Spring 4-23-2020

An Algorithm For Building Language Superfamilies Using Swadesh Lists

Bill Mutabazi

University of Nebraska - Lincoln, mutabazi@huskers.unl.edu

AN ALGORITHM FOR BUILDING LANGUAGE SUPERFAMILES USING SWADESH LISTS

By

Bill Jean Claudien Mutabazi

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Peter Revesz

Lincoln, Nebraska

May 2020

AN ALGORITHM FOR BUILDING LANGUAGE SUPERFAMILES USING SWADESH LIST

Bill Jean Claudien Mutabazi, M.S.

University of Nebraska, 2020

Adviser: Peter Revesz

To make sense of language originality and their evolution over the years is a daunting task. Most of scientific studies had been based on an asymmetric study of word convergence between different languages, borrowing words and common origin word meanings to define their linguistic family classification. This thesis presents an efficient algorithm for classifying language family based on cognate words of different languages. We used the Swadesh list-based database of various languages from different language family as a test case of words in a corpus. We use an agglomerative "bottom-up" hierarchical clustering methods to identify the interrelatedness of four different languages families (Afroasiatic, Bantu, Indo-European, and Uralic language family). Our study applies a comparative statistical methodology and a computational data analysis algorithm to quantitively identify the relatedness of ancient original

words and generate a phylogenetic tree of their relatedness based on Swadesh list cognate words.

DEDICATION

To My Loving Parents

ACKNOWLEDGMENTS

Firstly, I would like to thank wholeheartedly my advisor Dr. Revesz for his tremendous academic support. Without his guidance, support, and good nature, I would never have been able to pursue my research work in computational linguistics.

I also express my appreciation to my committee members, Dr. Samal and Dr. Riedesel, for their helpful comments and suggestions that can be further developed and improvised for bigger contribution of in the field of data mining, and computational linguistics. With a special mention to my colleagues, Colin Richards and Minal Khatri, thank you for your moral support and being there for me.

And finally, last but by no means least, I would like to dedicate this thesis to my lovely mum, dad, and my eternal cheerleader, late Best friend Rungano: I miss you. I am forever grateful for all of you have done for me to be at this stage. You are my whole world!

Thank you all for your encouragement!

Author

Bill Jean Claudien Mutabazi

Table of Contents

1 Introduction..... 1

1.1 List of contribution 4

2 Background..... 5

3 Research Problem..... 7

4 Data Source..... 10

5 Related work..... 14

6 Proposed New Method..... 18

7 Experimental Results..... 30

8 Conclusion..... 34

References I

Chapter 1

Introduction

Languages are a vital tool for communication of thoughts and ideas that builds friendships, economic relationships that preserves cultural ties in our societies from one generation to another. Archaeological evidences of complex group activities suggest that people have spoken languages for over 50,000 years, when modern humans started to disperse from Africa (Dimmendaal, 2007). Atkinson (Atkinson, 2003), found some evidence for the existence of a phonetically complex archaic language. In particular, Atkinson (Atkinson, 2003) found that the further humans travelled from Africa, the fewer number of different phonemes survived in various

languages. In this thesis, we investigate the possible evolution of four different language families, namely Afro-asiatic, Bantu, Indo-European, and Uralic from a common proto-language. For this purpose, we designed a preselection method to identify the interrelatedness of languages (Arabic, Finnish, German, Hungarian, Kinyarwanda, and Latin). We build their interrelatedness as a test case for a novel Family Tree Generation Algorithm.

We build a relative genetic classification of languages based on the percentages of cognate words. Cognate words have a common etymological origin (Atkinson, 2003). There are several alternative methods of studying and identifying cognate words, which involve data mining and neural data (Brown & Kass, 2014).

Cognate words are usually inherited from a shared parent language. In recent a study, Revesz (Revesz, 2018) identified a few cognate words that cut across language families in Africa and Eurasia. For example, Revesz (Revesz, 2018) identified *buda* to be an ancient cognate word. This name occurs as a mountain name in many places. There is a mountain name called *Buda* near Lake Victoria, which is a source of the Nile River, in Burundi in Africa as well as in Hungary in Europe. In fact, Buda Mountains are adjacent

to Budapest, the capital city of Hungary. Since Hungarian is a Uralic language (Revesz, 2017), the existence of Bantu and Uralic words that are apparently cognate is very interesting and surprising.

Previously the proposed Nostratic superfamily tried to link several Eurasian language families but did not include any from Africa (Ringe, 1995). In addition, Revesz (Revesz, 2019) proposed to add the Euphratic language, a Proto-Sumerian language to the Minoan Uralic language. Though, the analysis shows that the Minoan genes are composed of two originally distinct groups, the analysis raised the possibility of finding additional cognates between Bantu and the Eurasian language families (Revesz, 2017).

Computational linguistics often overlaps with the field of natural language processing because they share many common cognates. While natural language processing focuses on the tokens/tags and uses them as predictors in machine learning models, computational linguistics digs deeper into the relationships and links among them. Our approach is to mine data, to identify a computer-based preselection methods using the existing Swadesh lists and various online dictionaries, as aligned sequences of cognate words of different languages families and apply sequentially an improved unweighted

pair-group with arithmetic mean (UPGMA) clustering method (Hua, et al., 2017). The UPGMA method identifies the languages to output a phylogenetic tree that reflects the evolution of these languages. Therefore, we determine how closely these languages are related and their descendancy from a common protolanguage. This brings the previously disparate language families into a common superfamily.

1.1 List of Contributions

The main contributions of this thesis are the following:

i. Developing an algorithm to generate language families and superfamilies given for each input language a Swadesh list represented using the international phonetic alphabet (IPA) notation.

ii. The algorithm is novel in using the Levenshtein distance metric on the IPA representation and in the way it measures overall distance between pairs of Swadesh lists.

iii. Building a Swadesh list for the author's native Kinyarwanda language because a Swadesh list could not be found even after an extensive search for it.

Chapter 2

Background

In this section, we provide background information related to the data mining algorithms and the computational linguistics techniques used in this work, and motivating example that demonstrate the hierarchical approach that we used. In grouping cognate words from four different language families. A language family is a group of languages with a common ancestor. This common ancestor is referred to as a protolanguage.

The protolanguages are believed to have split up into two or more dialects, which gradually became more and more different from each other. For various reasons it is not possible to be precise about the number of languages in the world, but most linguists

agree that there are between 6,000 - 7,000 living languages. These languages are divided into about 10 major language families. The exact number is dependent on the classification paradigm. By classification paradigm, the alternative ways of classifying languages are according to their genealogy (genetic classification) or according to their linguistic features (typological classification). In this work, we use a computational linguistic based method to study the interrelatedness of Afro-asiatic, Bantu, Indo-European, and Uralic language families.

The widely known word interrelated study, lexicostatistics, analyses the proportion of shared words between languages and treats the proportion of shared words as a similarity measure (Hinkka, 2018). Within the lexicon of any language there exists a particular section that may be called “basic” or “stable”, so that it is possible to provide a list of meanings which in any language of the world will be represented by words from this section (the so-called “Swadesh list”, consisting of 200 items in its large version and of 100 items in its “compressed” version, represents an approximate, somewhat idealized version of this part of the lexicon). Swadesh list are words of basic vocabulary used in lexicostatistical studies to identify comparable approximate number of cognate words present in the words making up the list.

Chapter 3

Research Problem

To determine the interrelatedness of words by their meaning, sound, and lexicon similarity we use both qualitative and quantitative approaches. We assume that common words in languages are maintained at a definite rate, i.e., some parts of the vocabulary are much less subject to change than other parts. There are tendencies that make it possible to determine language relationships based on data mining techniques and Swadesh lists. The data mining identifies the cognate words originality. Cognates between languages usually have similarities in pronunciation and meaning but not necessarily spelling. In this work, we identify a new method to generate language family tree based on their string distances to identify their cognates. We considered the following:

i. The international phonetic alphabet (IPA) notation word which is a devise a system for transcribing the sounds form of a given speech.
ii. Word initial, medial and final lexicon (meaningful units) position in the given words
iii. The presence of same phoneme pairs, prefixes, and indexes (diphthongs).
iv. Sounds do not change randomly but regularly! two or more languages which are related will show regular sound correspondences. That means that two languages are related if there is a consistent of regularity of sound change between interrelated words of those languages.

The regularity of sound change implies that when a certain sound X changes to a slightly different sound X' in one word, the same change tends to take place in all words where sound X occurs, in all words where sound X occurs in a similar context. The regularity of sound change is the prerequisite for the comparative method. Because the sound changes from the protolanguage to its descendants regularly, there are also regular sound correspondences between languages with a common protolanguage (If regular sound correspondences can be established between two or more languages,

these languages are genetically related, that is, they belong to the same language family and are descendants of the same protolanguage).

Based on these rules, we adopted a methodology from our paper “A Quantitative Lexicostatistics Study of the Evolution of the Bantu Language Family” (Mutabazi & Revesz, 2019). For Studying languages that shares lexicon by position (languages of the same family) using Hamming distance, that was found to be inefficient compared to quantitatively the Levenshtein distance. The Hamming distance based only on the number of positions with same symbol in both strings compared, while the Levenshtein distance calculates the minimal number of insertions, deletions and replacements needed for transforming string X into string X’ of words. We comparatively identify the Hamming distance pitfalls to Levenshtein distance in our Section V. Since, by only considering the distance there might be many false cognates preselected, we identify an Adjusted Cognate Distance Score (Adj_Score) to identify a firm range of one word interrelatedness to another based on their phonetic sound similarity adjusted by the length of the given word.

Chapter 4

Data Source

A major data source is the online Global Lexicostatistical Database (GLD), a hierarchical system of wordlists organized from bottom to top. GLD classifies annotated Swadesh 100-compressed word list data in various families (Starostin, 2016). We identified Swadesh list cognate words for Afroasiatic's Arabic; Bantu's Swahili and Kinyarwanda; Indo-European's German and Latin; Uralic languages Finnish and Hungarian. Although Bantu languages' Kinyarwanda and Swahili are least documented, we identify IPA notation of the identified Swadesh words using knowledge as a native speaker. Our database contains fields for various word features such as a word's grammatical form

describing whether it is an adjective, adverb, noun, verb, cardinal numbers, conjunction, preposition, pronoun, verb, etc. Our dataset contains cognate words that are identified by at least these essential properties:

i. They are always structural units.

ii. They are words that have a similar but not necessarily identical meaning.

iii. They always share a formal resemblance.

We first comparatively study the structural and syntactic IPA notation similarities between all the languages. We will explain this in the latter sections. We focus on non-previously studied Bantu Languages. Table on Fig.2 below shows the high-level illustration of languages used from each language family studied; we collected data fields of 200-Swadesh words in each of the selected languages that are comparatively studied across languages of Kinyarwanda which is a tonal Bantu language spoken mostly in the country of Rwanda (Habumuremyi, 2006), Swahili which is another Bantu language widely used as a *lingua franca* in Eastern Africa and having official status in several countries. This database also has Arabic as representation of Afroasiatic language, German and Latin for the Indo-European family and Hungarian and Finnish for Uralic language family.

	BA		AF	IE		UR	
English	**Kinyarwanda**	**Swahili**	**Arabic**	**German**	**Latin**	**Hungarian**	**Finnish**
I	jewe	ˈmɪmɪ	ˈænæ	ic	ˈe.goː	ˈeːn	ˈminæ
you	woːwe	wewe	ˈentæ	ziː	tuː	ˈtiː	ˈsinæ
he	we-e	yeye	ˈhowwæ	eːɐ̯	id	ˈøː	ˈse
we	tue	sisi	ˈeɦnæ	viːɐ̯	noːs	ˈmiː	ˈme
you	muewe	ninyi	ˈento	iːɐ̯	woːs	ˈtiː	ˈte
they	bo	wao	ˈhommæ	ziː	ˈe.aeˌ	ˈøːk	ˈhe
this	iki	huyu	doːl	diːs	hik	ˈɛz	ˈtæmæ
that	iki'o	ˈl̪ʲɪlʲə	doːlæt	das	ˈil.le	ˈɒz	ˈse
here	ha^ano	hapa	ˈhenæ	hiːɐ̯	hik	ˈit	ˈtæːlːæ
there	haa-riya	ˈpa.lɛ	heˈnæːk	daː	ˈil.lik	ˈot	ˈtuo̯lːɑ
who	indê	ˈnʌni	miːn	veːɐ̯	kwis	ˈkiː	ˈkukɑ
what	ikie	ˈnini	ʔeːh	vas	kwid	ˈmiː	ˈmikæ
where	hêhê	wapi	feːn	voː	ˈu.bi	ˈɦɔl	ˈmisːæ

Table 1.: An illustration of our Database IPA notation of selected languages in each language family. (where the language Families Bantu language, Afro-Asiatic, Indo-European and Uralic as BA, AF, IE, and UR respectively) [Wik17].

Words are similar semantically if they are used in the same context and same type of each other (Gomaa, et al., 2013). Hymes (Hymes, 1960) holds that "lexicostatistics is not a short-cut, it does not replace other methods and information, but must be incorporated with them into a consistent body of knowledge". This is why our study of interrelatedness incorporates both statistics and data mining techniques to lexicostatistic to affirm more why word might be spoken almost exactly the same way in a small village in Africa as well as another village in Europe.

Chapter 5

Related Work

Linguistically, the major reason for the systematic comparison of languages is the desire to establish their relationships. This is to determine what languages have descended from a common protolanguage and how closely these languages are related. The similarity of words can either be semantical lexical sequence. Language models have been used to calculate language distances before, but these studies have been done on text (Hinkka, 2018). In earlier studies the text has commonly been normalized to some kind of a simpler Latin alphabet by removing at least some letter diacritics to improve language modeling performance (Batagelj, et al., 1992).

The motivation behind this is the assumption that removing diacritics makes the languages comparable and does not significantly change the meaning of the letters in terms of phonology. This may make sense in cases where the identity of the letter does not change when diacritics are removed, but it presents a new set of problems when the identity of the letter is tied to the diacritic or word stress (Hinkka, 2018). In that, most of the past studies have been based on qualitative study of one language (Gamallo, et al., 2007). We believe that both qualitative and quantitative study closes a big gap in identifying relativeness of common words across different language families.

A. Lexicostatistics using Neural Data

From a natural language processing (NLP) point view, measuring the lexical similarity between words, sentences, paragraphs and documents is an important component in various tasks such as information retrieval, document clustering, word-sense disambiguation, automatic essay scoring, short answer grading, machine translation and text summarization. Lexicostatistics works on the assumption that in related languages a part of the vocabulary is regarded basic. A test-list of meanings are sampled from the basic vocabulary. The common, everyday equivalents for this list are obtained from various languages and the degree of their relationships is quantified. This can help

to establish the relationship between languages, to classify related languages, and to establish the times at which related languages began to diverge. Based on NLP techniques, the interrelatedness of words can be studied by three approaches; String-based, Corpus-based, and Knowledge-based similarities (Gomaa, et al., 2013).

Daggumati and Revesz (Daggumati & Revesz, 2018) aids decipherment efforts of the index valley scripts by finding a similar but already known script with which the unknown symbols could be matched. In that their approach uses neural networks and two known scripts to find tentative phonetic assignments to the Indus Valley script symbols based on the Phoenician alphabets and the Brahmi syllabary script.

B. Past Statistical and Mathematical Work to Study Word Similarity

To find originality and similarity of words, archaeologists dig up documents that no modern person can read, use decipherment methods to find written characters that are familiar (say, the Phoenician alphabet), but the language is unknown. Other times, it is the reverse: The written script is unfamiliar, but the language is known. It may also happen that both script and language is unknown.

To construct word similarities in Meroitic language as an aid to decipherment, Smith (Smith, 2018) uses a mathematical approach of creating an alphabetical index of Meroitic and also comparing Meroitic words to possible cognates in Nubian or other known ancient and modern languages from the region. In his work he analyzed many of the longest texts by ranking words according to frequencies, to verify whether the current texts we have follow the mathematical relation known as Zipf's Law, where the word frequencies vary with their respective rank.

Chapter 6

Proposed New Method

For the context of this research, an embedded study takes advantage of both qualitative and quantitative approaches to identify cognate words in which are comparatively studied to identify to how the languages families are interrelated. Qualitatively, we used a Swadesh list-based database of all aspect types of words. We also use the international phonetic alphabet rules (IPA rules) to identify how the word are related by now only their lexicon block but their phonetics. We use online dictionaries to identify more of word meaning in the least documented languages. Quantitatively we use the Levenshtein distance approach mainly as a function in our algorithm to calculate a

string metric (pairwise string alignments) for measuring the difference between two lexicons sequence (row score). We used the following methodology:

Step 1: We translate each of the 200 collected word of Swadesh list for all seven language from four different language family to its IPA notation.

Step 2: We calculate the edit distance. This follows identifying the similar phonemes, as shown table 2 below, to what are set to a half metric distance compare to other lexicon (vocal, or sound) that are very different. Thus, they are identified by linguist to have similar tongue placement, pitch, length, and almost same voicing by their mental grammar.

Similar phoneme pairs
/b/ and /p/
/d/ and /t/
/g/ and /k/
/f/ and /v/
/s/ and /z/
Short and long vowels.
/w/ and /v/
/k/ and /h/ at the beginning of words

Table 2.: Similar phoneme pairs taken to be less distance than between other pairs of phonemes

Levenshtein distance is a simple metric which can be an effective string comparison tool. For preselection purpose, our algorithm compares each word to its corresponding word of the different family. Though for some languages i.e.: Latin, German most of Swadesh list word's suffix ends are voided for metric distance calculation, while some prefixes in Kinyarwanda are voided. Therefore, these morphemes added at the end or in front of a word form a derivative and does not identify the originality of the words. Fig.1 below illustrates these morphemes voided.

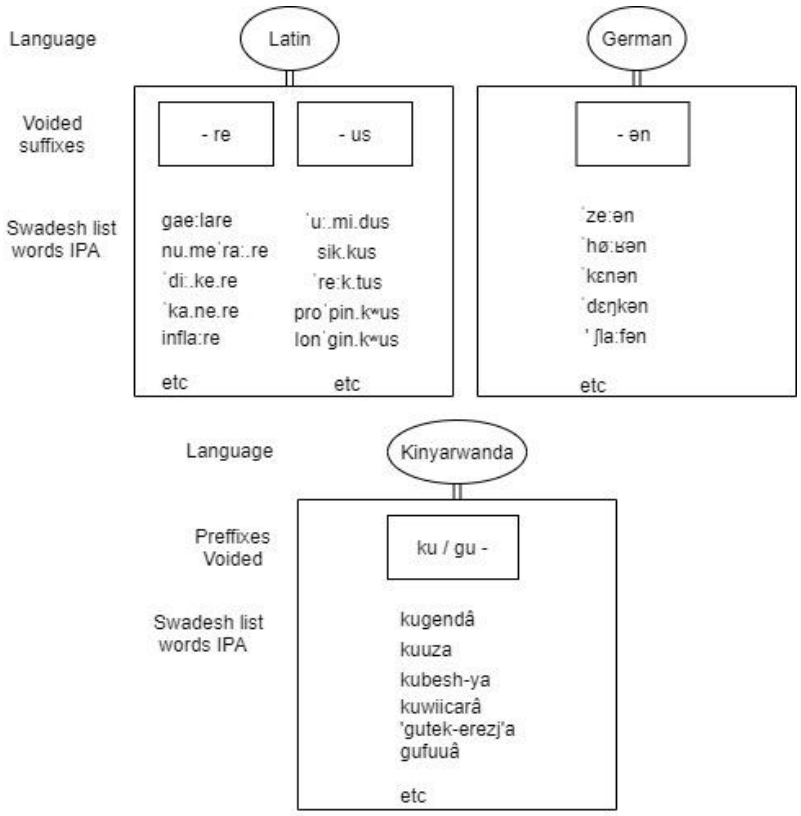

Fig.1: Illustration of suffixes and prefixes voided for distance metric calculation in each language.

A. Cognate string distance

String distance as a metric that measures the inverse similarity or the matching in the sequence of lexicon and characters from one cognate X word to another X' of the same meaning in a different language family previously used famous technique based on calculating the Hamming distance.

The Levenshtein distance counts the minimal number of substitutions needed to edit one string into another of equal length. Therefore, many of the words to be compared do not have the same length, using Levenshtein distance creates padding that generates space bits around an element's content which does not support another lexicon to be considered. Other technique includes Jaro-Winkler distance, a string-edit distance that gives a floating-point response in [0,1] where 0 represents two completely dissimilar strings and 1 represents identical strings.

We use the Levenshtein distance to calculate the distance of interrelatedness between cognate. The Levenshtein distance (Levenshtein, 1966) counts the minimal number of substitutions, insertions, and deletions to edit one string into another of any length.

Mathematically, the Levenshtein distance between two strings, a and b (of length |a| and |b| respectively), is given by lev a, b (|a|, |b|) where:

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Our approach uses the Levenshtein distance (LV.D) to compensate different sample sizes and number of total positions. Thus, the normalized function takes summations of all cognates with differences in word length. we assign a standard cost of 1 metric distance to each of the edit operations (insertion, deletion, and substitution).

In that: **string dist = 1 – (distance /length)**

Where:

Length = max (length of source expression, length of destination expression)

Distance = min (number of insertions, deletions and substitutions required to match given word A's lexicon to word B).

A. 1.: Example: Finding the Levenshtein distance between two words found in the Swadesh list of different language: Let X be a string of IPA notation "ˈ*multa*" of the Latin Swadesh list words and Y be a string of the same word in Finnish that translates

"ˈ*montæ*". As shown in table 3. below the distance yield is 2.5 that is rounded to 2* instead of 3 as phoneme *"æ"* and *"a"* are of similar phoneme they differ of short and long vowel hence they bear a half distance metric compared to other phonemes.

	[ ]	*m*	*u*	*l*	*t*	*a*
[ ]	0	1	2	3	4	5
m	1	0	1	2	3	4
o	2	1	1	2	3	4
n	3	2	2	2	3	4
t	4	3	3	2	2	3
æ	5	4	4	4	3	***2****

Table 3.: Calculation of edit distance between two words based on insertion, delete or substitution (Levenshtein distance) where [ ] is the string index of the compared words.

The bigger the word the likely to have more edits operations and hence the bigger string distance. Qualitative a normalized Levenshtein distance algorithm as shown below:

Steps	*Normalized (by standard cost metric distance) LV. D*
1	Set n to be the length of s. Set m to be the length of t. If n = 0, return m and exit. If m = 0, return n and exit. Construct a matrix containing 0…m rows and 0…n columns.
3	Initialize the first row to 0…n. Initialize the first column to 0…m.
4	Examine each character of s (i from 1 to n). and then Examine each character of t (j from 1 to m)
5	If s[i] equals t[j], the cost is 0. If s[i] doesn't equal t[j], the cost is 1.
6	Set cell d[i,j] of the matrix equal to the minimum of: a. The cell immediately above plus 1: d[i-1,j] + 1. b. The cell immediately to the left plus 1: d[i,j-1] + 1. c. The cell diagonally above and to the left plus the cost: d[i-1, j-1] + cost.
7	After the iteration steps (3, 4, 5, 6) are complete, the distance is found in cell d[n, m].

Fig.5: Computation Algorithm of Normalized Levenshtein distance as a simple metric string approximation tool to calculate string distance between Cognate word list.

Based on the lexicon similarity and string distance between words per each language in

the Swadesh list, we identify a symmetric matrix of the four-language family interrelatedness. Our study used a statistical approach, to generate a phylogenetic tree of the language family.

Using normalized function of Levenshtein distance, each word is paired to calculate how it is interrelated to cognate words of other language. This result in a 200 × 16 matrix of all the Swadesh list words. We identify our algorithm to study the interrelatedness of the superfamily of the words by generating a phylogenetic tree, Language Family Generation Algorithm. Comparatively, we cross-examine our approach to other statistical approach and the famously used unweighted Pair-Group Method with Arithmetic mean (UPGMA) which has been used to generate phylogenic tree mainly in genome biology and genealogy fields.

Step 3:

We calculate the Adjusted Cognate Distance Score (Adj_Score) to identify a firm range of one-word interrelatedness to another based on their phonetic sound similarity not the average length of the words being compared. In that:

$$\text{Adj\_score} = \text{row\_score} / (0.5\ (\text{length}\ (w1) + \text{length}(w2))$$

For every comparison of a given word we calculate the overall adjusted similarity to the same words in different language in that we use the row score and precalculated Adj_score. Thus:

$$\text{Overall Adj\_simm}(L_1, L_2) = \text{SUM}\ (1/\text{adj\_score})$$

$$\textit{where}\ \text{adj\_score(w1, w2)} < 1.5$$

B. Unweighted Pair-Group Method with Arithmetic mean (UPGMA)

The UPGMA is the method of tree construction that employs a sequential clustering algorithm, in which local topological relationships are identified in order of similarity, and the phylogenetic tree is built in an agglomerative manner (Kita & Kenji, 1999).

UPGMA consist of more pitfall that its ultrametricity distances are defined by the satisfaction of the 'three-point condition". (Li & Xu, 2010). For any three lexicon taxa (A,B and C): *dist AC <= max (dist AB, dist BC)* or in words: the two greatest distances are equal, or UPGMA assumes that the evolutionary rate is the same for all branches, in that *dist AC* the new node as the mean of the two nodes that were joined to create it; which ignores a lot of words that are interrelated.

C. Our Approach

Our approach, Language Family Tree Generating Algorithm (LFTGA), is based on the qualitative properties that characterize the relative position of lexicons and to what is their normalized string distance. We identify how interrelated the cognate words in the Swadesh list are related. The smaller the string distance; the more the words are related. We apply this methodology as follow: Initially That we start by assigning all clusters (initial samples) to a star-like tree, which are represented in a symmetric matrix N × N string distance of Swadesh word of then do the following steps:

Steps:	*Preselection methodology: Building Language Superfamilies Algorithm*
0	Call Levenshtein distance function to generate string distance of words to be studied
1	Initialize a symmetric matrix (n × n) of string distance d[i,j].
2	Find that pair (cluster i and j) with the smallest adjusted score distance value in the distance matrix: d[i,j].
3	Create a new cluster d'(i,j), which has d(i,j) = d_i + d_j members. d' comprises of i and j: Cluster i is connected by a branch to the common ancestor node. The same applies for cluster j. Therefore, the distance d[i, j] is split onto the two branches. So, each of the two branches obtains a length of d[i, j]/2.
4	If i and j are the last cluster, Exist.

	Else find a new cluster.
5	(a) Combine $d_{i,j}$ – $d'_{i,j}$ as $D_{i,j}$ as a new cluster (b) Go back to step 0, recalculate string distance of D'_{ij} to the remainder clusters
7	Define the distance from u to each other cluster (k, with k <> D_i or D_j) to be the minimum distances d_{ki} and d_{kj}. For both complete and single linkage': dku= min(Dki, Dkj).
8	Go back to step 1 with one less cluster. Clusters i and j are eliminated, and cluster k is added to the tree. As, a N-1 × N-1 matrix. Repeat the algorithm n times until there are no more clusters, then exit.

Complexity: *The time and space complexity are O(n2), since there are n-1 iterations, with O(n) work in each one, Where N is the number of Languages.*

By using data analysis techniques, we improve the lexicostatistical analysis (as well as any other formal statistical or probabilistic methods) that always goes hand-in-hand with rigorous comparative research. Based on the Swadesh list of cognate words from various languages and cluster them in super-family we deploy our quantitative approach as well. Based on the above data to identify the phylogenetic trees of the

language family by the root mean square error, their string distance average mean and more importantly the inverse distance of their cognate relatedness.

Chapter 7

Experimental Results

We apply both UPGMA and language family tree algorithm to identify to robustness of our approach. The operation of word interrelatedness normalizes the distance metric at the same time; in that identical words return a Levenshtein distance of 0. In simulation, we use only words of string distance 1 to 3, thus they are more related than others. Hence, the more the derived words are interrelated they are considered to be from the same root (origin) and cognates.

Referring to Table 2. By using our algorithm and preselection rules, we are able to minimize the string distance between word that contain short and long vowel ***"a"*** and "***æ".*** We consider these to have a short distance metric. Instead of having the distance of 3, our methodology outputs 2.5 that rounds to 2 as the metric distance between two cognate words " *'multa*" and " *'montæ".* Hence, using the same methodology we calculated the metric distance between each word in the database and minimized the distance of each row score using the adjusted score metric. Which derived matrix of cognate words paired of string that are accurate cognate. The preselection rules reduce a margin error to a difference of ~= .5 robust as one can argue that it is of rounding string distance, misspelling of one or two cognate words which would not cause any misplace the whole super family.

In our comparative study of cognate words paired where string distance(d)<=3, we identify only cognate words with the smallest string that we use to apply the language family tree algorithm and the UPGMA algorithm as both methodologies are distance method and therefore need a distance matrix.

	BA	*AF*	*IE*
AF	V (10,8,2) 180*		
IE	V (2,6,2) 190*	V (0,2,1) 197*	
UR	V (2,3,5) 190*	V (1,2,4) 193*	V (5,9,4) 182*

Table 4: A minimized matrix of cognate words paired of string distance(d)<=3, where V- are the vector matrix to the number of most related cognate words of string distance 1,2, and 3 respectively. BA, AF, IE, UR are Bantu, Afroasiatic, Indo-European and Uralic Superfamilies respectively and X values are the optimistic estimate string distance.*

Using our algorithm, we were able to build a phylogenic tree for superfamilies. As a test case, we used only seven languages from four different families. We based on the optimistic estimate string distance between families to identify which family is interrelated to another. Therefore, the smaller the optimistic estimate string distance value, the more the languages are related as they will have more accurate cognates

among them. Fig.3 below shows the relationship of the language s we used for our research and to how they are related:

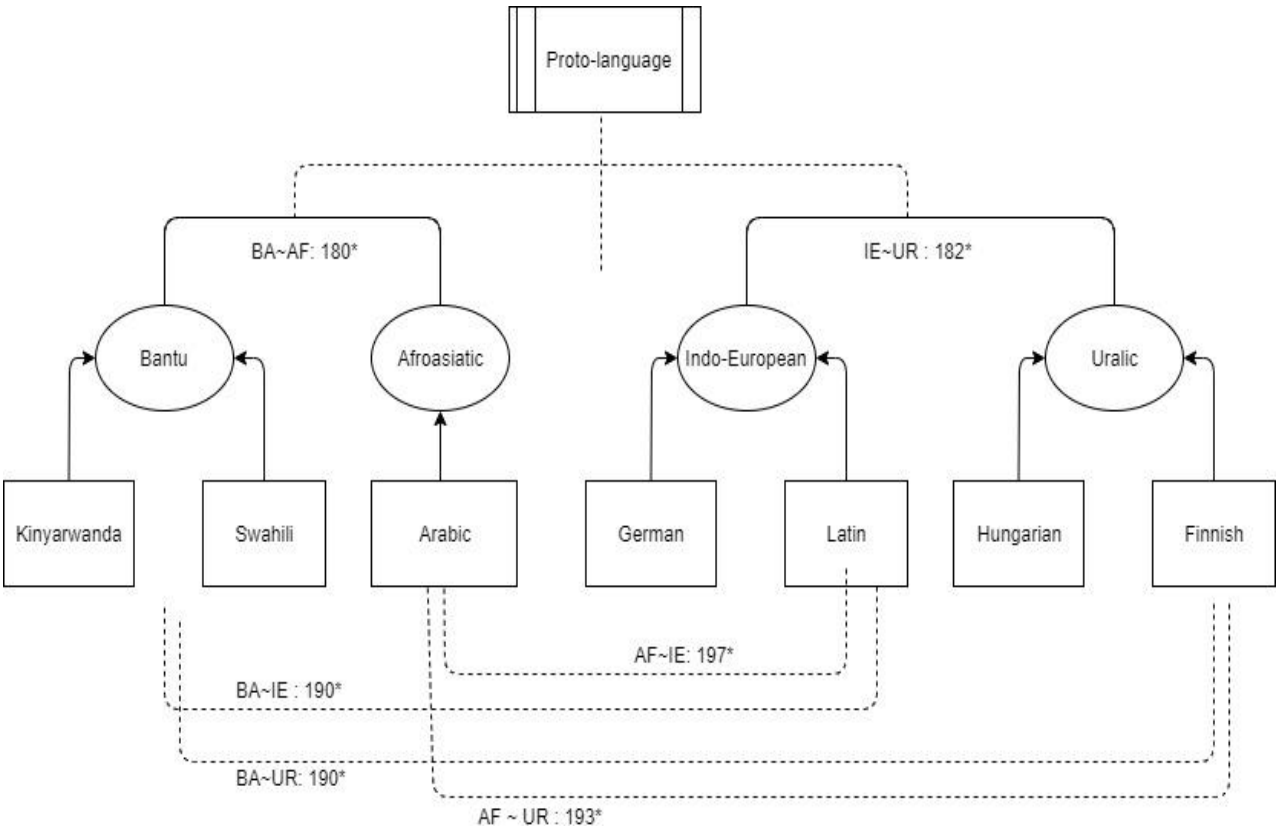

Fig.3: The implementation phylogenetic tree of Language Family Tree Generating Algorithm of Bantu, Afro-asiatic, Indo-European, and Uralic based on Swadesh list (is the optimistic estimate string distance).*

Chapter 8

Conclusion

In this thesis, we identify a well-founded preselection-based algorithm (LFTGA) to identify interrelatedness of languages based on the similarity of the cognate words shared. The test case of the algorithm bases on the model of using Swadesh list of various languages of four family (Afroasiatic, Bantu, Indo-European, and Uralic family). We applied a Levenshtein distance to identify the similarity of string words. Unlike previous studies, our method uses a sophisticated adjusted computer-based score calculation to preselect the cognate words based on their phonetic sound similarity not the length of the word compared.

Our analysis method shows the differences in item translation and cognate judgments that have a great impact on the topology of the trees calculated from lexicostatistical datasets. Based on the minimized optimistic string distance of the cognate words we identify an exemplary language interrelatedness as a test case. Datasets encoded in this way can then further used for phylogenetic calculations, and we hope that they will provide a more objective basis for stochastic calculations on linguistic datasets and may reveal interesting aspects and new insights into the complexity of language history to how languages are related to language family and pro-languages.

REFERENCES

1. Atkinson, Q. D. (2003), *"Language-tree divergence times support the Anatolian theory of Indo-European origin",* pages: 426, 435-439.
2. Batagelj, Vladimir, Pisanski, Tomaž, and Keržič, Damijana. (1992), "Automatic Clustering of Languages," *Computational Linguistics*, 18(3):339–352, 1992.
3. Brown, D., Kass, R., Uri, E., & Bown, E. (2014), *"Analysis of Neural Data".* New York: Springer, pages: 73, 710-713.
4. Daggumati, S. & Revesz, P. (2018), "Data Mining Ancient Script Image Data Using Convolutional Neural Networks," *Proc. 22nd International Database Engineering and Applications Symposium*, ACM Press, pp. 267-272, Villa San Giovanni, Italy.
5. Daggumati, S. & Revesz, P. (2019), Data mining ancient scripts to investigate their relationships and origins, *Proc. 23rd International Database Engineering and Applications Symposium*, ACM Press, pp. 209-218, Athens, Greece.
6. Dimmendaal, G. (2007), "The Wadi Howar diaspora: Linking linguistic diffusion to alaeoclimatological and archaeological findings. In Atlas of cultural and environmental change in arid Africa," *Africa Praehistorica* 21, ed. Olaf Bubenzer, Andreas Bolten, and Frank Darius, pp. 148-149. Cologne: Heinrich-Barth-Institut.

7. Gamallo, Pablo, Pichel, José Ramom, & Alegria, Iñaki. (2017), " From language identification to language distance," *Physica A: Statistical Mechanics and its Applications*, 484:152–162.

8. Gomaa, W.H., & Fahmy, A.A. (2013). *"A Survey of Text Similarity Approaches".* International Journal of Computer Applications, Volume 68– No.13

9. Habumuremyi, E. (2006). *"Iriza ry' ikinyarwanda",* English Kinyarwanda Dictionary. Vol 1.

10. Hinkka, A. (2018), *" Data-driven Language Typology",* University of Helsinki, Faculty of Science, Department of Computer Science. 3:4-7.

11. Hua, Guan-Jie, Che-Lun Hung, Chun-Yuan Lin, Fu-Che Wu, Yu-Wei Chan, and Chuan Yi Tang (2017), "MGUPGMA: A Fast UPGMA Algorithm with Multiple Graphics Processing Units Using NCCL." *Evolutionary Bioinformatics.*

12. Kita, Kenji. (1999), *"*Automatic Clustering of Languages Based on Probabilistic Models*," Journal of Quantitative Linguistics*, 6(2):167–171.

13. Li Y. and Xu L. (2010), "Unweighted Multiple Group Method with Arithmetic Mean," Fifth IEEE International Conference on Bio-Inspired Computing: Theories and Applications, Changsha, pp. 830-834.

14. Mutabazi, B. & Revesz, P. (2019), "A Quantitative Lexicostatistics Study of the Evolution of the Bantu Language Family," WSEAS Transactions on Computers, 18: 97-100.

15. Revesz, P. Z. (2016), "A computer-aided translation of the Cretan Hieroglyph script," *International Journal of Signal Processing*, 1:127-133, 2016.

16. Revesz, P. Z. (2017), "Establishing the West-Ugric language family with Minoan, Hattic and Hungarian by a decipherment of Linear A," *WSEAS Transactions on Information Science and Applications*, 14, 306-335.

17. Revesz, P. Z. (2018), "Spatio-temporal data mining of major European river and mountain names reveals their Near Eastern and African origins," *22nd European Conference on Advances in Databases and Information Systems*, Springer LNCS 11019, 20-32.

18. Revesz, P. Z. (2019), "Sumerian contains Dravidian and Uralic substrates associated with the Emegir and Emesal dialects," *WSEAS Transactions on Information Science and Applications,* 16, 8-30.

19. Ringe, D. (1995), "Nostratic' and the factor of chance," *Diachronica* 12:55–74.

20. Smith, R. (2018), "Constructing word similarities in Meroitic as an aid to decipherment", *Bouchet-Franklin Institute.*

21. Starostin, George (ed.) 2011-2016. *"The Global Lexicostatistical Database"*, Russian State University for the Humanities, & Santa Fe: Santa Fe Institute. Available online at http://lexstat.tk/databases/