A Quantitative Lexicostatistics Study of the Evolution of the Bantu Language Family

BILL MUTABAZI and PETER Z. REVESZ
Department of Computer Science and Engineering
University of Nebraska-Lincoln
348 Avery Hall, Lincoln, NE 68588
United States of America
mutabazi@huskers.unl.edu, revesz@cse.unl.edu

Abstract: This paper utilizes lexicostatistics and agglomerative, bottom-up hierarchical clustering methods to to identify the interrelatedness of thirty-two different Bantu languages and to computer generate an evolutionary tree for the Bantu language family.

Key-Words: - Bantu, Cognate, Hierarchical Clustering, Language Evolution, Lexicostatistics, Phylogenetic Tree

1 Introduction

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 [3]. Atkinson [1] found some evidence for the existence of a phonetically complex archaic language. In particular, Atkinson [1] found that the further humans travelled from Africa, the fewer number of different phonemes survived in various languages.

In this paper we study the Bantu language family, which is a major language family in Africa using *lexicostatistics*, using a method originally proposed by Morris Swadesh. He proposed to build relative genetic classifications of languages based on the percentages of cognate words in their basic lexicons. Alternative methods of study involve data mining and neural data [2] that we do not use in this paper.

Linguists identify *cognates* as words that have a common origin. Cognate words are usually inherited from a shared parent language. Revesz [9] identified recently a few cognate words that cut across language families in Africa and Eurasia. For example, Revesz [9] 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 (see Revesz [8]), the existence in it of a word, which is apparently cognate with a Bantu word, is surprising. Previously the proposed Nostratic superfamily tried to link several Eurasian language families but did not include any from Africa (see Ringe [12]). In addition, Revesz [10] proposed to add the Euphratic language, a Proto-Sumerian language, into Uralic family tree. The above results raised the possibility of finding additional cognates between Bantu and the Eurasian language families.

Computational linguistics often overlaps with the field of natural language processing because most of the tasks are common to both fields. While natural language processing focuses on the tokens/tags and uses them as predictors in machine learning models, computational linguistics digs further deeper into the relationships and links among them. Our approach is to mine data from various online dictionaries, as aligned sequences of cognate words of different languages families and apply sequentially the unweighted pair-group with arithmetic mean (UPGMA) clustering method [5, 6]. The UPGMA method identifies the inter-relatedness of languages to output a phylogenetic tree that reflects the evolution of these languages.

2 Problem Formulation

2.1 The database used

Cognate words are usually identified by at least these essential properties:

- They are always structural units.
- They are words that have a similar but not necessarily identical meaning.
- They always share a formal resemblance.

Swadesh used 200-words long lists of basic vocabularies in his lexicostatistical studies. The Swadesh list comprises names of body parts, names of some domestic and wild animals, simple verbs and nouns for everyday activities. Swadesh estimated the relatedness of two languages to be approximately related to the number of cognate words present in his word lists.

In this paper, we use an online raw database called the *Global Lexicostatistical Database (GLD)*, a hierarchical system of wordlists organized from bottom to top. GLD classifies annotated Swadesh 100-word list word data in all various families. We put together our own database based on annotated Swadesh list cognate words of Bantu languages with their respective online dictionaries from ancient languages. This database contains fields for various word features such as a word's grammatical form describing whether it is an adjective, adverb, noun, or pronoun, verb, etc.

2.2 Structural and syntactic similarities between Bantu Languages

Although the exact historical dispersion of Bantu speaking peoples is still a mystery, today's Bantu languages can be traced to a common origin. Structural and syntactic similarities between Bantu languages are evidenced in the following examples using the verb "to love" showing how most of the other Bantu languages relates or has mainly borrowed a lot of similarities from Swahili. Bantu languages remain one of the secrets of Africa's unwritten past - unless the secret is waiting to be revealed in the translations of the ancient Meroe script [4, 11] that is shown in Fig. 1.

Although there are still many scripts that have not yet been deciphered yet, there will perhaps come the day when linguists will decode and discover the ancient past. Phylogenetic algorithms can play a role in their decipherments, as shown in Revesz [7] by comparing a script whose phonetic values are unknown to a script with known phonetic values.

Fig 1. Early scripts signs used in cursive hieroglyphs and most of Bantu languages borrowed from the extinct meroitic language, they differ only in the shape of their signs. (Yahya, M., 2004).

3 Data Analysis

First, we calculate the Hamming distance δ between each pair of the 32 Bantu languages. The distance value between a pair of languages is calculated to be the percent of cognates in the Swadesh lists associated with those two languages, as shown in

	Swahili	Krongo-Kadugli	Nyimang	…	Kinyaranda	Luganda	Daju	Beni Choko	Molo
Swahili	0	397	409		138	142	256	401	364
Krongo-kaduguli	397	0	401		386	338	203	398	222
Nyimang	409	401	0		187	165	111	384	154
:									
.									
Kinyarwanda	138	386	187		0	118	322	187	338
Luganda	142	359	134		122	0	146	198	403
Daju	256	203	205		384	408	0	232	134
Beni Choko	401	385	308		390	356	268	0	155
Molo	411	123	416		277	288	253	373	0

Fig 2. The Hamming distance matrix for 32 Bantu languages.

Fig. 2. Second, we apply the UPGMA clustering algorithm [5, 6] to generate an evolutionary tree of the Bangtu language family as shown in Fig. 3.

That means, that we start by assigning all clusters (initial samples) to a star-like tree, then do the following steps:

1. Find that pair (cluster i and j) with the smallest distance value in the distance matrix: D[i,j].

2. Define a new cluster comprising cluster 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.

3. If i and j were the last 2 clusters, the tree is finished. If not the algorithm finds a new cluster called u.

4. Define the distance from u to each other cluster (k, with k <> i or j) to be an average of the distances dki and dkj. For 'Weighted PGMA (WPGM)': dku = dki+dkj/2. For 'Complete linkage': dku = max(dki, dkj). For 'Single linkage': dku = min(dki, dkj).

5. Go back to step 1 with one less cluster. Clusters i and j are eliminated, and cluster u is added to the tree.

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 Bantu languages. Based on the above corpus of data which gave a Hamming distance matrix.

5 Conclusion and Future Work

Languages are important to study because they are vital tools for the communication of thoughts and ideas, building friendships, economic relationships and preserving cultural ties in our societies from one generation to another. Our study focused on the mapping of the evolution of the Bantu language family, which is still not well understood. We used in our study lexicostatistics together with other techniques. Lexicostatistic dating of the separation of cognate languages on the basis of the percentage of common basic vocabulary was inspired by

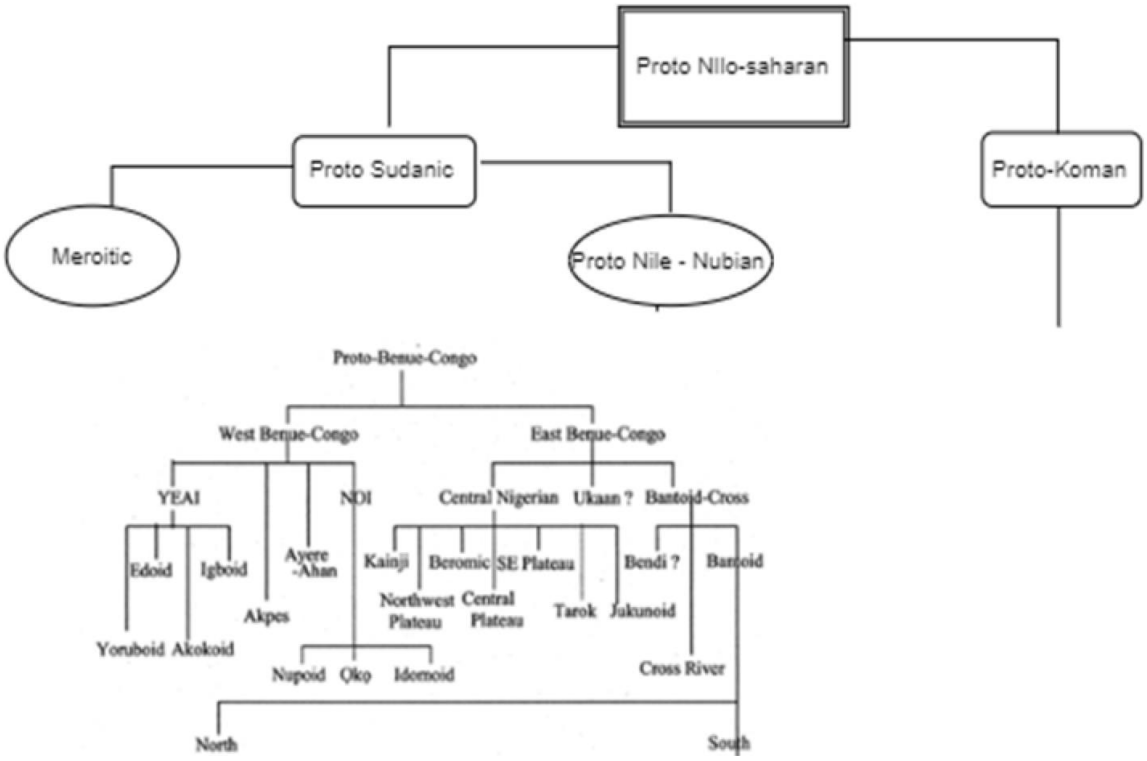

Fig 3. Phylogenetic tree classification of Bantu languages.

radiocarbon-dating of organic matter, and was based upon the alleged discovery of "the fact that fundamental vocabulary changes at a constant rate." As many authors have pointed out, this simple assumption needs to be further investigated because it may not hold in all cases of human language evolution. In the future, we can experiment with more sophisticated mathematical formulas for the estimation of the distance between two languages.

References:

[1] Atkinson, Q. D. (2003), "Language-tree divergence times support the Anatolian theory of Indo-European origin", pages: 426, 435-439.

[2] Brown, D., Kass, R., Uri, E., and Bown, E. (2014) Analysis of Neural Data. New York: Springer Science, Business Media: Book Review. , Page: 73, 710-713.

[3] Dimmendaal, G. (2007). The Wadi Howar diaspora: Linking linguistic diffusion to palaeoclimatological 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.

[4] Griffith, F L. (1911). The Meroitic inscriptions of Shablûl and Karanóg. University of Pennsylvania E. B. Coxe Jr. Expedition to Nubia VI. Philadelphia: University of Pennsylvania Museum.

[5] 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.

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

[7] Revesz, P. Z (2016), "Bioinformatics evolutionary tree algorithms reveal the history of the Cretan Script Family," International Journal of Applied Mathematics and Informatics, 10, 67-76.

[8] 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.

[9] 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.

[10] 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.

[11] Rilly, C. (2016), "Meroitic," In Julie Stauder-Porchet, Andréas Stauder and Willeke Wendrich (eds.), UCLA Encyclopedia of Egyptology, LA. Online Version: http://digital2.library.ucla.edu/viewItem.do?ark=21198/zz002k7ghs

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