Volume 6 ▪ Issue 1 ▪ March / April 2015

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

Publication Frequency: 6 Issues / Year

ISSN : 2180-1266

CSC PUBLISHERS

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

VOLUME 6, ISSUE 1, 2015

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 2180 - 1266

International Journal of Computational Linguistics (IJCL) is published both in traditional paper form and in Internet. This journal is published at the website http://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJCL Journal is a part of CSC Publishers

Computer Science Journals

http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCL)

Book: Volume 6, Issue 1, March / April 2015
Publishing Date: 30-04-2015
ISSN (Online): 2180-1266

IJCL Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2015

EDITORIAL PREFACE

The International Journal of Computational Linguistics (IJCL) is an effective medium for interchange of high quality theoretical and applied research in Computational Linguistics from theoretical research to application development. This is the *First* Issue of Volume *Six* of IJCL. The Journal is published bi-monthly, with papers being peer reviewed to high international standards. International Journal of Computational Linguistics (IJCL) publish papers that describe state of the art techniques, scientific research studies and results in computational linguistics in general but on theoretical linguistics, psycholinguistics, natural language processing, grammatical inference, machine learning and cognitive science computational models of linguistic theorizing: standard and enriched context free models, principles and parameters models, optimality theory and researchers working within the minimalist program, and other approaches.

IJCL give an opportunity to scientists, researchers, and vendors from different disciplines of Artificial Intelligence to share the ideas, identify problems, investigate relevant issues, share common interests, explore new approaches, and initiate possible collaborative research and system development. This journal is helpful for the researchers and R&D engineers, scientists all those persons who are involve in Computational Linguistics.

Highly professional scholars give their efforts, valuable time, expertise and motivation to IJCL as Editorial board members. All submissions are evaluated by the International Editorial Board. The International Editorial Board ensures that significant developments in image processing from around the world are reflected in the IJCL publications.

IJCL editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Scribd, CiteSeerX Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCL. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJCL provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Computational Linguistics (IJCL)

EDITORIAL BOARD

TABLE OF CONTENTS

Volume 6, Issue 1, March / April 2015

Pages

Classification of Oromo Dialects: A Computational Approach

Feda Negesse *feda.negesse@aau.edu.et*
Language Technology Group
Department of Linguistics
Addis Ababa University
Addis Ababa, PoBox: 176, Ethiopia

Abstract

Oromo is a lowland east Cushitic language which has tens of millions of native speakers in Ethiopia and in neighboring countries such as Kenya and Somalia. In the past, some attempts have been made to subjectively divide the language into different dialects or genetic units based on some phonological and lexical features. However, this study is intended to automatically compute lexical distances among varieties of the language spoken in Ethiopia and to objectively classify them into dialect areas. One hundred sixty basic words were used to calculate the normalized lexical distances with the Levenshtein Algorithm and an agglomerative clustering method was employed to classify the linguistic varieties into dialect areas. It is observed that the objective method has yielded a good result in dividing the linguistic varieties into six clusters and this classification is similar to some of the previous subjective classifications. It is also noted that the linguistic varieties have formed hierarchical clusters based on their geographical proximities, showing the dialectological fact that a geographical proximity predicts a linguistic similarity. A new classification of dialects of the language has been proposed but further research is needed to validate it with more lexical data and other clustering techniques.

Keywords: Oromo Language, Oromo Dialect, Levenshtein Algorithm, Lexical Distance, Computational Methods.

1. INTRODUCTION

The standard procedure in the traditional study of dialects has been classifying languages into dialects and demarcating their boundaries on a map depending on a subjective judgment of the dialectologist. The use of isoglosses has been another commonly used procedure to divide languages areas into dialect areas. An isogloss is a line on a map which divides areas whose dialects differ in some specific linguistic features (1). Nevertheless, the application of isoglosses has three major limitations (1, 3). First, it is difficult to have isoglosses which coincide because they may run parallel, or even cross each other, resulting in contradicting binary divisions. The dialectologist has to choose the best isoglosses which form bundles and this makes the procedure subjective. Second, the use of isoglosses may give a wrong impression that dialects are categorically different as it is impossible to indicate degrees of differences. Finally, speakers of dialects might be displaced by migration, war and natural disasters so that speakers of neighboring dialects may not live in adjacent areas. Computational techniques have been developed to measure the linguistic distances between dialects in order to solve some of the limitations (4).

One of the computational techniques is the Levenshtein distance, which has been improved over the past at different times to widen its applications and to increase its efficiency (5, 6). The Levenshtein distance between two strings is defined as the minimum number of edits which transforms one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character (7). However, the cost of mapping a string into other strings with different length is not the same and thus researchers very often employ normalization to solve this problem but the normalization procedures are different (8). For instance, Marzal & Vidal (5) tried to normalize the Levenshtien distance in terms of path rather than edit transformation while

other researchers (9) normalized the distance by dividing the distance of two strings by their mean length to avoid length bias. Recently, Higuera & Mico (10) proposed a contextual normalization in which every edit operation is divided by the length of the string where the edit operation takes place. However, the cubic complexity of the algorithm developed to compute a contextual distance is so great that it is not as commonly used as the generalized normalization.

The Levenshtien distance is a widely used metric in dialectometry for dialect comparison (4,11, 12,13) and in spellchecker(14).Kessler (3) used the Levenshtien distance to compare the distances between different Gaelic dialects using linguistic data collected from Ireland, Scotland and the Isle of Man; and he reported satisfactory results in clustering those Gaelic dialects into their natural classes. Heeringa, Kleiweg, Gooskens, and Nerbonne (15) analyzed the effects of n-grams on the distance measure using both examples from Norwegian and German, and they indicated that the *n*-grams only slightly improve the result. Furthermore, they confirmed that the approach which is based on phones instead of feature vectors could lead to better results. Nerbonne (13) measured Levenshtein distances on the basis of the entire set of Lowman's Southern states pronunciations in the Linguistic Atlas of the Middle and South Atlantic States. He also measured distances on the same data set using only vowels. When the two sets of linguistic distances were correlated, a high correlation (r = 0.94) was found, suggesting that vowel distance is a good predictor of a linguistic distance.

The studies reviewed above reveal that the advantage of the Levenshtien distance over the isogloss method lies in its ability to compare linguistic varieties in an objective way on the basis of an aggregate of linguistic features rather than on the basis of just a single one. The added advantage of Levenshtien distance is its significant correlation with other dialect measures. Gooskens & Heeringa (11) stated that linguistic dialect distances measured with Levenshtien correlate significantly with perceptual distances for 15 Norwegian varieties (r= 0.67, p <0.001).The study of 15 Norwegian dialects also showed that significantly stronger correlation was found between pronunciation (P < 0.001) and the perceptual distances (r=0.68) than between pronunciation and the lexical (r=0.30) distances (11).Nevertheless, Moberg, Gooskens and Nerbonne (16) indicated that linguistic distance is symmetrical but intelligibility score is often asymmetrical and thus Levenshtein distance cannot be always a strong predictor of intelligibility among dialects.

2. CLASSIFICATION OF OROMO DIALECTS

Oromo is one of the major languages in Ethiopia and spoken in other African countries such as Kenya and Somalia (17). According to the 2007 census of the country, Oromo is spoken as the first language by 33.3 % of the Ethiopian population (18). Some past and recent studies (19, 20, 21, 22, 23, 15, 24, 25) are available on the classification of dialects of Oromo and the studies reported inconsistent results on the number of Oromo dialects. Bender, Mulugeta and Stinson (19) divided the dialects into Macha, Tulama, Wollo, Rayya, Arsi, Guji, and Borana on the basis of geographical boundaries. Their classification did not take into account Oromo dialects spoken outside Ethiopia and the Oromo dialects in Kenya are Garba, Ajuran, Orma, Munyo, Garre and Waata (21). On the other hand, Lloret (23) broadly divided the dialects into just two, Western and Eastern groups and this classification is definitely crude.

However, some writers (22, 17) divided the language into clusters, which contain different dialects. For instance, Wako (22) classified the dialects of the language into five clusters such as Southern (Arsi, Guji and Borana), Central (Karayu, Selale), Mecha (Jimma, Wollega and Ilubabor), Eastern (Harar and Bale) and Nortern (Raya and Wollo). Similarly, Kebede (17) classified Oromo into four clusters such as North Western (Tulma, Mecha), Eastern (Harar, Arsi-Bale, Wallo, Rayya), Central (Arsi-Zeway, Guji, Borana, Munyo, Orma) and Southern (Waata). However, nine years later, Kebede (24) categorized Oromo into Wollo, Raya, Tulema, Mecha, Arsi, Hararge, Guji, Borana of southern Ethiopia and northern Kenyan, Orma, Gabra, Ajuran, Sakuye, Garreh, Munyo, and Waata varieties. The Sakuye variety was ignored in the previous classification by Kebede and his colleagues (21). In addition, he did not mention why his

classifications of Oromo dialects is inconsistent but as stated earlier, it is understandable that a subjective classification can yield inconsistent results.

FIGURE 1: An Ethiopian map showing the regional states in the federal state and Oromo dialects are spoken in Oromia (also spelled Oromiya), which is the biggest regional state in the country. The other dialects of Oromo, Wollo and Rayya, are spoken in Amhara and Tigray regional states respectively (*26*).

Two years later, Kebede (25) conducted extensive research on Oromo dialects as part of his doctoral study. Based on morphophonemic and phonetic-lexical data, he constructed a genetic tree of Oromo dialects, whereby the dialects were divided into ten genetic groups such as western, eastern, central, south-east-north, Waata, northeast east, north, Wollo and Raya. He reported that four of the ten genetic units did not exist because they gave way to the present dialects. The western genetic unit includes the variety spoken in Mecha and Tulema while the central genetic unit encompasses the variety used in Arsi (Rift-Valley Area), Gujii, Borena and Orma of Kenya. The Oromo spoken in the south eastern of Kenya is included in the Waata genetic unit. On the other hand, the East subgroup contains a variety which is used in an area stretching from Asri (High Land) to Jijjiga. The Wollo and the Raya subgroups are related because they have evolved from the same genetic family but they are at risk of endangerment being encircled by two major languages, Amharic and Tigrigna.

In general, review of the studies indicates that, to date, no computational methods have been used to compare linguistic distance among the dialects and no attempt has been made thus far to classify the dialects objectively. Consequently, the objectives of this study are to compute lexical distance among Oromo varieties using the Levenshtien algorithm and to automatically classify the dialects based on the values obtained from the computation. Finally, it is also intended to validate the previous impressionistic classifications attempted by different researchers.

3. METHODS

3.1. Materials

Eleven speakers (one from each variety) of Oromo varieties were asked to translate 160 strings (basic vocabulary) written in English into their own dialects. Then three other speakers of the same dialects were requested to edit the translation of the vocabulary for spelling and

representativeness of the dialects. When there was disagreement among them regarding how a word should be translated into Oromo, the translation accepted by two of them was taken as a correct one. The problem is though Oromo speakers can be influenced by a *quasi-standard Oromo,* which appears to be composed of different dialects. The quasi-standard Oromo seems to be emerging as one of the dialects of the language and is used in media, education and other public domains. It is practically impossible to find speakers who have not been exposed to this dialect but the participants of the study were told to translate the list of words based on their native dialects and the edition of the translation was also done accordingly.

3.2. Calculating Lexical Distance

As discussed earlier, pairs of longer strings have on average have a larger Levenshtein distance than that of pairs of shorter strings. Consequently, the normalized edit distance was used to calculate the lexical distance between two linguistic varieties. The two words had to be taken into account when the distance between them was computed (6). 160 pairs of words in two linguistic varieties were aligned to compute the path costs, the following mathematical procedure:

> An editing path P between two words, S and R, of lengths n and m, respectively ($n \leq m$), is a sequence of ordered pairs of integers (ik, jk), where $0 \leq k \leq m$, that satisfies the following.
>
> $0 \leq k \leq , 0 \leq jk \leq R$;
> $(i0 , j0) = (0,0) , (im , jm) = (|S| , |R|)$
> $0 \leq ik - ik-1 \leq 1, 0 \leq jk - jk-1 \leq 1 , \forall k \geq 1$
> $ik - ik-1 + jk - jk-1 \geq 1$
>
> The weights can be associated to paths as follows:
> $\omega(Ps, R) \sum_{k-1}^{m} \gamma\left(S_{i_{K-1}} k - 1 + 1 \ldots i_k \rightarrow R_{j_{k-1}} + 1 \ldots j_k\right)$
> $ED\ (S, R) = min\ \{\omega(P) |\ P \text{ is an edit transformation of } S \text{ into } R\}$
>
> Let $\omega\hat{}(p) P = \omega(P) / L\ P$, where L is the length of P , the normalized edit distance NDE is defined as : $NDE(S, R) = min\{\omega\hat{}(P)\}$

In order to counter a word length bias, the function is expressed in terms of path costs and not in terms of the edit operation costs. To automatically calculate lexical distance between pairs of eleven varieties of Oromo language, the above mathematical function was coded in the Levenshtien algorithm written by Schauerte & Fink (27) and implemented in the MATLAB version 2011a.

3.3. Classification of Dialects

The data obtained from Levenshtien distance were summarized with MATLAB version 2011a and the dialects were automatically classified into groups with a hierarchical clustering method. Clustering is a well-known procedure to seek groups of close varieties, and has been used in dialectometry to classify languages into dialects (28, 29). This procedure attempts to identify relatively homogeneous groups of cases based on selected characteristics, using an algorithm that starts with each case in a separate cluster and combines clusters until only one is left (30).It is an iterative procedure that selects the shortest distance in a matrix and fuses the two data points that give rise to it. Therefore, based on the squared Euclidean lexical distance, the hierarchical clustering method could produce hierarchically structured clusters of the dialects in the form of dendrogram.

4. RESULTS

4.1 Lexica Distances among the Linguistic Varieties

One of the objectives of this study is to compute lexical distance among the Oromo varieties spoken in Ethiopia in order to determine their linguistic distance. It is always true that the distance between a linguistic variety and itself is zero, which shows that the smaller the distance, the closer the dialects.TABLE1 indicates that the distance among the varieties ranges from zero to

2.2037, with the greatest distance obtained between Borana and Wollo dialects as the speakers of these dialects live in regions located far from each other. Speakers of the Borana dialect inhabit the southern part of Ethiopia, which shares a geographical boundary with the northern part of Kenya while the Wollo Oromo live in northern part of the country. When compared to the lexical distance between Borana and Guji, Wollo and Rayya have a longer distance and this is not surprising as the varieties are heavily influenced by Amharic and Tigrigna.

Varieties	Borana	Guji	Arsi	Bale	Harar	Wollo	Rayya	Showa	Wollega	Jimma	Ilubabor
Borana	0.0000										
Guji	0.8723	0.0000									
Arsi	1.8644	1.6269	0.0000								
Bale	1.9649	1.6923	0.1428	0.0000							
Harar	1.7419	1.8833	1.2000	1.0122	0.0000						
Wollo	2.2037	2.1273	1.1600	1.2192	1.0769	0.0000					
Rayya	1.8361	2.1091	1.5625	1.5873	1.2817	0.8736	0.0000				
Showa	1.5441	1.5588	1.7833	1.7541	1.4412	1.2105	1.4412	0.0000			
Wollega	2.1754	2.3962	1.9000	1.8833	2.1111	1.6567	1.9167	1.1379	0.0000		
Jimma	1.7846	1.8438	1.9655	2.0351	1.6719	1.5652	1.6716	0.9468	0.9468	0.0000	
Ilubabor	2.3148	2.2143	1.9655	1.9828	2.2500	1.7188	2.0172	1.6716	0.2769	0.4138	0.0000

TABLE 1: A matrix of lexical distances among Oromo varieties automatically computed with the Levenshtein Algorithm.

TABLE 1 also reveals that the closest distance was found between Oromo speakers of Arsi and Bale, and it confirms the common observation that speakers living in adjacent places use closely related linguistic features. The Oromo varieties spoken in Wollega, Jimma and Ilubabor also have short Levenshtien distance ranging from 0.2769 to 0.9468, with the shortest being between Wollega and Ilubabor. In the previous classifications, the speakers of these three areas have been consistently included in one of the Oromo dialects called Macha (21,22). In addition, it can be seen that the Macha variety has a longer Levenshtein distance with Borana, Guji and Harar varieties than with the Macha variety. Similarly, Heeringa, Kleiweg, Gooskens & Nerbonne (15) stated that it is an established fact in dialectology that geographic proximity among varieties can generally predict their linguistic similarity. In other words, dialects which are geographically closer exhibit more linguistic similarities than dialects which are far from each other. For instance, the study conducted on Norwegian dialects indicated that municipalities (geographical areas) which are geographically closer were also found to be linguistically closer (29), which suggests that linguistic varieties carry important information about their geographical locations. In other words, linguistic features (e.g., lexical items) speakers use may tell where speakers are from, and knowing the geographical areas where speakers live may help linguists to safely guess the linguistic features used in the areas.

4.2. Automatic Classification of the Linguistic Varieties

This study is also intended to objectively classify the Oromo varieties into dialects based on the Levenshtein distances among them. The dendrogram reveals that the linguistic varieties are classified into hierarchies of small and big clusters. Consistent with the data in TABLE1, the language is divided into three big clusters which are difficult to label but if the Borana-Guji cluster

had not been separated, the language would have been divided into two big clusters which could be named Barentuma and Borana. The small clusters in the dendrogram include Wollega-Ilubabor-Jimma, Arsi-Bale, Harar-Wollo-Rayya, Borana-Guji, and Showa. The language is also divided into smaller clusters when the Harar variety stands as an independent dialect separating itself from the Wollo-Rayya group. It is worth noting that the Showa variety forms a separate group on its own because of its contact with Addis Ababa and Showa dialects of Amharic. Although empirical data are unavailable, it is a common observation that this variety has frequently borrowed lexical items from the adjacent dialects of Amharic. Understandably, the Showa variety has distinctive linguistic features of its own, which contribute to its independence as a separate dialect of Oromo.

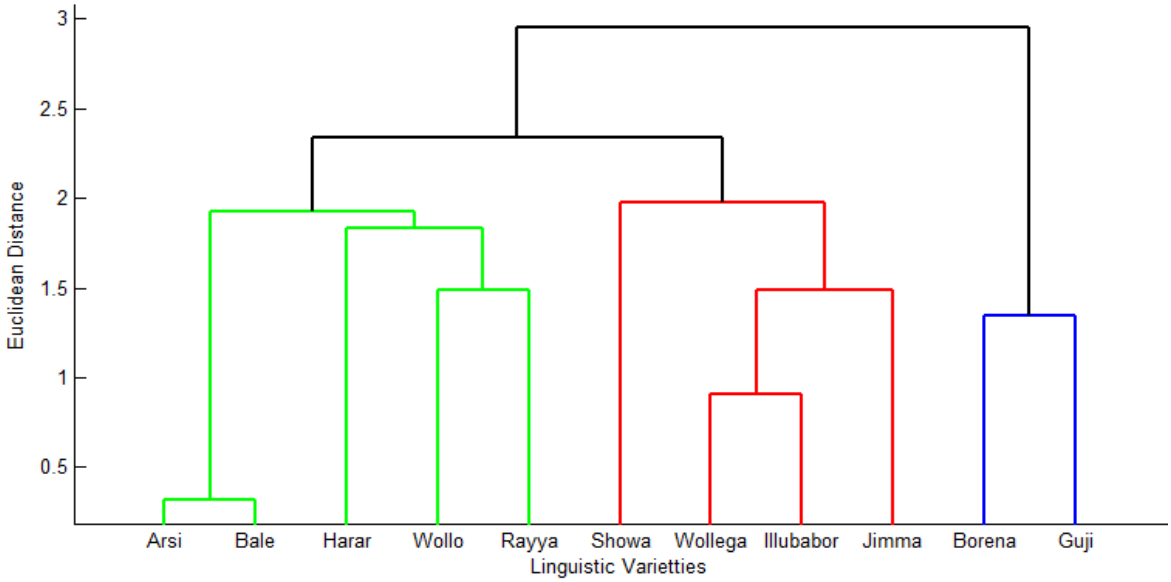

FIGURE 2: Dendrogram derived from the 11 × 11 matrix of average Levenshtein distances showing the clustering of Oromo dialects. The scale distance is given as a Euclidean distance and the Dendrogram was constructed based on Average Linkage of Within Group. The varieties will be more and more dissimilar as the distance between them increases.

The classification in FIGURE 2 is similar to those in the previous literature (19, 20, 21, 25). For instance, Wako (22) grouped dialects of the language into southern cluster (Arsi, Guji and Borana), Central cluster (Karayu, Selale), Mecha cluster (Jimma,Wollega and Ilubabor), eastern cluster (Harar and Bale) and northern cluster (Raya and Wollo). The label *Tulama* is very often used to refer to the Oromo variety spoken in the central part of Oromia, which include the Salale and Karayu varieties (24). Similarly, Kebede (25) divided Oromo spoken in Ethiopia into west, central, east and north genetic groups based their geographical locations, which take into account varieties used in Kenya. Worth noting is the Arsi and Bale varieties which have been usually considered as one dialect named Arsi or Arsi-Bale variety (22) and interestingly, these two varieties have the shortest distance as indicated in FIGURE 2. However, the classification with the Dendrogram is inconsistent with the classification done subjectively by different researchers (22, 23, 24). For example, Lloret (23) divided the language into western and eastern dialect areas but the dialects of the language formed two clusters as this researcher did but none of them could be considered as western and eastern groups. Generally, the computational method performed very well in classifying the language into different dialect areas and this classification could agree with the subjective judgment or intuition of linguists who attempted to classify the language into dialects or clusters.

Therefore, it is important to propose a new classification of Oromo dialects for three major reason; the previous classifications were done based on subjective comparisons of mainly lexical differences. The second prime reason is that the past classifications produced inconsistent results causing confusions to readers. Finally, in the old classifications, the labels that are used to identify the dialects are not inappropriate as the names of Oromo subgroups such as Macha and Tulama have been used to refer to the western and the central varieties but names of geographical areas have been used to identify the other varieties. In the new classification, the Oromo language have been divided into five dialects such as west, central, northern, southern, southeast and eastern dialects as indicated in the following tree-diagram. Thought it is well known that the computational methods is better than the traditional (subjective) method, further research is certainly needed to validate the results of the current study. The language may be classified into various dialect areas differently if all Oromo dialects inside and outside Ethiopia are included. The Borena variety may form cluster with Oromo varieties spoken in Northern Kenya as speakers have good business interaction across the porous border.

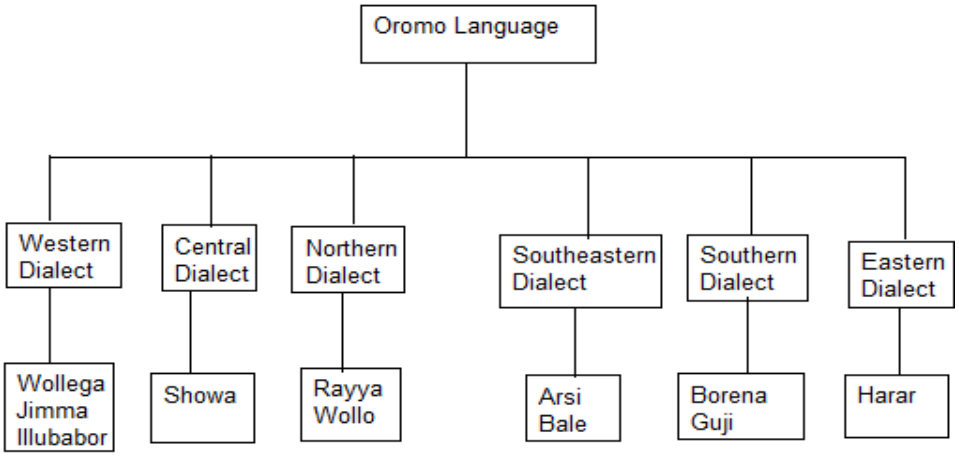

FIGURE 3: A tree-diagram demonstrating the newly proposed classification of Oromo dialects spoken in Ethiopia.

This classification can be indicative of mutual intelligibility and perceptual distance among the dialects. The previous studies on dialects of European languages revealed that Levenshtein distance is strongly correlated with mutual intelligibility scores and perceptual distance. For instance, Gooskens & Heeringa (11) indicated that Levenshtein distance is correlated significantly with perceptual distance for 15 Norwegian varieties (r= 0.67, p <0.001). In addition, Beijring, Gooskens & Heeringa (31) reported that the normalized Levenshtein distance of Danish dialects was strongly correlated to intelligibility scores obtained from 351 native speakers of the language (r = 0.8, P< 0.01). Therefore, one can hypothesize that the western dialect of Oromo is more intelligible to the speakers of the central dialect than to the speakers of the other dialects. Clearly, a rigorous study is needed to assess the mutual intelligibility of the dialects in order to compare the intelligibility scores with the Levenshtein distance of the dialects.

5. CONCLUSIONS

As stated earlier, this study sets out to determine the Levenshtien distance among the Oromo dialects and to ultimately classify the dialects objectively on the basis of the distance values. Normalized Levenshtien distances were computed for 11- by-11 matrix of eleven Oromo dialects spoken in Ethiopia. It was observed that the linguistic varieties are closer to one another based on their physical proximities and their closeness has an important implication for their linguistic homogeneity or mutual intelligibility. It was found that the Arsi and Bale varieties has the shortest linguistic distance while the Borana and Guji varieties have the longest distance measured with Levenshtien algorithms. The varieties spoken in Ethiopia were automatically classified into western, northern, central, southern, eastern and southeastern dialects. However, it is possible

that the eastern variety will be included in the southeastern dialect, forming a big linguistic variety but the speakers of this variety seem to use distinctive and perceptible phonetic features. Further research that will have more data is important to validate the results of the current research.

This study has some limitations though it has attempted to automatically classify the Oromo varieties into dialects with their appropriate labels. The regular clustering was used to classify the dialects into their classes but it is broadly accepted that it lacks stability. This is due to the fact that clustering looks for the minimum distance between two points in a matrix, and sometimes several pairs of elements may show similar distances. As a consequence, small differences in the input data matrix can lead to considerably different clusters (9). In future research, different clustering methods (e.g. K-means clustering, multidimensional scaling) can be used to supplement the limitation of the clustering technique employed in the current research. In addition, the study was limited in scope because it did not consider the Oromo dialects spoken in Kenya and Somalia due to financial and security constraints. A comprehensive study which will include data from all dialects of the language spoken in different parts of Africa is needed to present a complete classification of the dialects. Despite the limitations, the current study is interesting as it has employed such an objective procedure that it can be replicated and validated by a further investigation.

6. REFERENCES

[1] P. Matthews.The Concise Oxford Dictionary of Linguistics. Oxford: Oxford University Press, 1997.

[2] J. Chambers and P. Trudgill. Dialectology. Cambridge: Cambridge University Press, 1980.

[3] B. Kessler. "Computational dialectology in irish gaelic." in Proc. of the European Associaton for Computational Linguistics, 1995, pp. 60–67.

[4] W. Heeringa. "Measuring Dialect Pronunciation Differences using Levenshtein Distance." Ph.D.thesis, University of Groningen, 2004.

[5] R. Wagner and M. Fisher. "The string-to-string correction problem." Journal of the ACM, vol. 21, pp. 168–178, 1974.

[6] A. Marzal and E. Vida. "Computation of Normalized Edit Distances and Applications." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, pp. 926–932,1993.

[7] L. Yujian and L. Bo. "A normalized Levenshtein distance metric." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, pp 1091–1095, 2007.

[8] V. I. Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals." Soviet Physics Doklady , vol.10, pp. 707–10,1966.

[9] J. Nerbonne, W.Heeringa and P. Kleiweg. "Edit distance and dialect proximity." in Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, ed, D. Sankoff and Kruskal J. Stanford:CSLI Press, 1999, pp. v–xv.

[10] C. Higuera and L. Micó (2015,Jan.) " A Contextual Normalised Edit Distance." Researchgate.[On-line]. 23(2). Available: www.researchgate.net/Higuera/contextual.[Jan.2,2015].

[11] C. Gooskens and W. Heeringa. "Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data." Language Variation and Change, vo.16, pp. 189–207, 2004.

[12] J. Nerbonne. "Computational Contributions to the Humanities." in Conference of the Association for Literary and Linguistic Computing and The Association for Computers and the Humanities, Gothenburg, Sweden, 2004.

[13] J. Nerbonne. "Identifying linguistic structure in aggregate comparison." Literary and Linguistic Computing, vol. 21, pp.463–75, 2006.

[14] L. Salifou and H. Naroua. "Design of A Spell Corrector For Hausa Language." International Journal of computational Linguistics, vol.5, pp.14-26, 2014.

[15] W. Heeringa, P.Kleiweg, C. Gooskens and J. Nerbonne." Evaluation of string." in proc. the Workshop on Linguistic Distances, 2006.

[16] J. Morberg, C.Goosken and J.Nerbonne. "Conditional entropy as a measure of linguistic remoteness between related languages" . in Proc. Computational Linguistics, 2007.

[17] H. Kebede. "Raayaa Oromo Phonology: Aspects of Palatalization" in Ethiopia in Broader Perpespectives, 1997, vol. pp.469-91.

[18] Central Statistical Agency. Population and Housing Census of Ethiopia. Addis Ababa: Central Statistical Agency, 2007.

[19] M. L. Bender, E. Mulugeta and D. L. Stinson. Two Cushitic languages. in Language in Ethiopia, M. L. Bender, J. D. Bowen, R. L. Cooper and C. A.Ferguson, ED,. London: Oxford University Press, 1976, pp. 130-54.

[20] G. Gragg. Oromo of Wellega. in Language in Ethiopia, M. L. Bender, J. D. Bowen, R. L. Cooper and C. A.Ferguson, ED, London: Oxford University Press, 1976, pp. 166-95.

[21] B. Heine. "The Waata Dialect of Oromo: Grammatical Sketch and Vocabulary, Language and Dialect Atlas of Kenya". Journal of the International African Institute, vol. 55, pp. 228-232, 1980.

[22] T. Wako. "The phonology of Mecha Oromo". Unpublished MA Thesis. Institute of Language Studies: Addis Abeba University, Ethiopia, 1981.

[23] M. Lloret (1994). A Comparative Study of Consonant Assimilation in Some Oromo Dialects. in the 3rd International Symposium on Cushitic and Omotic Languages, Berlin.

[24] H. Kebede. "Causative Verb and Palatalization in Oromo".The Journal of Oromo Studies vol 14, pp.95-109, 2007.

[25] H. Kebede. "Towards the Genetic Classification of the Afaan Oromoo Dialects." Published PhD Thesis, Department of Linguistics and Scandinavian Studies: The University of Oslo, Norway, 2009.

[26] Discover Ethiopia, http://hayo.co/discover-ethiopia, Feb. 2015.

[27] B. Schauerte, G. A. Fink, "Focusing Computational Visual Attention in Multi-Modal Human-Robot Interaction," in Proc. ICMI, 2010.

[28] D. Shaw. "Statistical analysis of dialect boundaries." Computers and the Humanities, pp.173-177,1974

[29] S. Hyvonen, A Leino, and M. Salmenkivi. "Multivariate Analysis of Finnish Dialect Data:An Overview of Lexical Variation." Literary and Linguistic Computing, vol. 22, 2007.

[30] J.Verma, and V. Richhariya. "A Review: Salient Feature Extraction Using K-Mediods Clustering Technique." Journal of Computer Science and Information Technology, pp.23 – 25, 2012.

[31] K. Beijering, C. Gooskens and W. Heeringa ." Predicting intelligibility and perceived linguistic distance by means of the Levenshtein algorithm." Linguistics in the Netherlands, pp.13-24, 2008.

INSTRUCTIONS TO CONTRIBUTORS

Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective. Today, computational language acquisition stands as one of the most fundamental, beguiling, and surprisingly open questions for computer science. With the aims to provide a scientific forum where computer scientists, experts in artificial intelligence, mathematicians, logicians, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists can present research studies, International Journal of Computational Linguistics (IJCL) publish papers that describe state of the art techniques, scientific research studies and results in computational linguistics in general but on theoretical linguistics, psycholinguistics, natural language processing, grammatical inference, machine learning and cognitive science computational models of linguistic theorizing: standard and enriched context free models, principles and parameters models, optimality theory and researchers working within the minimalist program, and other approaches. IJCL is a peer review journal and a bi-monthly journal.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCL.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Started with Volume 6, 2015, IJCL aims to appear with more focused issues related to computational linguistics studies. Besides normal publications, IJCL intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJCL List of Topics:
The realm of International Journal of Computational Linguistics (IJCL) extends, but not limited, to the following:

- Computational Linguistics
- Computational Models
- Computational Theories
- Corpus Linguistics
- Formal Linguistics-Theoretic and Grammar Induction
- Information Retrieval and Extraction
- Language Generation
- Language Learning
- Linguistics Modeling Techniques
- Linguistics Theories
- Machine Translation
- Models of Language Change and its Effect on Lingui
- Models that Address the Acquisition of Word-order
- Models that Combine Linguistics Parsing
- Models that Employ Statistical/probabilistic Gramm
- Models that Employ Techniques from machine learning
- Natural Language Processing
- Quantitative Linguistics
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Web Information

CALL FOR PAPERS

Volume: 6 - **Issue:** 2

i. Paper Submission: April 30, 2015 **ii. Author Notification:** May 31, 2015

iii. Issue Publication: June 2015

CONTACT INFORMATION

Computer Science Journals Sdn BhD

B-5-8 Plaza Mont Kiara, Mont Kiara

50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6204 5627

Fax: 006 03 6204 5628

Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2015
COMPUTER SCIENCE JOURNALS SDN BHD
B-5-8 PLAZA MONT KIARA
MONT KIARA
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6204 5627
FAX: 006 03 6204 5628
EMAIL: cscpress@cscjournals.org