Gadaa Journal/Barruulee Gadaa

Vol. 2 No. 2 June 2019
https:journals.ju.edu.et

e-ISSN: 2616-3985
p-ISSN: 2616-3977

A Bilingual Journal of the Institute of Oromoo Studies
(IOS) Jimma University

Full Length Research Paper

Revisiting the Swadesh's Wordlist: How long should it be?

Feda Negesse (PhD) *
Addis Ababa University
Department of Linguistics
Email: fedan2010@gmail.com

Submission Date: 13/07/2019
Acceptance Date: 20/10/2019

Abstract

The Swadesh's wordlist has been used for more than half a century to collect data for studies in comparative and historical linguistics. The current study compares the classification results of the Swadesh's100 wordlist with those of its subsets to determine if reducing the size of the wordlist impacts its effectiveness. In the comparison, the 100, 50 and 40 wordlists were used to compute lexical distances of 29 Cushitic and Semitic languages spoken in Ethiopia and neighboring countries. Gabmap, a based application, was employed to compute the lexical distances and to divide the languages into related clusters. The comparison shows that the subsets are not as effective as the 100 wordlist in clustering languages into smaller related subgroups, but they are equally effective in dividing languages into bigger groups such as subfamilies. It is observed that the subsets may lead to an erroneous classification whereby unrelated languages by chance form a cluster which is not attested by a comparative study. The chance to get a wrong result will be higher when the subsets are used to classify languages which are not closely related. Though a further study is still needed to settle the issues around the size of the Swadesh's wordlist, this study indicates that the 50 and 40 wordlists cannot be recommended as reliable substitutes for the 100 wordlist under all circumstances. The choice seems to be determined by the objective of a researcher and the degree of affiliation among the languages to be classified.

***Keywords**: Swadesh, Wordlist, Classification, Cushitic, Semitic*

---

* Corresponding author.

Axareeraa

Tarreen jechootaa Iswaadiish qorannoolee Seerluga seenaafi waldorgomsiisaatiif ragaalee sassaabaadhaaf jaarraa walakkaadhaa ol hojiirraa oolaa ture. Kaayyoon qorannoo kanaas bu'aalee qoqqooddii tarree jechootaa Iswaadiish 100 tuutaalee xixiqqaasaa waliin waldorgomsiisuudhaan hanga tarree jechootaa xiqqeessuun bu'aa qorannoorratti dhiibbaa inni qabu agarsiisuudha. Waldorgomsiisuu kana keessatti garaagarummaa fageenyaa jechootaa (lexical distances) afaanota Kuushiifi Seemii 29 Itoophiyaafi biyyoota ollaa keessatti dubbataman gidduutti mul'atu agarsiisuuf tarreen jechoota 100, 50fi 40 hojiirra ooleera. Dalagaa kanaaf tooftaan herreeguu Gapmap jedhamu fageenya jechootaa (lexical distance) herreeguudhaafi afaanota kanneen maatiilee walfakkaatan jalatti qooduuf tajaajileera. Bu'aan dorgomsiisuu kun kan inni agarsiisu tuutaaleen xixiqqaa afaanota gara garee maatii xixiqqaasaaniitti qoqqooduudhaaf hanga tarree jechootaa Iswaadiish bu'a qabeessa akka hintaanedha; garuu afaanota kanneen gara maatiilee gurguddootti qoqqooduurratti tarree jechootaa Iswaadiish waliin qixa bu'a qabeessadha. Haaluma kanaan tarreen jechootaa tuuta xixiqqaa afaanota wal hinfakkaannefi hinfiroomne maatiilee sirrii hintaane jalatti qooduudhaan dogoggora qooddii uumuu akka danda'u bu'aan qorannoo kanaa agarsiiesera. Hangamtaa tarree jechootaa Iswaadiish daangessuudhaaf qorannoon bal'aa barbaachisullee tarreen jechootaa 50fi 40 haala kamiinillee tarree jechootaa Iswaadiish bakka bu'ee qorannoodhaaf akka hintajaajille qorannoon kun mirkaneesseera. Ta'us filannoon tarree jechootaa kan inni murteeffamu kaayyoo qorataafi sadarkaa qooddii afaanotaarratti hundaa'a.

Jechoota Ijoo: *Iswaadiish, tarree jechootaa, qoqqooddii, Kuushii, Seemii*

---

1. Introduction

One way of keeping research quality is employing an effective data collection tool because it undoubtedly determines the quality of data to be gathered (Vorndran & Botte 2008; Sharma 2012). As a data collection tool for historical and comparative linguistics, the Swadesh's wordlist was developed by Swadesh (1952) and has been used for more than half a century. The review of related works indicates the wordlist has been used to calculate lexical or phonetic similarities in order to classify languages or varieties based on their similarities (Chumbow et al. 2007; Blench et al. 2008; Kitchen et al. 2009; Starostin 2010). The limitations of the wordlist are that lexical similarity can happen due to borrowing, chance and universal tendency of languages to have similar lexical items such as onomatopoeic words (Minett & Wang 2003; Muller et al. 2009). However, it is assumed that borrowing does not equally affect words in a language and this is universal applying to all languages. The borrowing of lexical items among non-related languages is so rare that it accounts for about 2.5 to 5 percent of the 40 wordlist compiled by Holman et al. (2008). Moreover, it is assumed that the replacement of the basic words through time is less stable and same for all languages at any time. A study conducted on thirteen languages estimated that the average vocabulary retention is about 80.5 percent every thousand years (Holman et

al. 2008), and this percentage is consistent with Swadesh's (1971) claim that the replacement rate per million years is 15 percent. The retention rate of vocabulary depends on frequency of use, semantic field and culture while the frequency of use accounts for 50 percent of the retention variance (Atkinson 2010; Vejdemo 2010).

The Swadesh's 200 words can split into two groups: the 100 words initially created by Swadesh, and the remaining 100 words. The first 100 words are called high rank while the remaining 100 words low rank (Chen 1996). Words in the high rank are postulated to be more stable and loan-resistant than those in the low rank (Chen 1996). Moreover, it is known that more retention of a proto-language will be kept in the high rank while borrowed items will be brought into the low rank more quickly and easily. Based on this point, Chen (1996) proposed a method to judge genetic relationships between languages. He stated that languages with genetic affinity have a greater number of related words in the high rank than in the low rank and these words are presumably resistant to a lexical change. The comparison of the 200 wordlist and the 100 wordlist revealed that the presence of loanwords in the full list does not significantly affect the classification results (Syrjanen et al. 2013). The impact of loanwords may have been offset by the size of the wordlist and relative weight of size of the wordlist and the number of loanwords.

In spite of the attempts to improve the wordlist, there has been a perceived arbitrariness in determining the number of words in the Swadesh's list because the items on the list have increased 100 to 200 (Sarah 1962; Starostin 2000; Kitchen et al. 2009). The free ride in the length of the wordlist seems to have inspired other researchers to create a very long list for a comparative study (Snider & Roberts 2006). Some researchers have started wondering if the subsets of the longest wordlist could be used to get satisfactory results. For instance, Chumbow et al. (2007) used 50 words from the 100 wordlist and reported a good result on the classification of Cameroon and Equatorial Guinea languages, but the selection of the words in the subset was not systematic. However, other researchers tried to calculate stability index of the 100 words in the Swadesh's list and took only those which could yield an optimal classification. Among them are Holman et al. (2008) who reported that the first 40 words which have a higher stability index are enough to classify languages into subgroups. On the other hand, others still argue that even a shorter list containing the first (35 or 15) most stable words can produce a good classification result as the 100 wordlist (Holman et al. 2008).

Unfortunately, investigators (Holman et al. 2008, 2010) employed different procedures in calculating stability index, which resulted in different ranks of stability index for the same items; case in point is louse which ranks first in Holman's et al. (2008) list but 17th in Starostin's (2010) list. The words in the 40 stable lexical items are not the same though not altogether different. Starostin (2010) argues that the first 50 stable items (according to his rank of stability index) are sufficient to obtain a good classification result for remotely related languages but he contends that the 100 wordlist or 200 wordlist may be needed to classify dialects and closely related languages. Despite the difference in their computational techniques, both investigators might have considered many world languages when they

calculated stability rate because it could be affected by geographic settings of languages or their families (Greenhill et al. 2008).

The 40 wordlist is frequently used in published works because the subset has been experimented on different world languages and produced reliable results (Kitchen et al. 2009; Wichmann 2012). The experiments indicated that increasing the list size does not affect a classification result, producing no significant gain or loss. It is also known that a short list can result in an erroneous classification of languages with an attested genetic relationship (Wang & Wang 2004; Syrjanen et al. 2013). The Swadesh's wordlist has two associated problems: the length of the list is far from being settled as there are various proposals and the ranking of the lexical items is based on different stability indexes. The current study is concerned with the first problem as it sets out to test Starostin's (2010) and Holman's et al. (2008) proposals by applying them on 29 Cushitic and Semitic languages spoken in Ethiopia and neighboring countries. As far as the online survey of published works is concerned, the proposals have not been tested on different languages. Therefore, the main aim of the current study is to compare the classification results of the 100 Swadesh wordlist with those of its subsets to determine if reducing the length of the wordlist impacts its effectiveness.

2. Method

2.1. Languages and the wordlists

The languages in the study belong to two main branches of the Afroasiatic family: Cushitic and Semitic. Amharic, Ge'ez, Gurage, Tigrigna and Silte are members of the Ethio-Semitic group according to linguistic studies (Blench et al. 2008; Voigt 2009). Amharic, Chaha, Geto, Gafat, Inor, Kistane, Mesqan, Mesmes, and Silte are grouped under the South Semitic branch while Tigrigna, Tigre and Ge'ez under the North Semitic branch (Bender 1976). Some of the languages have disappeared, and others are used by small and threatened communities. For instance, three of the Semitic languages such as Ge'ez, Gafat and Mesmes are dead while others continue to actively serve as media of a daily communication for their respective speakers. Language death is disputable, but it is generally accepted that a language is declared totally dead when no "speakers are left of a particular language variety in a population that had used it" (Mufwene 2004, pp. 204).

On the other hand, the Cushitic languages in the current study are Afar, Arbore, Awngi, Gedeo, Hadiya, Oromo, Sidama, Somali and Tsamay (Mous 2012). The sub-grouping of the Cushitic languages is not conclusive as the position of Awngi is debatable and Tsamay is not considered in the internal classification of the family (Wedekind & Wedekind 2002; Mous 2012). Ongota is also a very controversial language, generating a lot of proposals about its genetic affiliation (Sava & Tosco 2000). This classification problem could be attributed to a lack of stability as the language is shifting to Tsamay and other neighboring languages (Campbell et al. 2014), adding to the list of endangered languages of the world. The Cushitic languages exhibit a greater time-depth than the other Afroasiatic languages in spite of their good typological similarities (Mous 2012). It is believed that the close relationship among the Semitic languages and the distant relationship among the Cushitic

languages will give a good opportunity to try out the Swadesh's 100 wordlist and its subsets to assess their effectiveness in classifying different languages.

The Swadesh's 100 wordlists of 29 languages which are spoken in Ethiopia and neighboring such as Eritrea, Kenya, Somalia, Sudan and Egypt were used in the study. The languages for which wordlists were readily available online and those for which native speakers were willingly available for translation were included in this study. The wordlists for five of the languages were collected online from published works while the wordlists for the remaining languages were collected by the investigator from native speakers of the languages. Arbore, Awngi, Tsamay, Ongota, Mussiye and Gedeo have wordlists in the SIL linguistic surveys conducted by Wedekind & Wedekind (2002). The wordlist used in the SIL survey is long consisting of 320 items, including the 100 Swadesh's core vocabulary; thus, only those words which are in the original wordlist were considered for this study.

2.2. Classification and statistical analysis

The effectiveness of 40 wordlist of Holman et al. (2008) and the 50 wordlist of Starostin (2010) was tested on the 29 languages described above. The wordlists were transcribed phonemically and submitted to Gabmap (a web-based application developed at Information Sciences, Groningen and hosted by at Meertins Institute, Amsterdam) in order to classify them based on lexical distances. The software was selected for ease of use, assuming no high computational skills (Nerbonne et al. 2011). The lexical items were tokenized so that diacritics would not be considered as separate characters and diagraphs (e.g., /dʒ, tʃ/) were not used in the phonemic transcription because the software considers them as separate strings. Using the weighted Levenshtein distance, the software automatically computed the lexical distances of the wordlists of the languages. The lexical similarity matrices of the 100 wordlist and its subsets were used to automatically classify the languages based on weighted average hierarchal clustering. In order to validate the results, an attempt was made to compare them with the previous classifications done by comparative techniques.

3. Results

The main objective of this study is to determine if the first two stable subsets (consisting of 50 or 40 items) of the Swadesh's 100 wordlist are sufficient to classify languages into groups already identified by a comparative method. Based on the previous works of Holman et al. (2008) and Starostin (2010), it is expected that classification results of the two subsets will be as good as that of the 100 wordlist.

3.1. Swadesh's 100 wordlist

The languages are divided into different hierarchies which are consistent with the classifications in the published works on Ethiopian languages (Figure 1). At the highest level, the dendrogram branches off into two bigger groups, which are readily recognized as Semitic and Cushitic families by a linguist familiar with Ethiopian languages. The Semitic unit split into two groups which are often known as the North and the South Semitic; the

former includes Ge'ez, Tigre and Tigrigna while the latter consists of Amharic, the Guraghe languages, Gafat, Argoba, Harari and Zay. It might strike readers as odd that Mesmes has moved out of the Guraghe cluster and joined Gafat because the language is usually considered a member of the cluster in many previous works (Lewis 2005). Geto and Chaha have the shortest lexical distance which means they are closely related sharing many lexical similarities. Diachronically, they may be considered sister languages which recently descended from a common ancestor though unrelated languages can exhibit large lexical similarities due to an excessive borrowing (Müller et al. 2009).

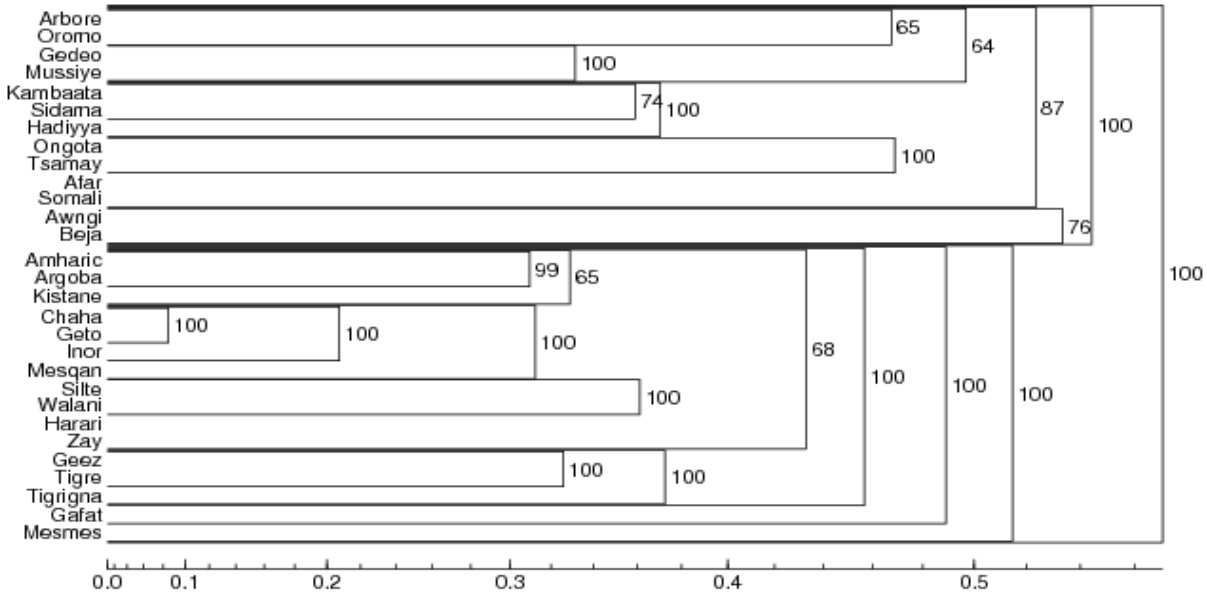

Figure 1: Classification of Cushitic and Semitic languages in the study

The numbers show the relative certainty of clusters of the 29 languages in the study and the clustering was made with Gaussian noise. The probabilistic clustering is based on a lexical similarity computed from the Swadesh's100 wordlist.

The Cushitic group has the largest of speakers inhabiting different regions, mostly the eastern and southeastern parts of Ethiopia and the dendrogram has divided the group into six subgroups, which consist of different number of languages. As expected, Awngi and Beja form a separate branch because the languages are different to be in one group. In the previous works, Awngi is reported to be part of the Central Cushitic which includes other varieties such as Qimant, Xamtanga and Bilen (Wedekind & Wedekind 2002). Beja is an isolate as it is the only language which has a separate branch, known as the North Cushitic (Eberhard, Simons & Fennig 2019).

Kambaata, Hadiyya and Sidama form a separate group while Tsamay and Ongota; Oromo, Arbore, Gedeo and Mussiye; and Afar and Somali make different clusters, which are comparatively homogenous. These languages are usually consistently reported to form the biggest Cushitic unit known as the East Cushitic which is further divided into Highland East Cushitic and Lowland East Cushitic. In the previous works (Tosco 2003), Highland East Cushitic consists of Kambaata, Hadiyya, Sidama, Gedeo and Burji, but Gedeo has left this group and joined Lowland East Cushitic, which includes Oromoid, Afar-Saho and Omo-

Tana groups. In the current classification, Somali, which is usually grouped together with Boni in the Eastern Omo-Tana, should have formed a group with Arbore, which belongs to the western linguistic unit in Omo-Tana (Mous 2012). The association of Afar with Somali should not be surprising because Somali is geographically the closest language to Afar, but the two languages are less closely related; Afar might have been classified with Saho if Saho had been included in the study (See Eperhard, Simons & Fennig 2019).

3.2. Starostin's Swadesh wordlist

The 50 wordlist is as effective as the 100 wordlist in classifying 29 Afroasiatic languages into families and subfamilies. Cushitic and Semitic families are clearly separated in the dendrogram, and they are further divided into small groups attested by past studies (Gragg & Hoberman 2012; Mous 2012). The internal classification of the Cushitic family also remained intact with only Oromo changing its membership, having joined the Highland East Cushitic, which is inconsistent with its classification reported in the previous studies (Tosco 2003).

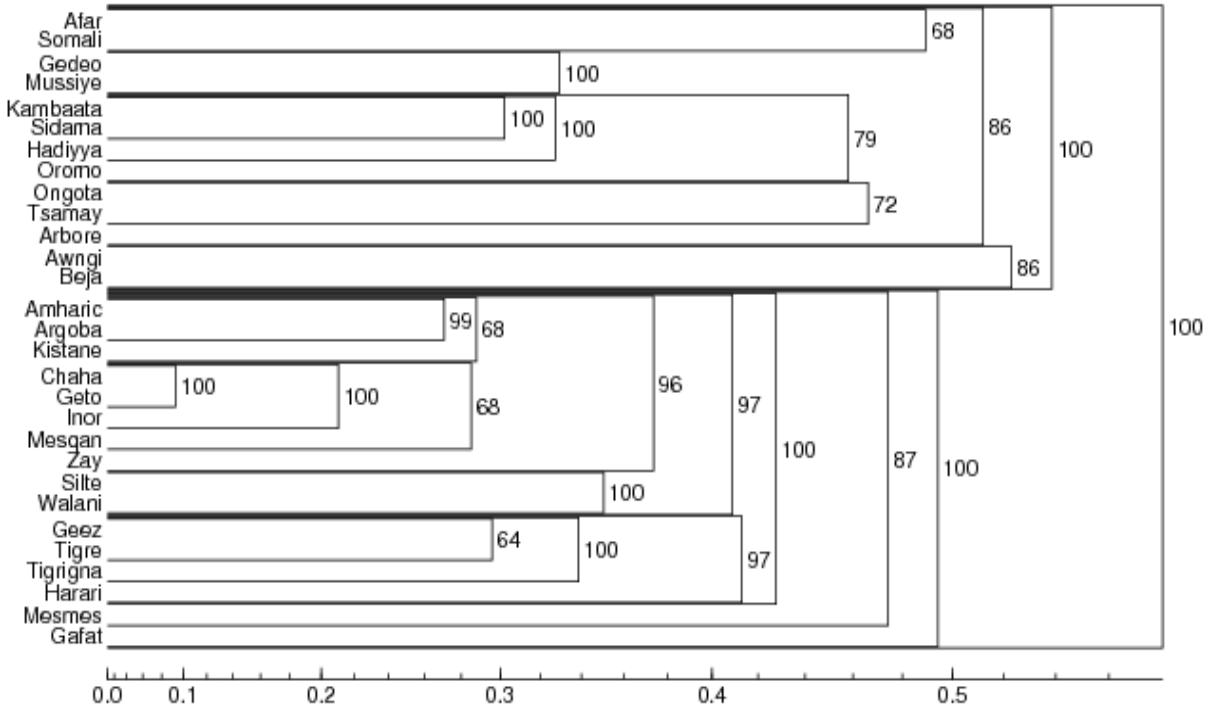

Figure 2: Classification of Cushitic and Semitic languages in the study

The numbers show the relative certainty of clusters of the 29 languages in the study and the clustering was made with Gaussian noise. The probabilistic clustering is based on a lexical similarity computed from the first 50 stable subset of the Swadesh's100 wordlist.

In addition, the 50 wordlist does well in the internal classification of the Semitic languages, but more instability is noted in this family, particularly in the Guraghe languages. For example, Kistane, Zay, Harari and Masqan changed clusters, making the hierarchical clustering unstable and unreliable (Figure 2). However, many groups such as the North Semitic, Amharic-Argoba, Chaha-Geto-Inor and Silte-Walani remained stable, and Gafat and Mesmes also remained in their clusters. It seems that the wordlist is more effective in

classifying languages which are either very closely or remotely related. Similarly, Starostin (2010) stated that this wordlist can be a better choice when one opts for a satisfactory classification instead of using the 200 or 100 wordlist which is relatively time-consuming especially if a large number of languages are involved. However, it is good to know that the 50 wordlist may not be as reliable as the 200 wordlist in classifying languages into clusters, which are attested by comparative studies.

3.3. Holman's et al Swadesh wordlist

Holman et al. (2008) and Starostin (2010) have tried to reduce the Swadesh's wordlist to 40 and 50 items respectively by taking only the first more stable words in the 100 wordlist proposed by Swadesh (1952). Obviously, the two wordlists are not altogether different because they contain words sampled from the same list but different indices of stability are used to select the words from the reservoir. Therefore, it is interesting to see if the two wordlists produce at least similar classification results as the 100 wordlist even though Holman et al. (2008) and Starostin (2010) have already claimed successful results, by having tried out their wordlists on several world languages.

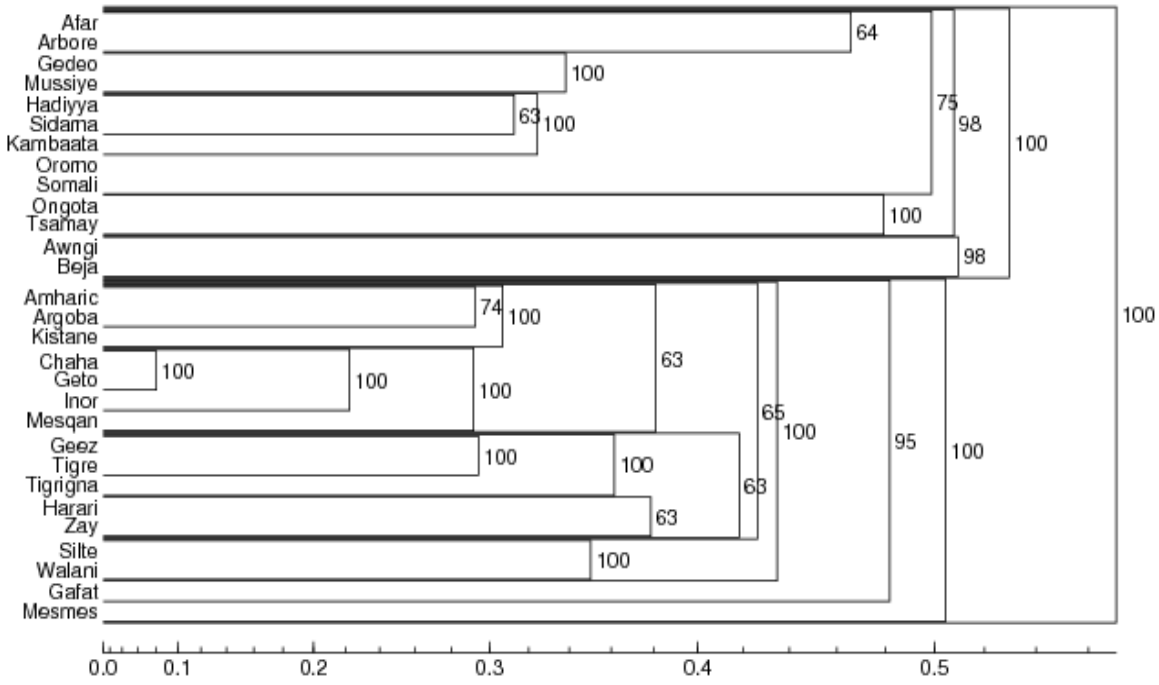

Figure 3: Classification of Cushitic and Semitic languages in the study

The numbers show the relative certainty of clusters of the 29 languages in the study and the clustering was made with Gaussian noise. The probabilistic clustering is based on a lexical similarity computed from the first 40 stable subset of the Swadesh's100 wordlist.

The classification generally remains intact, but very few lower nodes of the clusters become unstable. In the dendrogram, we can clearly see two big branches (Semitic and Cushitic) and the internal classification of both families has not changed essentially (Figure 3). Obviously, when compared to the performance of the 100 wordlist, the 40 wordlist has some limitations

in classifying languages which can be affiliated to different languages. For instance, Arbore is grouped with Afar and the value for certainty of the cluster is very low (64 percent), which casts doubt on the validity of the cluster representing languages which are closely related. In addition, the internal classification of the Highland East Cushitic is problematic as Sidama has become unstable by being grouped with Hadiyya or Kambaata. The dendrogram shows that the probability of Sidama forming a cluster with Hadiyya is lower (63 percent), but it is higher with Kambaata (100 percent). Interestingly, the two languages have not disintegrated when the size of wordlist is reduced to 40 items.

Similarly, the Semitic languages have not disintegrated with all languages remain in their clusters albeit the reduction of the size of the wordlist to 40 items. The languages are sorted out into the North and the South Semitic subfamilies, but the certainty for some languages to form a true cluster appears to have decreased. For instance, the certainty of Argoba and Amharic to branch off from the same node is reduced to 74 percent and the same is true for Zay and Harari. This pattern is also observed in Cushitic languages whereby the percentage of certainty for Sidama to form a true cluster with Hadiyya has decreased to 63 percent.

4. Discussion

Based on the results of previous studies, it was expected that the two subsets of the original Swadesh's wordlist compiled by Holman et al. (2008) and Starostin (2007) would have similar results in classifying 29 Cushitic and Semitic languages spoken in Ethiopia and other countries such as Kenya, Somalia and Eritrea. The classification results of the two subsets are generally similar, particularly in dividing the languages into families and subfamilies. When compared with the 50 wordlist, the 40 wordlist is more popular and thus more commonly used in language classifications (Kitchen et al. 2009; Wichmann 2012) but their classification results are not that different. The choice between the two wordlists may be guided by the availability of more works which have used the 40 wordlist. If one is to be guided by the principle that the more items in the list, the more accurate and reliable classification results will be expected, the 50 wordlist will be definitely a better option.

Generally, the 100 wordlist is more effective in separating the 29 languages into homogenous groups, which are consistent with classifications reported in the previous works on Cushitic and Semitic languages (Ethnologue, 2010) . The study conducted on Uralic languages showed that the 100 wordlist could yield a classification result which is similar to the 200 wordlist (Syrjanen et al. 2013). This is not surprising given the empirical evidence that the 100 wordlist is superior to the 200 wordlist in terms of containing words which are resistant to borrowing (Syrjanen et al. 2013). It is also known that genetically related languages have more cognates in the 100 wordlist which may contribute to its effectiveness. Therefore, it is evident that limitation in quantity of the wordlist can be compensated by its quality but it is yet not known where the balance between quantity and quality is optimally maintained.

5. Conclusion

The study assessed the effectiveness of the 50 and 40 wordlists in classifying languages into homogenous clusters and compares their classification results with that of the 100 wordlist. The intention is to see if the reduction of the size of the wordlist affects its classification efficacy. Accordingly, it showed that the subsets are not as effective as the 100 wordlist in clustering languages into smaller subgroups, but they are equally effective in dividing languages into bigger groups such as subfamilies. It is noted that the subsets may lead to erroneous classification results whereby unrelated languages by chance form a cluster not attested by a comparative study. It is important to know that the 100 wordlist can also result in a wrong classification, but the chance seems to be higher when the size of the wordlist is reduced. The chance is particularly greater when the subsets are used to classify languages which are not closely related. Though a further study is still needed to settle the issues around the size of the Swadesh's wordlist, this study indicates that Holman's et al. (2008) and Starostin's (2010) wordlists cannot substitute the 100 wordlist under all circumstances. The choice seems to be determined by the objective of a researcher and the closeness or affinity of the languages to be classified.

References

Atkinson, Q.D. (2010). The prospects for Tracing Deep Language Ancestry. *Journal of Anthropological Sciences*, 88, pp. 231–233.

Blench, R. et al. (2008). Links between Cushitic, Omotic, Chadic and the Position of Kujarge. *5th International Conference of Cushitic and Omotic languages*, pp.-pp.

Chumbow, S.B., Martin, D. & Bot, L. (2007). Classification of the Languages of Cameroon and Equatorial Guinea on the Basis of Lexicostatistics and Mutual Intelligibility. *African Study Monographs*, 28(4), pp.181–204.

Greenhill, S.J., Blust, R. & Gray, R.D. (2008). The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics. *Evolutionary Bioinformatics*, 2008(4), pp.271–283.

Holman, E.W. et al. (2008). Explorations in Automated Language Classification. *Folia Linguistica*, 42, pp.331–354.

Kitchen, A. et al. (2009). Bayesian Phylogenetic Analysis of Semitic Languages Identifies an Early Bronze Age origin of Semitic in the Near East. *Proceedings of Biological Sciences/The Royal Society*, 276 (1668), pp. 2703–10.

Lewis, M. (2005). Towards a Categorization of Endangerment of the World's Languages. Available at: http://ftp.sil.org/silewp/2006/

Minett, J.W. & Wang, W.Y. (2003). On Detecting Borrowing: Distance-based and Character-based Approaches. *Diachronica*, 20 (2), pp. 289–330.

Mous, M. (2012). Cushitic. In Z.Frajzyngier & E. Shay. *The Afroasiatic Languages*, pp. 342- 422. Cambridge: Cambridge University Press.

Gragg, G. & Hoberman, R. (2012). Semitic.In Z.Frajzyngier & E. Shay. *The Afroasiatic Languages*, pp. 342- 422. Cambridge: Cambridge University Press.

Mufwene, S. (2004). Language Birth and Death. *Annual Review of Anthropology*, 33, pp. 201-222.

Müller, A. et al. (2009). ASJP World Language Tree of Lexical Similarity: Version 2, pp.105–120.

Nerbonne, J., Collen, R., Gosskens, C., Kleiweg, P. & Leinonnen, T. (2011). Gabmap-A Web-based Application for Dialectology. *Dialectolgic*a, Special Issue II.

Sharma, O.P. (2012). Quality Indicators of Scientific Research. *Indian Journal of Microbiology*, 52(2), pp.305–306.

Starostin, G. (2010). Preliminary Lexicostatistics as a Basis for Language Classification: A New Approach. *Journal of Language Relationship*, 3, pp.79–116.

Swadesh, M. (1952). Lexico-statistical Dating of Prehistoric Ethnic Contacts: With Special Reference to North American- Indians and Eskimos. *Proceedings of the American Philosophical Society*,452–463, 1952.

Syrjanen, K. et al. (2013). Shedding more Light on Language Classification Using Basic Vocabularies and Phylogenetic Methods: A Case Study of Uralic. *Diachronica*, 30(3), pp.323–352.

Tosco, M. (2000). A Sketch of Ongota, a Dying Language of Southwest Ethiopia. *Studies in African Linguistics*, 29(2), pp. 59–134.

Tosco, M. (2003). Cushitic and Omotic Overview. *Selected Comparative-historical Afroasian Linguistic Studies in memory of Igor M. Diakonoff*, pp.87–92.

Vejdemo, S. (2010). The Effect of Semantic Properties on Rates of Cross-linguistic Lexical Change. pp.85–104.

Voigt, R. (2009). North vs. South Ethiopian Semitic. *Proceedings of the 16th International Conference on Ethiopian Studies*, pp.1375–1387.

Vorndran, A. & Botte, A. (2008). An Analysis and Evaluation of Existing Methods and Indicators for Quality Assessment of Scientific Publications.

Wang, F. & Wang, W.S.Y. (2004). Basic Words and Language Evolution. *Language and Linguistics*, 5(3), pp.643–662.

Wedekind, C. & Wedekind, K. (2002). Sociolinguistic Survey of the Awngi Language of Ethiopia. *SIL Electronic Survey Reports*, pp. 2002-044.

Wichmann, S. (2012). History, Contact and Classification of Papuan Languages: Part One. *Linguistic Society of Papua New Guinea*, pp.59–87.