameters should hold (Chen and Van Ness, 1996):

$$\begin{aligned} &\alpha_i + \alpha_j = 1 \\ &\beta = 0 \\ &|\gamma| < \alpha_i \end{aligned} \qquad (1.7)$$

A space-dilating clustering algorithm satisfies the following conditions, in terms of the Lance-Williams parameters (Chen and Van Ness, 1996):

$$\begin{aligned} &\alpha_i + \alpha_j \geq 1 \\ &\alpha_i + \alpha_j + \beta \geq 1 \\ &\gamma + \alpha_i \geq 1 \end{aligned} \qquad (1.8)$$

2.8.4 Choice of Linkage Method

The first five linkage methods suggested by Rencher (2002:456-471) in Section 2.8 only consider the distances between clusters, and do not take into consideration the distances between elements within clusters. These methods simply recalculate the distances between clusters based on different criteria. These methods will not be discussed in this paper,

however, an overview of the results yielded by using each of these methods is given in Section 3.3.

Ward's method, also referred to as the incremental sum of squares method (Rencher, 2002:466) or Ward's minimum variance method (Székely and Rizzo, 2005) takes into consideration, not only between-cluster distances when forming clusters, but also within-cluster distances. Ward's method states that, not only should the between-cluster distances be maximised, but the within-cluster distances should also be minimised. This method combines these two properties into one criterion (Ward, 1963). Milligan (1979:344) shows that Ward's method fits the Lance-Williams algorithm, gives appropriate parameters, and asserts that the monotonicity property does hold for this method. Vogt and Nagel (1992) claim that Ward's clustering algorithm is space-conserving. We therefore propose the use of Ward's method as clustering algorithm for our analysis, since, while adhering to the three desirable properties mentioned above, the method also accounts for both inter- and intra-cluster distances.

The use of Ward's Linkage, however, is limited to use with the squared Euclidean distance metric as the measure of original distances between observations. This is because the objective function is often chosen as the minimum variance, or minimum squared error. The Euclidean distance is related to the measurement of the sum of squared errors; hence the use of this metric when using Ward's linkage method.

We use the Manhattan distance as measure between observations. We therefore attempt to define Ward's method for use with other distance metrics, in particular the Manhattan distance metric.

2.9 A Variation on Ward's Minimum Variance Method

Many authors refer to Ward's linkage method as the minimum variance method: Rencher (2002:466), Cormack (1971) and Milligan (1979) to name but a few. Ward (1963) suggested that the decision on which a pair of clusters is to be joined should be based on the optimal value of an objective function. Ward (1963) then used the example of least squared error, or minimum variance, as an objective function. It is this example that has become famous as Ward's method or Ward's method of minimum variance. However, Ward did explain that the objective function "reflects the criterion chosen by the investigator" (Ward, 1963:236).

"Ward (1963) suggested a general hierarchical clustering procedure where the criterion for selecting the optimal pair of clusters to merge at each step is based on the optimal value of an objective function. The objective function could be any function that reflects the investigator's purpose" (Székely and Rizzo, 2005:160).

Ward's method is most commonly used with the objective function of minimum variance. If we, however, decide to use the Manhattan distance we propose using an objective function of minimum absolute deviation.

We discuss the objective function for Ward's minimum variance method, as well as the objective function used by Székely and Rizzo (2005). We then propose our own objective function. After we have identified an objective function, it is important to know how the distance measure will be updated after each step of clustering. For updating the distance matrix, we also discuss the Lance-Williams algorithm for each of the three objective functions.

2.9.1 Ward's Minimum Variance Method

Ward's minimum variance method joins the two clusters A and B that minimise the increase in the sum of squared errors (SSE):

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$

We define the SSE within and between clusters as follows:

$$SSE_A = \sum_{i=1}^{n_A} (\mathbf{a}_i - \bar{\mathbf{a}})' (\mathbf{a}_i - \bar{\mathbf{a}})$$

$$SSE_B = \sum_{i=1}^{n_B} (\mathbf{b}_i - \bar{\mathbf{b}})' (\mathbf{b}_i - \bar{\mathbf{b}})$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (\mathbf{y}_i - \bar{\mathbf{y}}_{AB})' (\mathbf{y}_i - \bar{\mathbf{y}}_{AB}) \quad (1.9)$$

Where:

- $\mathbf{a}_i$ represents the SLS vector for language i in cluster A, and $\bar{\mathbf{a}}$ the centroid of cluster A.
- $\mathbf{b}_i$ represents the SLS vector for language i in cluster B, and $\bar{\mathbf{b}}$ the centroid of cluster B.
- $\mathbf{y}_i$ represents the combined observation vector for language i in cluster AB, and $\bar{\mathbf{y}}_{AB}$ the centroid of cluster AB.

In other words, Ward's minimum variance method calculates the distance between cluster members and the centroid. The centroid of a cluster is defined as the point at which the sum of squared Euclidean distances between the point itself and each other point in the cluster is minimised. Rencher (2002:463) also refers to the centroids of the clusters as their mean vectors. The centroid of cluster A is defined as the sum of all points in A divided by the number of points in A, or mathematically: $\bar{\mathbf{a}} = \sum_{i=1}^{n_A} \mathbf{a}_i / n_A$. Rencher (2002) states that the

objective function to minimise when using Ward's minimum variance method can also be written as,

$$I_{AB} = \frac{n_A n_B}{n_A + n_B}\left(\bar{\mathbf{a}} - \bar{\mathbf{b}}\right)'\left(\bar{\mathbf{a}} - \bar{\mathbf{b}}\right), \qquad (1.10)$$

where $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$ represent the centroids of clusters A and B, respectively.

Because this objective function is based on the distances between the centroids of the clusters (Rencher, 2005:466-468; Lance and Williams, 1966) it is necessary to use the squared Euclidean distance as the metric to calculate distances between objects. Ward's minimum variance linkage method can therefore only be applied to distance matrices using the squared Euclidean distance metric.

Ward's minimum variance method satisfies the recurrence relation as proposed by Lance and Williams (1966). Cormack (1971), Milligan (1979) and Rencher (2002:470) provide the values for α_i, α_j, β and γ when using Ward's method of minimum variance:

$$\alpha_i = \frac{n_i + n_k}{n_i + n_j + n_k}$$

$$\alpha_j = \frac{n_j + n_k}{n_i + n_j + n_k}$$

$$\beta = \frac{(-n_k)}{n_i + n_j + n_k}$$

$$\gamma = 0 \qquad (1.11)$$

2.9.2 Székely and Rizzo

Székely and Rizzo (2005) extend the use of Ward's method by showing that the same Lance-Williams parameters are applicable, even if the objective function is not minimum variance (*i.e.* when the distance metric is not squared Euclidean). They still use the Euclidean metric, but show that these parameters are also applicable to any power α of Euclidean distance

where $0 < \alpha \leq 2$, by generalising the objective function. Thus, Székely and Rizzo (2005) propose an objective function using the Euclidean distances between all the observations within a cluster and all the observations between clusters. They define a distance, the *e*-distance, $e(A, B)$, between clusters $A = \{\mathbf{a}_1, \dots, \mathbf{a}_{n_A}\}$ and $B = \{\mathbf{b}_1, \dots, \mathbf{b}_{n_B}\}$ with each vector in A or B consisting of p different values:

$$e(A,B) = \frac{n_A n_b}{n_A + n_B}\left(\frac{2}{n_A n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(\mathbf{a}_i, \mathbf{b}_j) - \frac{1}{n_A^2}\sum_{i=1}^{n_A}\sum_{j=1}^{n_A} d(\mathbf{a}_i, \mathbf{a}_j) - \frac{1}{n_B^2}\sum_{i=1}^{n_B}\sum_{j=1}^{n_B} d(\mathbf{b}_i, \mathbf{b}_j)\right). \quad (1.12)$$

If the objective function is minimum variance, then $d(\mathbf{a}_i, \mathbf{b}_j)$ denotes the squared Euclidean distance:

$$d(\mathbf{a}_i, \mathbf{b}_j) = \left(\sqrt{\sum_{l=1}^{p}(a_{il} - b_{jl})^2}\right)^2$$

and $\frac{1}{n_A^2}\sum_{i=1}^{n_A}\sum_{j=1}^{n_A} d(\mathbf{a}_i, \mathbf{a}_j)$ represents the mean squared error within cluster A.

If the objective function is not minimum variance, but rather the function defined in Székely and Rizzo (2005), then $d(\mathbf{a}_i, \mathbf{b}_j)$ denotes the Euclidean distance to the power α where $0 < \alpha \leq 2$:

$$d(\mathbf{a}_i, \mathbf{b}_j) = \left(\sqrt{\sum_{l=1}^{p}(a_{il} - b_{jl})^2}\right)^{\alpha}$$

and $\frac{1}{n_A^2}\sum_{i=1}^{n_A}\sum_{j=1}^{n_A} d(\mathbf{a}_i, \mathbf{a}_j)$ now represents the mean error to the power α within cluster A.

Székely and Rizzo (2005) show that the Lance-Williams parameters for their objective function (Equation 1.12) are the same as the parameters for the minimum variance method.

2.9.3 Least Absolute Deviation

Székely and Rizzo (2005) defined their objective function using the distance between all elements in a cluster and were no longer restricted to the use of the sum of squared errors as objective function. They can, therefore, generalise Ward's method for the use of any power of Euclidean distance. Since Székely and Rizzo (2005) show that using the distance between every single observation is also acceptable in Ward's clustering algorithm, we generalise the method of Székely and Rizzo (2005) even further. We now use a 1-norm distance, for instance the Manhattan metric, to calculate the distances between observations. Our objection function will be least absolute error. With this objective function, Ward's method should join the two clusters A and B that minimise the increase in absolute deviation or absolute error (AE):

$$I_{AB} = AE_{AB} - AE_A - AE_B.$$

We define the within cluster and between cluster absolute error as follows:

$$AE_A = \sum_{l=1}^{p} |a_{il} - a_{jl}|$$

$$AE_B = \sum_{l=1}^{p} |b_{il} - b_{jl}|$$

$$AE_{AB} = \sum_{l=1}^{p} |a_{il} - b_{jl}| \qquad (1.13)$$

where:

- a_{il} represents the l^{th} element of the SLS vector $\mathbf{a}_i$ for language i in cluster A,
- b_{il} represents the l^{th} element of the SLS vector $\mathbf{b}_i$ for language i in cluster B.

We use the *e*-distance, $e(A, B)$, that Székely and Rizzo (2005) defined between clusters $A = \{\mathbf{a}_1, \dots, \mathbf{a}_{n_A}\}$ and $B = \{\mathbf{b}_1, \dots, \mathbf{b}_{n_B}\}$ in Equation 1.13. However, we now have a different objective function and therefore the measure $d(\mathbf{a}_i, \mathbf{b}_j)$ is no longer Euclidean, but describes a Manhattan distance:

$$d_M(\mathbf{a}_i, \mathbf{b}_j) = \sum_{l=1}^{p} |a_{il} - b_{jl}|.$$

If we can prove that the distance $e(A, B)$ in Equation 1.13 suggested by Székely and Rizzo (2005) can be used with our measure of $d_M(\mathbf{a}_i, \mathbf{b}_j)$, we generalise Ward's method even further and show that it can be used with non–Euclidean distances as well. If we are able to prove this, it follows that the same Lance-Williams parameters are applicable to our objective function. Thus, the proof given by Székely and Rizzo (2005) should also hold when we use an objective function based on an L1 Distance like the Manhattan distance.

2.10 Generalising Ward's Method: Least Absolute Error Method

Suppose $A = \{\mathbf{a}_1, \dots, \mathbf{a}_{n_A}\}$, $B = \{\mathbf{b}_1, \dots, \mathbf{b}_{n_B}\}$, and $C = \{\mathbf{c}_1, \dots, \mathbf{c}_{n_C}\}$ are distinct clusters with all the vectors $\mathbf{a}_i$, $\mathbf{b}_i$ and $\mathbf{c}_i$ consisting of p elements:

Székely and Rizzo (2005) defined the constants δ_{AA}, δ_{BB} and δ_{AB}:

$$\delta_{AA} = \frac{1}{n_A^2} \sum_{i=1}^{n_A} \sum_{j=1}^{n_A} d(\mathbf{a}_i, \mathbf{a}_j)$$

$$\delta_{BB} = \frac{1}{n_B^2} \sum_{i=1}^{n_B} \sum_{j=1}^{n_B} d(\mathbf{b}_i, \mathbf{b}_j) \qquad (1.14)$$

$$\delta_{AB} = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(\mathbf{a}_i, \mathbf{b}_j)$$

If we can justify that these constants can be defined similarly with our distance measure, we can continue the proof in the same way as Székely and Rizzo (2005).

The constants, defined by Székely and Rizzo (2005), represent the mean squared error within and between clusters for the minimum variance method, and the mean error to the power α within and between clusters for the Extended Method that Székely and Rizzo (2005) defined. It is clear that when we use the Manhattan distance we have,

$$d(\mathbf{a}_i, \mathbf{b}_j) = d_M(\mathbf{a}_i, \mathbf{b}_j) = \sum_{l=1}^{p} |a_{il} - b_{jl}| \qquad (1.15)$$

By replacing the distance $d(\mathbf{a}_i, \mathbf{b}_j)$ with $d_M(\mathbf{a}_i, \mathbf{b}_j)$ as the distance, we just define the mean absolute error within and between clusters. This is exactly what we want to achieve, as our objective function is minimum absolute error. We are therefore able to use our distance measure $d_M(\mathbf{a}_i, \mathbf{b}_j)$ in the constants defined by Székely and Rizzo (2005) in a way that makes sense, and we continue to show that the rest of the proof now also holds for our distance metric.

We first define the constants in terms of d_M our distance measure:

$$\delta_{AA} = \frac{1}{n_A^2} \sum_{i=1}^{n_A} \sum_{j=1}^{n_A} d_M(\mathbf{a}_i, \mathbf{a}_j)$$

$$\delta_{BB} = \frac{1}{n_B^2} \sum_{i=1}^{n_B} \sum_{j=1}^{n_B} d_M(\mathbf{b}_i, \mathbf{b}_j) \qquad (1.16)$$

$$\delta_{AB} = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d_M(\mathbf{a}_i, \mathbf{b}_j)$$

where:

- δ_{AA} represents the mean absolute deviation ***within*** cluster A: distance between all the vectors $\mathbf{a}_i$ and $\mathbf{a}_j$,
- δ_{BB} represents the mean absolute deviation ***within*** cluster B: distance between all the vectors $\mathbf{b}_i$ and $\mathbf{b}_j$,
- δ_{AB} represents the mean absolute deviation ***between*** clusters A and B: distance between all the vectors $\mathbf{a}_i$ and $\mathbf{b}_j$.

We note that Székely and Rizzo (2005) used the constant $\frac{n_A n_b}{n_A + n_B}$, as also used by Rencher (2002:468). Then, similar to the $e(A, B)$ definition from Székely and Rizzo (2005) in Equation 1.13, we define$e_M(A, B)$:

$$e_M(A,B) = \frac{n_A n_b}{n_A + n_B} \left(\frac{2}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=2}^{n_B} d_M(\mathbf{a}_i, \mathbf{b}_j) - \frac{1}{n_A^2} \sum_{i=1}^{n_A} \sum_{j=1}^{n_A} d_M(\mathbf{a}_i, \mathbf{a}_j) - \frac{1}{n_B^2} \sum_{i=1}^{n_B} \sum_{j=1}^{n_B} d_M(\mathbf{b}_i, \mathbf{b}_j) \right)$$

$$= \frac{n_A n_B}{n_A + n_B} (2\delta_{AB} - \delta_{AA} - \delta_{BB}) \qquad (1.17)$$

Similar to δ_{AA}, δ_{BB} and δ_{AB}, we define the constants δ_{CC}, δ_{AC} and δ_{BC}:

$$\delta_{CC} = \frac{1}{n_C^2} \sum_{i=1}^{n_C} \sum_{j=1}^{n_C} d_M(\mathbf{c}_i, \mathbf{c}_j)$$

$$\delta_{AC} = \frac{1}{n_A n_C} \sum_{i=1}^{n_A} \sum_{j=1}^{n_C} d_M(\mathbf{a}_i, \mathbf{c}_j) \qquad (1.18)$$

$$\delta_{BC} = \frac{1}{n_B n_C} \sum_{i=1}^{n_B} \sum_{j=1}^{n_C} d_M(\mathbf{b}_i, \mathbf{c}_j)$$

where:

- δ_{CC} represents the mean absolute deviation ***within*** cluster C: distance between all the vectors $\mathbf{c}_i$ and $\mathbf{c}_j$,
- δ_{AC} represents the mean absolute deviation ***between*** clusters A and C: distance between all the vectors $\mathbf{a}_i$ and $\mathbf{c}_j$,
- δ_{BC} represents the mean absolute deviation ***between*** clusters B and C: distance between all the vectors $\mathbf{b}_i$ and $\mathbf{c}_j$.

Now we have, similar to $e_M(A, B)$ in Equation 1.18:

$$e_M(A, C) = \frac{n_A n_C}{n_A + n_C}(2\delta_{AC} - \delta_{AA} - \delta_{CC})$$

$$e_M(B, C) = \frac{n_B n_C}{n_B + n_C}(2\delta_{BC} - \delta_{BB} - \delta_{CC})$$

Consider cluster $A \cup B$ formed by merging clusters A and B. We denote $A \cup B$ by K, and define the following constants δ_{KC} and δ_{KK} :

$$\delta_{KC} = \frac{1}{n_K n_C} \sum_{i=1}^{n_K} \sum_{j=1}^{n_C} d_M(\mathbf{k}_i, \mathbf{c}_j)$$

δ_{KC} is the mean absolute deviation between clusters C and $A \cup B$ (the distance between all vectors $\boldsymbol{c}_j$ in C and all vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ in $A \cup B$).

Therefore δ_{KC} should represent the mean absolute deviation between:

1. all vectors $\mathbf{a}_i$ in $A \cup B$ and $\boldsymbol{c}_j$ in C (equivalent to all vectors $\mathbf{a}_i$ in A and $\mathbf{c}_j$ in C), and

2. all vectors $\mathbf{b}_i$ in $A \cup B$ and $\boldsymbol{c}_j$ in C (equivalent to all vectors $\mathbf{b}_i$ in B and $\mathbf{c}_j$ in C).

$$\therefore \delta_{KC} = \frac{1}{(n_A + n_B)n_C}\left(\sum_{i=1}^{n_A}\sum_{j=1}^{n_C} d_M(\mathbf{a}_i, \mathbf{c}_j) + \sum_{i=1}^{n_B}\sum_{j=1}^{n_C} d_M(\mathbf{b}_i, \mathbf{c}_j)\right) \quad (1.19)$$

Similar to Equations 1.17 and 1.19, constant δ_{KK} is defined as:

$$\delta_{KK} = \frac{1}{n_K^2}\sum_{i=1}^{n_K}\sum_{j=1}^{n_K} d_M(\mathbf{k}_i, \mathbf{k}_j),$$

where, δ_{KK} is the mean absolute deviation within cluster $A \cup B$ (the distance between all vectors in $(A \cup B)_i$ ($\mathbf{a}_i$ and $\mathbf{b}_i$) and all vectors in $(A \cup B)_j$, *i.e.* $\mathbf{a}_j$ and $\mathbf{b}_j$).

Therefore, δ_{KK} should represent the mean absolute deviation between:

1. all vectors $\mathbf{a}_i$ in $(A \cup B)_i$and $\mathbf{a}_j$ in $(A \cup B)_j$,

2. all vectors $\mathbf{a}_i$ in $(A \cup B)_i$and $\mathbf{b}_j$ in $(A \cup B)_j$,

3. all vectors $\mathbf{b}_i$ in $(A \cup B)_i$and $\mathbf{a}_j$ in $(A \cup B)_j$, and

2. all vectors $\mathbf{b}_i$ in $(A \cup B)_i$and $\mathbf{b}_j$ in $(A \cup B)_j$.

$$\therefore \delta_{KK} = \frac{1}{(n_A + n_B)^2}\left(\sum_{i=1}^{n_A}\sum_{j=1}^{n_A} d_M(\mathbf{a}_i, \mathbf{a}_j) + 2\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d_M(\mathbf{a}_i, \mathbf{b}_j) + \sum_{i=1}^{n_B}\sum_{j=1}^{n_B} d_M(\mathbf{b}_i, \mathbf{b}_j)\right) \quad (1.20)$$

In terms of the original constants, we now have:

$$\delta_{KC} = \frac{n_A n_C \delta_{AC} + n_B n_C \delta_{BC}}{(n_A + n_B)n_C}$$

$$\delta_{KK} = \frac{n_A^2 \delta_{AA} + 2n_A n_B \delta_{AB} + n_B^2 \delta_{BB}}{(n_A + n_B)^2} \quad (1.21)$$

We define$e_M(A \cup B, C)$ similar to the way we defined $e_M(A, B)$ in Equation 1.18:

$$e_M(A \cup B, C) = e_M(K, C) = \frac{n_K n_C}{n_K + n_C}(2\delta_{KC} - \delta_{KK} - \delta_{CC})$$

$$\therefore e_M(A \cup B, C) = \frac{(n_A + n_B)n_C}{n_A + n_B + n_C}\left(\frac{2n_A n_C \delta_{AC} + 2n_B n_C \delta_{BC}}{(n_A + n_B)n_C}\right.$$

$$\left. - \frac{n_A^2 \delta_{AA} + 2n_A n_B \delta_{AB} + n_B^2 \delta_{BB}}{(n_A + n_B)^2} - \delta_{CC}\right)$$

$$\therefore e_M(A \cup B, C) = \frac{(n_A + n_B)n_C}{n_A + n_B + n_C}\left(\frac{2n_A n_C \delta_{AC} + 2n_B n_C \delta_{BC}}{(n_A + n_B)n_C}\right)$$

$$- \frac{(n_A + n_B)n_C}{n_A + n_B + n_C}\left(\frac{n_A^2 \delta_{AA} + 2n_A n_B \delta_{AB} + n_B^2 \delta_{BB}}{(n_A + n_B)^2}\right)$$

$$- \frac{(n_A + n_B)n_C}{n_A + n_B + n_C}\delta_{CC} \qquad (1.22)$$

Székely and Rizzo (2005) then simplify the second term, by using Equation 1.18:

$$e_M(A, B) = \frac{n_A n_B}{n_A + n_B}(2\delta_{AB} - \delta_{AA} - \delta_{BB})$$

$$\therefore 2\delta_{AB} = \frac{n_A + n_B}{n_A n_B} e_M(A, B) + \delta_{AA} + \delta_{BB}$$

Now:

$$-\frac{(n_A+\ n_B)n_C}{n_A+\ n_B+\ n_C}\left(\frac{n_A^2\delta_{AA}+2n_An_B\delta_{AB}+\ n_B^2\delta_{BB}}{(n_A+n_B)^2}\right)$$

$$=\ -\frac{n_C}{n_A+\ n_B+\ n_C}\left(\frac{n_A^2\delta_{AA}+n_An_B(\frac{n_A+n_B}{n_An_B}e_M(A,B)+\ \delta_{AA}+\ \delta_{BB})+\ n_B^2\delta_{BB}}{(n_A+\ n_B)}\right)$$

$$=\ -\frac{n_C}{n_A+\ n_B+\ n_C}\left(\frac{n_A^2\delta_{AA}+(n_A+\ n_B)e_M(A,B)+\ n_An_B\delta_{AA}+\ n_An_B\delta_{BB}+\ n_B^2\delta_{BB}}{(n_A+\ n_B)}\right)$$

$$=\ -\frac{n_C}{n_A+\ n_B+\ n_C}\left(\frac{n_A(n_A+\ n_B)\delta_{AA}+(n_A+\ n_B)e_M(A,B)+\ \ n_B(n_A+n_B)\delta_{BB}}{(n_A+\ n_B)}\right)$$

$$=\frac{1}{n_A+\ n_B+\ n_C}(-\ n_An_C\delta_{AA}-n_Ce_M(A,B)-\ \ n_Bn_C\delta_{BB})$$

This is substituted into $e_M(A\cup B,C)$ in Equation 1.23:

$$e_M(A\cup B,C)$$

$$=\ \frac{(n_A+\ n_B)n_C}{n_A+\ n_B+\ n_C}\left(\frac{2n_An_C\delta_{AC}+\ 2n_Bn_C\delta_{BC}}{(n_A+n_B)n_C}\right)$$

$$+\frac{1}{n_A+\ n_B+\ n_C}(-\ n_An_C\delta_{AA}-n_Ce_M(A,B)+\ \ n_Bn_C\delta_{BB})$$

$$-\frac{(n_A+\ n_B)n_C}{n_A+\ n_B+\ n_C}\delta_{CC}$$

$$=\ \frac{1}{n_A+\ n_B+\ n_C}[(2n_An_C\delta_{AC}+\ 2n_Bn_C\delta_{BC})-\ n_An_C\delta_{AA}$$

$$-\ n_Ce_M(A,B)-\ n_Bn_C\delta_{BB}-\ (n_A+n_B)n_C\delta_{CC}]$$

$$=\frac{1}{n_A+\ n_B+\ n_C}[(n_An_C)(2\delta_{AC}-\ \delta_{AA}-\delta_{CC})$$

$$+\ (n_Bn_C)(2\delta_{BC}-\delta_{BB}-\delta_{CC})-\ n_Ce_M(A,B)]$$

$$\therefore\ e_M(A \cup B, C)$$

$$= \frac{(n_A + n_C)}{n_A + n_B + n_C} e_M(A, C)$$

$$+ \frac{(n_B + n_C)}{n_A + n_B + n_C} e_M(B, C) \qquad (1.23)$$

$$- \frac{n_C}{n_A + n_B + n_C} e_M(A, B)$$

We have now shown that the proof used by Székely and Rizzo (2005) also holds when using $d_M(\mathbf{a}_i, \mathbf{b}_j)$, a non-Euclidean distance. The same Lance-Williams parameters as used in Ward's minimum variance method therefore also apply to this least absolute error version of Ward's method. We can therefore continue using Ward's method while we are using the Manhattan metric. We are now able to construct dendrograms to graphically show the clustering.

2.11 Dendrograms

A dendrogram graphically represents the results obtained from performing cluster analysis and is similar to the phylogenetic tree constructed by Turchi and Cristianini (2006). A dendrogram "shows all the steps in the hierarchical procedure, including the distances at which clusters are merged" (Rencher, 2002:456). Our trees will be unrooted. This means that we make no assumptions on the evolutionary ancestry of the languages. We let the data speak for itself; we are purely interested in the relative proximity between pairs of languages.

We perform a cluster analysis, using Ward's least absolute error method as clustering algorithm. This is done in MATLAB (MATLAB, 2010) by using the "linkage" function and specifying the option "ward". The linkage function using Ward's method applies the Lance-Williams updating function. Since we have shown that the Lance-Williams parameters are

the same for the least absolute error method, the use of this function in MATLAB is justified. We then construct a dendrogram by using the “dendrogram” function in MATLAB.

3. Results/Application

3.1 Statistical Language Signature (SLS)

We obtained a frequency table for all di-grams in the file ‘English.txt’. Table 2 provides an excerpt of this table.

Table 2. Frequency table of di-grams for English

	‘a’	*‘b’*	*‘c’*	*‘d’*	*‘e’*	*‘f’*	*‘g’*	*‘h’*	*‘i’*	…	*‘s’*	*‘t’*	*‘u’*	…	*‘y’*	*‘z’*	…	*‘_’*
‘a’	0	8	30	9	0	1	16	0	13	…	68	92	2	…	7	0	…	20
‘b’	5	0	0	0	52	0	0	0	7	…	3	0	2	…	13	0	…	0
‘c’	19	0	5	0	43	0	0	32	30	…	0	38	10	…	1	0	…	12
‘d	12	0	0	0	45	0	2	1	34	…	3	0	15	…	1	0	…	181
‘e’	32	1	51	84	38	12	3	0	13	…	66	21	0	…	3	0	…	363
‘f’	11	0	0	0	16	9	1	0	4	…	0	0	16	…	0	0	…	97
‘g’	12	0	0	0	22	0	0	61	8	…	3	1	4	…	0	0	…	33
‘h’	75	0	0	0	174	0	0	0	60	…	0	56	14	…	1	0	…	36
‘i’	22	6	66	6	25	8	71	1	0	…	69	92	0	…	0	4	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮
‘s’	11	0	9	2	47	0	0	33	15	…	24	46	16	…	1	0	…	214
‘t	35	0	0	0	75	0	0	199	161	…	33	5	10	…	37	0	…	121

'u'	17	9	14	6	3	1	4	0	6	…	11	16	0	…	0	0	…	0
'y'	0	0	0	0	0	0	0	0	1	…	1	0	0	…	0	0	…	127
'z'	3	0	0	0	1	0	0	0	0	…	0	0	0	…	0	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮
'_'	270	68	57	43	93	89	16	92	100	…	95	256	20	…	0	0	…	0

As expected, the frequency of the di-gram "th" in the English text document is relatively large. Words ending on "e" occur 363 times, which makes this di-gram "e_" the most common occurrence in the English text document. Words starting with "a" or "t" are also relatively common with 270 and 256 occurrences, respectively. A portion of the SLS Matrix RF, with cells defined by Equation 1.1, is shown in Table 3.

Table 3. Statistical Language Signature matrix table of relative frequencies of di-grams for English[1]

	'a'	*'b'*	*'c'*	*'d'*	*'e'*	*'f'*	*'g'*	*'h'*	*'i'*	…	*'s'*	*'t'*	*'u'*	…	*'y'*	*'z'*	…	*'_'*
'a'	0	0.08	0.29	0.09	0	0.01	0.15	0	0.13	…	0.65	0.89	0.02	…	0	0.07	…	0.19
'b'	0.05	0	0	0	0.5	0	0	0	0.07	…	0.03	0	0.02	…	0	0.13	…	0
'c'	0.18	0	0.05	0	0.41	0	0	0.31	0.29	…	0	0.37	0.1	…	0	0.01	…	0.12
'd	0.12	0	0	0	0.43	0	0.02	0.01	0.33	…	0.03	0	0.14	…	0	0.01	…	1.74
'e'	0.31	0.01	0.49	0.81	0.37	0.12	0.03	0	0.13	…	0.64	0.2	0	…	0	0.03	…	3.5

[1]It is worth noting that the entries of both the SLS matrix and the SLS vector have been multiplied by 100, and rounded to 2 decimals for presentation purposes.

'f'	0.11	0	0	0	0.15	0.09	0.01	0	0.04	⋯	0	0	0.15	⋯	0	0	⋯	0.93
'g'	0.12	0	0	0	0.21	0	0	0.59	0.08	⋯	0.03	0.01	0.04	⋯	0	0	⋯	0.32
'h'	0.72	0	0	0	1.68	0	0	0	0.58	⋯	0	0.54	0.13	⋯	0	0.01	⋯	0.35
'i'	0.21	0.06	0.64	0.06	0.24	0.08	0.68	0.01	0	⋯	0.66	0.89	0	⋯	0	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮
's'	0.11	0	0.09	0.02	0.45	0	0	0.32	0.14	⋯	0.23	0.44	0.15	⋯	0	0.01	⋯	2.06
't	0.34	0	0	0	0.72	0	0	1.92	1.55	⋯	0.32	0.05	0.1	⋯	0	0.36	⋯	1.17
'u'	0.16	0.09	0.13	0.06	0.03	0.01	0.04	0	0.06	⋯	0.11	0.15	0	⋯	0	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮
'y'	0	0	0	0	0	0	0	0	0.01	⋯	0.01	0	0	⋯	0	0	⋯	1.22
'z'	0.03	0	0	0	0.01	0	0	0	0	⋯	0	0	0	⋯	0	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮	⋮	⋱	⋮
'_'	2.6	0.65	0.55	0.41	0.9	0.86	0.15	0.89	0.96	⋯	0.91	2.46	0.19	⋯	0	0	⋯	0

Since we describe our data as a vector in $p = 3600$ dimensions, we show a part of our SLS Vector for English as follows in Table 4.

Table 4. Statistical Language Signature vector for English

$\mathbf{a}_{English} =$

aa	ab	ac	ad	⋯	be	bf	bg	⋯	tg	th	ti	⋯	_a	_b	_c	_d	...	_ _
0	0.08	0.29	0.09	⋯	0.5	0	0	⋯	0	1.92	1.55	⋯	2.6	0.65	0.55	0.41	...	0

3.2 Distance Matrix

The distance between a pair of languages is obtained by calculating the Manhattan distance between the SLS vectors of those languages. The Manhattan distance matrix calculated for some of the languages analysed is shown in Table 5.

Table 5. Manhattan distance matrix for 9 Indo-European languages

	Afrikaans	Asturian	Bosnian	Breton	Catalan	Corsican	Czech	Danish	…	Serbian	…
Afrikaans	0.0000	0.2879	0.3571	0.2664	0.2757	0.3576	0.3683	0.1940		0.3564	
Asturian	0.2879	0.0000	0.3048	0.2878	0.1297	0.2556	0.3230	0.2920		0.3067	
Bosnian	0.3571	0.3048	0.0000	0.3449	0.3132	0.3017	0.2366	0.3592		0.0362	
Breton	0.2664	0.2878	0.3449	0.0000	0.2816	0.3418	0.3462	0.2698		0.3420	
Catalan	0.2757	0.1297	0.3132	0.2816	0.0000	0.2547	0.3410	0.2633		0.3126	
Corsican	0.3576	0.2556	0.3017	0.3418	0.2547	0.0000	0.3606	0.3468		0.3052	
Czech	0.3683	0.3230	0.2366	0.3462	0.3410	0.3606	0.0000	0.3715		0.2332	
Danish	0.1940	0.2920	0.3592	0.2698	0.2633	0.3468	0.3715	0.0000		0.3543	
…											
Serbian	0.3564	0.3067	0.0362	0.3420	0.3126	0.3052	0.2332	0.3543		0.0000	
…											

In order to discuss the difference in results when using the Euclidean distance rather than the Manhattan distance, we also provide the Euclidean distance matrix for the same 9 languages in Table 6.

Table 6. Euclidean distance matrix for 9 Indo-European languages

	Afrikaans	Asturian	Bosnian	Breton	Catalan	Corsican	Czech	Danish	…	Serbian	…
Afrikaans	0.0000	0.0941	0.1170	0.0873	0.0948	0.1316	0.1073	0.0690		0.1163	
Asturian	0.0941	0.0000	0.0937	0.0881	0.0484	0.1010	0.0880	0.0907		0.0928	
Bosnian	0.1170	0.0937	0.0000	0.1082	0.0943	0.0942	0.0727	0.1131		0.0141	
Breton	0.0873	0.0881	0.1082	0.0000	0.0843	0.1174	0.0965	0.0828		0.1072	
Catalan	0.0948	0.0484	0.0943	0.0843	0.0000	0.0991	0.0922	0.0842		0.0934	
Corsican	0.1316	0.1010	0.0942	0.1174	0.0991	0.0000	0.1046	0.1244		0.0931	

Czech	0.1073	0.0880	0.0727	0.0965	0.0922	0.1046	0.0000	0.1013		0.0706	
Danish	0.0690	0.0907	0.1131	0.0828	0.0842	0.1244	0.1013	0.0000		0.1119	
…											
Serbian	0.1163	0.0928	0.0141	0.1072	0.0934	0.0931	0.0706	0.1119		0.0000	
…											

In both of the complete distance matrices, the minimum distance was between Bosnian and Serbian:

$$D_E(\mathbf{a}_{Bosnian}, \mathbf{a}_{Serbian}) = 0.0141$$

$$D_M(\mathbf{a}_{Bosnian}, \mathbf{a}_{Serbian}) = 0.0362$$

This is where the first cluster will form, in both cases, regardless of the linkage method used. The new distance between the cluster C_{BS} (Consisting of Bosnian and Serbian) and any other point k will be updated by using the Lance-Williams updating algorithm, as defined in Equation 1.5:

$$d_{k(BS)} = \alpha_B d_{kB} + \alpha_S d_{kS} + \beta d_{BS} + \gamma |d_{Bk} - d_{Sk}|$$

We use Ward's linkage method and the values of the Lance-Williams parameters are as follows:

$$\alpha_B = \frac{n_B + n_k}{n_B + n_S + n_k}$$

$$\alpha_S = \frac{n_S + n_k}{n_B + n_S + n_k}$$

$$\beta = \frac{(-n_k)}{n_B + n_S + n_k}$$

$$\gamma = 0$$

We construct dendrograms for both objective functions: minimum variance and least absolute error (*i.e.* using the Euclidean and the Manhattan distance, respectively.

3.3 Dendrogram

Our dendrogram shows the different sub-groups of the Indo-European language family and use languages from the following subdivisions: Baltic, Slavic (Balto-Slavic), Celtic, Germanic and Romance. The phylogenetic classification of the languages we analyse is as follows:

- Slavic: Bosnian, Serbian, Slovenian, Slovak, Czech, Polish
- Baltic: Lithuanian, Latvian
- Celtic: Breton, Welsh, Scottish, Irish
- Germanic: English, Icelandic, Swedish, Danish, Norwegian, Luxembourgish, Frisian, Dutch, Afrikaans, German
- Romance: French, Italian, Spanish, Asturian, Catalan, Friulian, Galician, Portuguese, Corsican, Romanian

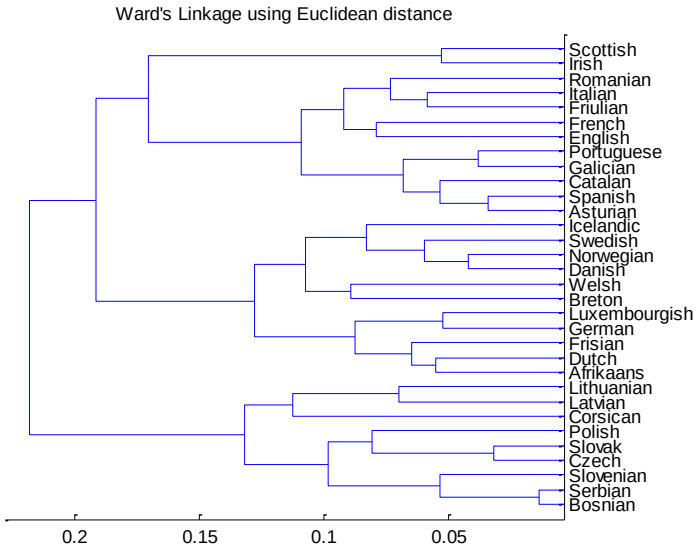
An overview of the dendrogram results yielded by other linkage methods suggested by Rencher (2002:456-471) follows. The centroid and median methods produce dendrograms containing reversals. This is not appropriate for our analysis; we are interested only in monotonic linkage algorithms. With its propensity to 'chain' the single linkage method produces large, elongated clusters with languages such as Corsican and Afrikaans on opposite sides of the same cluster.

The complete linkage method produces results that were more acceptable. The only concerns were the misclassification of Breton and English, and that the Balto-Slavic cluster is never formed. The average linkage provides an improvement on the complete linkage

method in such a way that the Balto-Slavic cluster is formed in this case. Nevertheless, the problem regarding the misclassification of English and Breton is still present.

Ward's method is considered the most appropriate for our analysis. Although both the average linkage function and Ward's method are known to be monotonic and space-conserving, the average linkage function only considers distances between clusters and disregards the importance of intra-cluster similarity. We provide the dendrogram resulting from clustering with Ward's linkage using Euclidean distances in Figure 1. In Figure 2 we provide a dendrogram resulting from clustering with Ward's linkage using Manhattan distances.

Figure 1. Clustering with Ward's Linkage using Euclidean distance

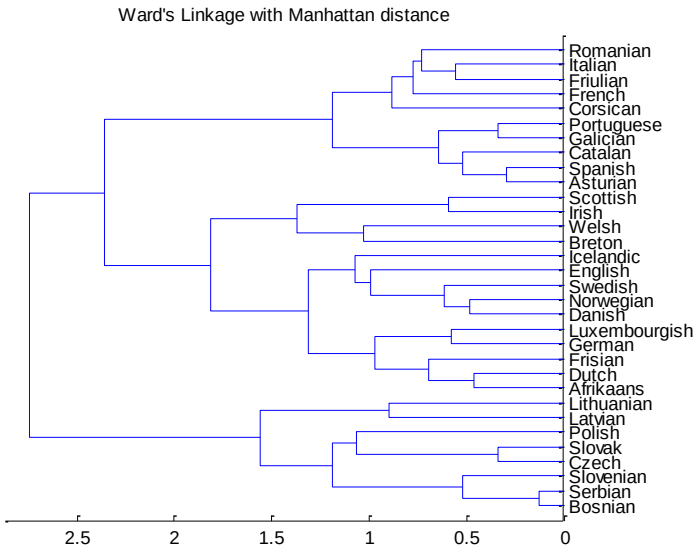


We note that using the Euclidean distance matrix with Ward's method has a few drawbacks:

- Although the Baltic and Slavic languages are clustered together to obtain the Balto-Slavic language cluster, we observe Corsican, a Romance language (Paul, 2009) clustered with them.

- Breton and Welsh, two Celtic languages (Paul, 2009) are clustered with the Germanic languages. They only join with Scottish and Irish after these two have been clustered with the Romance languages.
- English is clustered with the Romance languages, an error which we have encountered with the average linkage method.

Figure 2. Clustering with Ward's Linkage using Manhattan distance



Ward's method of linkage with the Manhattan distance yields similar results to the average linkage method with Manhattan distances. There is one important difference in the results of these two methods. English and Breton, misclassified by the average linkage method, are now correctly clustered with the Germanic and Celtic language groups, respectively.

The clustering can also be considered as more or less intuitive from a linguistic and sometimes even geographical point of view. An example of this is that the Noridic Languages or Scandinavian Germanic Languages (Danish, Norwegian and Swedish) are clustered together, before they are joined with the West Germanic (*e.g.* Afrikaans, Dutch and German). Using Ward's method of linkage provides us with 5 distinct logical clusters, that

also make sense phylogenetically: Baltic Languages, Slavic Languages, Romance Languages, Germanic Languages and Celtic Languages.

When using Ward's method with the Manhattan distance, there seems to be a misclassification of English and Icelandic. English is joined with the Nordic Languages before Icelandic is, where Icelandic is also regarded as Scandinavian Germanic. English belongs with the West Germanic languages as it forms part of the Anglo-Frisian language family. Comparing this minor drawback of this method to the disadvantages of the other clustering methods, however, we see that this method produces the best results. Furthermore, if we are interested only in the four subdivisions of the Indo-European language family, using this method of clustering is sufficient.

The Manhattan distance yielded the best results when used in conjuction with Ward's method. Ward's method using the Manhattan distance produced the most accurate results. With this clustering algorithm, four distinct clusters were found. These clusters also have significant linguistic meaning. The results from this method were intuitive and logical.

4. Conclusion

This paper investigated the use of phenetic methods in language classification and classified languages solely based on content similarity Phenetic classification in the form of hierarchical cluster analysis clustered similar languages together, without assuming similar roots. We used phenetic classification methods and showed that the phylogenetic properties inherent to the languages were clearly reflected in our results.

The Manhattan distance method yielded the most accurate results when using this type of distance measurement. Ward's method was the best linkage method, because both inter- and intra-cluster distances were incorporated.

An important component of this research was the extension of Ward's method for use with a 1-norm distance such as the Manhattan distance. The use of Ward's method can be expanded to include Manhattan distances. Use of another objective function (in this case least absolute error) with Ward's function produces more accurate results than using Ward's method with the Euclidean distance metric in our situation.

This research could be extended beyond Indo-European languages and applied to other language families. Another possibility for further research is the use of tri-gram frequencies to form the SLS. This would mean having a three-dimensional SLS. The Manhattan metric could still be used in this case, as well as Ward's method of linkage within the hierarchical clustering process, making this adaptation a natural extension of this research.

Ultimately, we established that grouping of languages, similarities between languages and language traits well known in the field of linguistics can be extracted or independently observed using unsupervised machine learning techniques. It is indeed possible to autonomously classify languages without any prior linguistic knowledge or assumptions.

References

BENEDETTO, D., CAGLIOTI, E., and LORETO, V., (2002), "Language Trees and Zipping," *Physical Review Letters*, *88*(4), 048702.

BOYCE, A.J., (1964), "The Value of Some Methods of Numerical Taxonomy with Reference to Hominoid Classification," in *Phenetic and Phylogenetic Classification: A Symposium*, eds. V.H. Heywood, and J. McNeill, Systematics Association Publication no. 6, 47-65.

BURGMAN, M.A., and SOKAL, R.R., (1989), "Factors Affecting the Character Stability of Classifications," *Plant Systematics and Evolution*, *167*, 59-68.

CAIN, A.J., and HARRISON, G.A., (1960), "Phyletic Weighting," *Proceedings of the Zoological Society of London, 135,* 1-31.

CHEN, Z., and VAN NESS, J.W., (1996), "Space-Conserving Agglomerative Algorithms," *Journal of Classification*, *13,* 157-168.

CORMACK, R.M., (1971). "A Review of Classification," *Journal of the Royal Statistical Society. Series A (General)*, *134*(3), 321-367.

FARRIS, J.S., (1972), "Estimating Phylogenetic Trees from Distance Matrices, *The American Naturalist*, *106*(951), 645-668.

FERRER I CANCHO, R.F.", and SOLÉ, R.V., (2003), "Least Effort and the Origins of Scaling in Human Language," *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 788-791.

FITCH, W.M., and MARGOLIASH, E., (1967), "Construction of Phylogenetic Trees," *Science*, *155*(3760), 279-284.

LANCE, G.N., and WILLIAMS, W.T., (1966), "A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems," *The Computer Journal*, *9*(4), 373-380.

PAUL, L.M. (Editor), (2009), *Ethnologue: Languages of the World*, 16th Edition, Dallas: SIL International. Online version: http://www.ethnologue.com.

MANTEGNA, R.N., BULDYREV, S.V., GOLDBERGER, A.L., HAVLIN, S., PENG, C.K., SIMONS, M., and STANLEY, H.E., (1994), "Linguistic Features of Noncoding DNA Sequences," *Physical Review Letters 73*, 3169-3172.

MATLAB, (2010), *Version 7.10*, Natick, MA: The Mathworks Inc.

MILLIGAN, G.W., (1979), "Ultrametric Hierarchical Clustering Algorithms, *Psychometrika, 44*(3)", 343-346.

RENCHER, A.C., (2002), *Methods of Multivariate Analysis,* 2nd Edition, New York: John Wiley & Sons.

SCHLEICHER, A., (1848), "*Zurvergleichenden Sprachengeschichte"* (Towards a comparative history of languages), Bonn: H.B. König.

SCHLEICHER, A., (1863), "*Die Darwinsche Theorie und die Sprachwissenschaft"* (Darwinian theory and the science of language), Wiemar: Böhlau.

SHANNON, C.E., (1951), "Prediction and Entropy of Printed English," *Bell System Technical Journal, 30*(1), 50-64.

SOKAL, R.R., and SNEATH, P.H.A., (1963), *Principles of Numerical Taxonomy,* San Francisco: W.H Freeman and Company.

SOKAL, R.R., (1986), “Phenetic Taxonomy: Theory and Methods,” *Annual Review of Ecology and Systematics, 17*, 423-442.

SZÉKELY, G.J., and RIZZO, M.L., (2005), “Hierarchical Clustering via Joint Between-Within Distances: Extending Ward’s Minimum Variance Method,” *Journal of Classification, 22*, 151-183.

TAUB, L., (1993), “Evolutionary Ideas and 'Empirical' Methods: The Analogy between Language and Species in Works by Lyell and Schleicher,” *The British Journal for the History of Science*, *26*(2), 171-193.

THEODORIDIS, S., and KOUTROUMBAS, K., (2003), *Pattern Recognition,* 2nd Edition, London: Academic Press.

TURCHI, M., and CRISTIANINI, N., (2006), “A Statistical Analysis of Language Evolution,” *Proceedings of the 6th International Conference on the Evolution of Language*, 348-355.

VOGT, W., and NAGEL, D., (1992), “Cluster Analysis in Diagnosis,” *Clinical Chemistry*, *38*(2), 182--198.

WARD JR., J.H., (1963), “Hierarchical Grouping to Optimize an Objective Function,” *Journal of the American Statistical Association*, *58*(301), 236-244.

WARNOW, T., (1997), “Mathematical Approaches to Comparative Linguistics,” *Proceedings of the National Academy of Sciences of the United States of America, 94*, 6585-6590.

UNITED NATIONS GENERAL ASSEMBLY, (1948), *Universal Declaration of Human Rights, General Assembly Resolution 217 A (III)*, United Nations, Paris.