ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF ENGINEERING

DEPARTMENT OF COMPUTING

LEMMATIZATION FOR AFAN OROMO TEXT

BY

MOHAMMED TIYA BERISO

A THESIS SUBMITTED TO THE SCHOOL OF ENGINEERING DEPARTMENT OF COMPUTING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS OF SCIENCE IN INFORMATION SYSTEM.

ADAMA, ETHIOPIA

FEBRUARY, 2017

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF ENGINEERING

DEPARTMENT OF COMPUTING

LEMMATIZATION FOR 'AFAN' OROMO TEXT

BY

MOHAMMED TIYA BERISO

A Thesis Submitted to the School of Engineering Department of Computing in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Information Systems.

Adama, Ethiopia

February, 2017

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF ENGINEERING

DEPARTMENT OF COMPUTING

LEMMATIZATION FOR 'AFAN' OROMO TEXT

By

Mohammed Tiya

Signature of the Board of Examiners for Approval

Name	Signiture	Date
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

DECLARATION

This thesis is my original work and has not been submitted for a degree in any other University and source was duly acknowledged.

_____

Mohammed Tiya

February, 2017

The thesis has been submitted for examination with my approval as University advisor.

_____

Dr. Solomon Teferra

February, 2017

Acknowledgment

First, I would like to thank the Almighty Allah for helping me to finalize my work peacefully.

My deepest gratitude goes to my advisor Dr. Solomon Teferra for his unreserved support in critically commenting the thesis work and advising me for better academic performance. I would also thank Ato Debela Tesfaye and Dr. Teshome Balayneh for guiding me on morphological analysis of the language and providing relevant. I am also grateful to Adama Hospital Medical College staff that have supported me in morally and providing me with good working environment and materials.

Last but not the least my gratitude goes to my family who helped me financially and morally in supporting and encouraging me in this study time.

Dedication

This work is dedicated to my father who passed away during my Diploma Studies.

Tables of Content

Contents **pages**

List of Tables

List of Figures

List of Abbreviations

FFL-Full form lemma

FN –False negative

FP-False positive

FFL -from a Full Form-Lemma

GNU-Gnu's Not UNIX (free software and containing no UNIX code)

OOV- out- of-vocabulary

TN- True negative

TP- True positive

Weka - Waikato Environment for Knowledge Analysis

Abstract

This study has been intended to evaluate lemmatization of the Afan Oromo text performance in natural language processing like information retrieval system, machine learning. The objectives of the study are: to identify the various affixes and the lemma's from the collected written Afan Oromo surface words, to perform text lemmatization and its relationship with stemming and to evaluate its performance.

The study conducted on manually collected 1000 tokens with its candidate lemma by consulting language experts. The collected tokens were analyzed using weka package which was presented in hierarchal clusterer of the candidate lemma 231 was selected and tested in 10,15,and 20 clusters using the single-link criterion with edit distance similarity of the tokens. The study mainly focused on the lemmatization of Afan Oromo text that achieved good performance in its accuracy. The study findings revealed that the single-link of hierarchical clustering with edit distance produced 98.3 % of accuracy.

Using the above findings, it is proved that there is a strong lemmatization performance on Afan Oromo text. According to the study, lemmatization contributes towards a better performance than stemming when we compered the results of each other. The error rate was 1.7% where seen due the manually annotated lemma and the hierarchical clustering algorithms itself. Improvements should be made for the implementation of lemmatization on this language to make easy to all user of the language.

1. INTRODUCTION

1.1 Introduction

In today's era researchers live in a computerized world where almost everything is done by computers [1]. With this advantage, researchers have a lot of data to store and retrieve. This stored data can be accessed for retrieving information from anywhere. This is good application of information retrieval systems. Since Afan Oromo language is so vast and has its own grammar structure, it is very necessary to study the word structure[2]. Here a researcher needs some mechanism of dropping all the word forms to a common base form which is called root word or stem. This is done by stemming or lemmatization. Stemming is the process of reducing the words to a common stem by clipping off its prefixes and suffixes. These prefixes and suffixes are called as morphemes. This suffix stripping is done by generating various rules in stemming. For example – 'introduce' has many word forms like introduction, introducing, introduced. After stemming, all the forms are contracted to the word introduc- which is not a proper root word as it removed the last character e. This problem is resolved by lemmatization where the inflectional ending is detached and the base or dictionary form of the word is generated to 'introduce'. Thus in stemming, there are two kinds of problems. These are under-stemming and over-stemming that leads the stemmed words to meaningless for user. Under-stemming occurs mostly in case of semantically same words[3], For example – under-stemming in Afaan Oromo a word 'Lalisaa**'** which when stemmed gives 'Lal**'**, but there is no a proper root or must leave as 'lalisaa', which means noun of a person. Over-stemming occurs mainly in case of semantically distant words. Example the Afan Oromo word 'maxxantummaa' when stemmed gives 'maxx' which is again no a proper root word. The word maxxantummaa must be removed suffix 'ummaa' and t-> changed into zero morphemes and given correct headwords maxxana that where produced by lemmatization. Thus, the stemmer does not provide the contextual knowledge which is provided by lemmatization. Since English and other European languages are not highly inflected many stemmers and lemmatization have been developed [4], but when we see Afan Oromo, it is highly inflected. In addition, lemmatization for this language has not been found yet. In this paper, the researcher is

emphasizing on Afan Oromo, which is the official language of Oromia Regional State in Ethiopia. So, it is widely spoken in various parts of the region and in neighboring countries like Kenya and Somalia. So, in order to preserve the language and its root words researchers have tried to develop a lemmatization which contains various surface words, suffix, prefixes and found out other methods i.e. hierarchical clustering. The study however does not include all the methods but can be taken as a hierarchical cluster for extending the functionality of the lemmatization. The researcher made an attempt to make lemmatized surface words using single-link algorithm of hierarchical cluster with edit distance to measure the similarity of its words.

1.2. Statement of the Problem

Afan Oromo is one of the major Ethiopian languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries like Kenya and Somalia[5] and [6]. It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 34.5% of the total population[7]. Besides the first language speakers, a number of members of other ethnicities who are in contact with the Afan Oromo's speak it as a second language.

In Ethiopia there are different attempts made to develop stemming algorithms for Afan Oromo, but has never been develop lemmatization for this language. The gap of the stemmer is: the stemmed token has no meaning, above-stemmed and under-stemmed. In order to fill these gaps for this language that faced by stemming algorithm and/or not overcome to respond a real root words for user, and this study necessitates to bring a solution.

In spite of having this lemmatization in mind, this study aims to answer the following research questions:

1.3 Research Questions

- What are the unique features of word formation in Afan Oromo that may be used for lemmatization?
- What makes single-link of hierarchical clustering algorithm optimal for Afan Oromo text lemmatization?

1.4 Objectives of the Research

1.4.1 General Objective

The main objective of this study is to explore or investigate the development of a lemmatization for Afan Oromo text.

1.4.2 Specific Objective

- To review literatures on the topic area specifically, on previous works related to lemmatization of different language and digesting its concept.
- To understand basics of Afan Oromo and structure in writing system
- To develop lemmatization techniques for Afan Oromo text.
- To evaluate the performance of the lemmatization for Afan Oromo and recommend future research direction.

1.5 Significance of the study

The application of the result of this study will be for people who are retrieving information in Afan Oromo and those who are interested in learning Afan Oromo without waiting for any assistance or help from native speakers of the languages. It helps to understand the headword of Afan Oromo that where other words formed from it in lemmatization. It also supports native speakers to represent many words in single lemma or headwords and it also opens a door for other researchers to study on lemmatization of this language.

1.6 Research Methods

To achieve the objectives stated in Section 1.4, the researcher has used the following methods: Primarily, literatures related to study are conducted on lemmatization in different languages, the nature of Oromo language and the structure of the corpus to be lemmatized

for training and testing were investigated. To carry out this study, articles, journal, books, and language experts have been consulted. The various steps that are generally adopted by a researcher in studying his research problem along with the logic behind are method. It is necessary for the researcher to know not only the research methods/techniques but also the methodology. This research has been conducted in order to figure out challenges (under and above stemming and stemmed tokens have no meaning) of stemming in Afan Oromo text. Towards achieving the main objective, the following step by step procedures are followed:

1.6.1 Literature Review

To have conceptual understanding and identify the gap that is not studied or covered by previous studies; different materials, including journal articles, conference papers, and books, have been reviewed. In this study the review is mainly concerned works that have direct relation with the topic and the objective of the study. These include previous works done on local stemming algorithms of Afan Oromo. Journals, conference papers and books for other languages on lemmatization that are given more attention on Information Retrieval were reviewed.

1.6.2 Corpus Preparation

A corpus to evaluate the lemmatization for Afan Oromo text was selected and prepared as there is no previous research and corpora in Afan Oromo for evaluating lemmatization. The prepared corpus consists of thousand tokens, gathered from Afan Oromo-Amharic-English dictionary and Afan Oromo- Deutsch-English that are written in Afan Oromo language with different topics. This corpus has been manually lemmatized by removing the prefixes and suffixes then trained on weka.

1.6.3 Tools Used and Evaluation Technique of Lemmatization

Waikato Environment for Knowledge Analysis (Weka) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand[8]. It is free software licensed under the GNU General Public License. It was used as it contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions[9]. Now days it was used in many different application areas, in particular for educational purposes and

research. This tool has many advantages [9]. These are: free availability under the GNU General Public License, portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform, It is a comprehensive collection of data preprocessing and modeling techniques, ease to use due to its graphical user interfaces. More specifically, it supports data preprocessing, clustering visualization…etc.[10]. another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling. For this study the researcher has used weka 3.6.1 version to analyze the dataset by using different algorithms that enabled lemmatization of Afan Oromo by hierarchical clustering[10].

In the current study, we assigned 231 candidate lemma for 1000 words or tokens for lemmatization manually. Then those words and the candidate Lemmas were imported from attribute relation file format to weka tool. In the weka software, by using hierarchical cluster algorithm of single-link, complete-link and ward algorithms, and edit distance we measured similarity of words. From these algorithms the single-link with the edit distance is performed as the remaining two algorithms took high computation time or did not produce the result of classes of the cluster for the candidate lemma. Accuracy of hierarchical clustering lemma was calculated as sum of correctly clustered lemma generated by hierarchical clustering plus the candidate lemma that was not generated by hierarchical clustering and divided to all the dataset and then multiplied by hundreds.

1.7. Scope, Delimitation and limitations of the study

1.7.1 Scope of the study

The focus of this study is on lemmatizatin of Afan Oromo text that was annotated manually.

1.7.2 Delimitation of the Study

The lemmatization of information that was not in other types or format like as image, video, and graphics and also compound wards are out of focus of the research. The tokens out of thousands and two hundred thirty three lemma were excluded from this study.

1.7.3 Limitation of the Study

The absence of standard test corpus and evaluation tool for Afan Oromo language, financial aspects were limitation. Then, the researcher prepared the small corpus to carry out for the experimentation. The researcher also tested on pre-existing algorithms of weka for the evaluation of lemmatization. However, the amount of corpus prepared for this study is relatively small and requires further development.

1.8. Organization of the thesis

This thesis report is organized into five chapters. The first chapter talks about the motivation behind conducting the research and discusses: background of the study, statement of the problem, the objectives, methodology and scope and limitations of the study. The second chapter presents the basic concepts and related works on automatic text lemmatization. Concerning the basic concepts and discusses its process, types, approaches, techniques and evaluation methods of text lemmatization. Furthermore, has reviewed history of text lemmatization and global works and research works on text lemmatization on local language. Chapter three discusses Afan Oromo language features such as Afan Oromo writing system, morphology, and word formation. Chapter four describes the practical activities carried out on corpus preparation for the experimentation and evaluation and discussion of the result. Finally chapter five gives conclusions and recommendations based on the findings of the study.

2. LITERATURE REVIEW

2.1 Introduction

This chapter is concerned with the review of literature. It covers some general background of Afan Oromo and some typical characteristics of the language. It covers the overview of a lemmatization in relation with Information Retrieval and it approaches. Related works on lemmatization is also covered in this chapter. The chapter also deals with the challenges of lemmatization, and the evaluation technique that is employed for this research.

2.2. A Brief Overview of Afan Oromo

Afan Oromo is among the major languages that are widely spoken and used in Ethiopia[5]]. It is part of the Lowland East Cushitic group within the Cushitic family of the Afro-Asiatic phylum, unlike Amharic (an official language of Ethiopia) which belongs to Semitic family languages. Although it is difficult to identify the actual number of Afan Oromo speaking society (as a mother tongue) , due to lack of appropriate and current information sources.

Afan Oromo has a very rich morphology like other African and Ethiopian languages[11]].With regard to the writing system, 'Qubee' (Latin-based alphabet) has been adopted and became the official script of Afan Oromo since 1991[12]].It is widely used as both written and spoken language in Ethiopia and neighboring countries like Kenya and Somalia[13]]. Currently Afan Oromo is an official language of Oromia Regional State (which is the largest region in Ethiopia) and used as an instructional media for primary and junior secondary schools of the region. It is also delivered in colleges that are found in the regions and in some universities of the country forming its own department. Furthermore, many literature works, newspapers, magazines, educational resources, official documents and religious writings are written and published in this language[14]].

2.3 Basic Concepts of Lemmatization

In information retrieval literature,[15]] we use stemming or lemmatization for preprocessing of text collections in early step. It is very important in various information retrieval applications, such as text classification, information extraction, indexing and searching, because it brings out actual grammatical or semantic relations and it converts words into different forms using two techniques; stemming and lemmatization. The goal of both

stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form[15]].

Stemming usually refers to a process of removing affixes from a word, and returns the stem or root of a word. Users of information retrieval system generally do not worry about how a word has been stemmed, from the correct point. Then, information retrieval system uses stems to index and represent documents with these index terms[16]].

In linguistics, a lemma (from the Greek noun "lemma", "headword") is the "dictionary" or "canonical" form of a set of words[17]]. More specifically, a lemma is the canonical form of a lexeme, where lexeme refers to the set of all the forms that have the same meaning, and lemma refers to the particular form that is chosen as base form to represent the lexeme[18].

It is the task of finding the lemma which is the dictionary form of a given word form. In information retrieval, this unit is also used as the citation form or headword by which it is indexed[19]].

It is a decision in favor of one form of an expression which is considered its (proper) citation form, and against all the other forms which are not[20]. This decision is principled in many cases; in other cases it is more or less discretionary. However, even in the principled cases the dictionary user cannot be expected to know the lemmatization rules by heart, and much less can he be expected to guess the motives behind discretionary decisions. Moreover, many dictionaries, especially bilingual ones, are destined for users with imperfect knowledge of the language in question, who are not expected to know that *worse* is a form of *bad*. Therefore, in order not to frustrate the user who searches a certain expression in the dictionary, reference entries are set up. Such an entry reduces to a (pseudo-)lemma and a reference to the appropriate lemma[20]. For instance: worse: see bad.

2.4. Process of Lemmatization

There are two main processes used for derivation of new words in a language: the inflectional and the derivative process. The words are derived from the same morphological class (for example the form *cleared* and *clears* of the verb *clear*) in the inflectional process while in the derivative process are derived from other morphological classes (*clearly*)[21]. The creation of a new word can be reached by applying a set of derivation that consisted full form words and the candidate lemmas in the both processes. Then stripped the prefixes and

suffixes in added some character to derive a new word form. From this point of view, the lemmatization can be regarded as the inverse operation to the inflectional and derivative processes.

2.5 Training Data of Lemmatization

Data from different sources were used for training of the automatically constructed lemmatization that contains the full form of words and lemmas[22]. This manually annotated word with its candidate lemma has been imported from attribute relation file format to weka that it has been used for analysis. Then, it applies different algorithm of hierarchical clustering with edit distance that measured their similarity. At the end, the researcher has been carried out to evaluate the lemmatization's performance by calculating its accuracy in section 4.

2.6. Approaches of lemmatization

The researchers were reviewed five different lemmatization approaches for experiments. For this purpose, the researcher's used these five different lemmatizes: hand crafted lemmas for lemmatization, morphological Analysis, radix tree based, fixed-length word truncation approach [24]. The last approach was hierarchical clustering algorithm with edit distance that measure words similarity among different words. In this study, the researcher used the hierarchical clustering lemma.

2.6.1 Handcrafted Lemmatization

Morphological analysis of Afan Oromo is not available as part of the selected as the representative of handcrafted lemmatization and after discussion has been carried out with the experts of the language to formulate the full form words and the lemmas. Therefore, Afan Oromo that contains full word forms and Lemmas were arranged in ascending order to provide all possible lemmas for a given word and also a set of all conceivable morphological analysis. The analysis is used the handcrafted lemma tokens and manually stripped set of affix patterns (As is the author's best source of knowledge).

4.6.2. Morphological Analyzer Based Approach

Morphological Analyzer gives all possible analyses for a given word which is based on finite state technology and it produces the morphological analysis of the word form as its output using finite state tools[23]. This approach uses finite state methods or automata and two-level morphology to build a lexicon of a language with essentially an infinite vocabulary[23]. The application of this approach, given different word as input parameter was processed and its feature form representation is generated by a full-scale two-level morphological analyzer. In the next section 4, the researchers have used this method in order to get the lemmas from different words as input; the only single lemma would be displayed as output for all words. All the words and the appropriate lemmas of these words were manually annotated by different linguistics experts or teachers who have thought this subject in the school and universities.

4.6.3. Radix Trie Based Approach

A radix tree (also compact prefix tree) is a space-optimized trie data structure where each node with only one child is merged with its parent [25, 26]. Trie is a data structure which allows retrieving all possible lemmas [26]. Here each node is having single character. Two nodes connected with the edges. Word is retrieved byte by byte. This approach is also involved backtracking for getting appropriate result [25].

4.6.4 Fixed Length Truncation

In this approach, [24] simply truncate the words and use the first 5 and 7 characters of each word as its lemma and words with less than n characters are used as a lemma without truncation. It is the most appropriate for the language like Turkish which has average length of words was of 7.07 letters. This approach is also applicable in Afan Oromo Language as it leaves or not truncates the words that were less or equal to seven in its length as it is considered as lemma. It is the simplest approach not dependent on any language or grammar. For example, words like ‘Bishaan’ water,’ shan’ five, ‘ilkaan’ teeth… were not lemmatized in this approach.

4.6.5 Hierarchical clustering Approach

Hierarchical clustering output a hierarchy; structure that is more informative than the unstructured set of clusters returned by flat clustering [24]. This may be either all the way from single documents up to the whole text set or any part of this complete structure. There are two natural ways of constructing such a hierarchy: bottom-up and top-down. The first principle is used in agglomerative algorithms. The stopping criterion for both algorithms may be that the desired number of cluster is reached or some limitation a criterion function or any internal evaluation measure. The result of hierarchical agglomerative clustering is strongly dependent on the similarity measure[24]. The single-link method defines the similarity between two clusters as the similarity between the two most similar objects, one from each cluster. This may result in elongated, locally similar clusters. Complete-link method is similarity between two clusters is defined as the similarity between the two most dissimilar objects, one from each cluster. Ward algorithm finds the distance of the change is caused by merging the cluster. The information of a cluster is calculated as the error sum of squares of the centroids of the cluster and its members[9].The agglomerative algorithms are deterministic, generating the same cluster hierarchy every time. The similarity definition can be viewed as the criterion function, although it is used locally for each merging and not as a global score[24]] for a discussion of criterion functions).The time complexity of the agglomerative algorithms are $O(n^2)$ as they all need to compute the similarity between all objects to find the pair of objects that are most similar.

2.7. Evaluation Methods of Lemmatization

Approaches of lemmatization are viewed and three algorithm of hierarchical clustering were trained on weka. The first method involves manually lemmatization of tokens by linguistics expert the candidate lemma was assigned to each tokens. The second method was looking-up the form of tokens in a morphological analysis to determine their base form and the methods of each suffix and prefixes leads to single lemmas by stripping their suffixes and prefixes. It was started from individual words or tokens then stripping these suffices based on morphological analysis and applied some morphological rule i.e. some character might be changed into zero morpheme and also remain the lemma. The base form is generated from a stripped words by the application of morphological rules that was produced[25]]. The

fourth methods also applied on different corpus of Afan Oromo but the researcher has not been used for this specific study. The fifth approach was very important. For this study, the researcher used the edit distance to measure tokens similarity or dissimilarity among different tokens that were resulted in clustering lemma. The similarity of the tokens were hierarchically clustered into a single lemma in using the following formulas of string edit distance or the edit distance is the cost of the cheapest sequence of operations(script) turning a string into another[26]]. Allowable operations are insertion, deletion and substitution of symbols. Costs are gathered in a matrix C. The edit distance (Levenshtein)[27],[27]]:

Edit distance: $$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$ ------Equation 1[27]]

The above equation 1 were applied the three operation of edit distance D(i-1)+1 to delete the suffices, D(i,j-1)+1 to insert the some characters and D(i-1,j-1)+ 2;0 was substituted the characters.

$(|\sum| + 1) * (|\sum| + 1)$ ----------Equation (2)

Based on the edit distance measure of similarity a confusion matric would be derived and evaluate the performance lemmatization's precision, recall, and accuracy was calculated by using the following equations:

$$\text{Precision} = TP/(TP + FP)$$----------Equation (3)

The precision measure evaluates the coherence of a cluster, that is, the degree to which a cluster contains lemma cluster from a hierarchical clustering. It can be interpreted as the clustering rate under the assumption that all samples of the cluster are predicted to be members of the actual dominant class for the classes of cluster.

$$\text{Recall} = \frac{TP}{TP+FN}$$ ----------Equation (4)

This equation was calculated the recall of clusters which assigned correctly to the classes of cluster from all the given lemma. It measure the distribution of clusters assigned to the classes of cluster in a given cluster.

Moreover, the accuracy of the overall solution is defined to be the weighted sum of the TP and TN and divided by the overall data value of each cluster, that is,

$$\text{Accuracy} = \left(\frac{\text{TP+TN}}{\text{FP+FN+TP+TN}}\right)$$ ----------Equation (5)

This equation was computed and calculated the performance of hierarchical clustering lemmatization by using single-link algorithm with edit distance measure.

In addition, the evaluation of results should be done manually as well because the hierarchical cluster was assigned the correctly lemma from all lemma that to the classes to cluster,

2.8. Review on Related Text Lemmatization Studies

The researcher has given a general overview of the classical techniques used in lemmatization in the previous section and different techniques that are employed in the next section. The researcher was reviewed different researches that focused on lemmatization that applied different techniques. In this section first focused on reviewing some earliest works, related to global researches and then reviewed all local works in the area of text Lemmatization.

2.8.1 History of Text Lemmatization and Global Related Works

There have been some attempts in creating lemmatization of the texts.

A Novel Arabic Lemmatization Algorithm proposed by[28]] is the first Arabic lemmatization algorithm, and[28]] hypothesis that lemmatization will be more efficient in tokenizing Arabic documents than stemming. In addition to the general stemming benefits, lemmatization can overcome the stemming errors and reduce stemming cost by reducing unnecessary stemming. In their[29]] method they group stop words into useful this can help them in identifying nouns and verbs and direct them into the appropriate stemming and useless stop words are that used extensively and give no benefits to the subsequent words. They also consider their algorithm as advanced stemmer, in which identified nouns and verbs are used to generate global nouns and verbs dictionaries. The benefit of these dictionaries is to find same nouns in the corpus that were used differently in other sentences.

The algorithm they develop were consists of different phases. These are: first phase removes useless stop words, second phase identify nouns by either locating noun preceding stop words or words starting by definite articles. These nouns are lightly stemmed by removing suffixes and prefixes and then added to the global nouns dictionary. At this level, these words are flagged as nouns as preparation to the stemming phase. In parallel to that process they[30]] find verbs by locating verbs preceding stop words. Similar to the nouns, the verbs are added to the global verb dictionary and tagged as verbs. If the word is not a stop word then the word is added to the noun dictionary and flagged as nouns. Before they [31] direct a word to the appropriate stemming by the word flag all the stop words are removed since they offer no further advantage. Other words that do not belong or match in any category will be treated as nouns and stemmed lightly. Finally, root extraction by comparing verbs to Arabic root patterns, words (tokens) with missing tags are considered nouns and lightly stemmed. They[28]] evaluate their experiments in comparing Arabic lemmatization algorithm to the leading root-based stemmer presented by khoja. The design of their experiment's algorithm effects on improving the performance of the k-means clustering in the contrast to the khoja stemmer's effect. They[28]] used the term frequency inverse document frequency weight function and created three different experiments sets. The first and second experiments are to cluster documents belonging two different groups. The last experiment is to study the effect on clustering three different groups on lemmatizing algorithms[30]].

They[31]] deal with the automatic construction of lemmatizes from a Full Form-Lemma (FFL) training dictionary and with lemmatization of new, in the FFL dictionary unseen, i.e. out-of- vocabulary (OOV) words. They[31]] used an automatic technique called OOV Words Lemmatization to train their lemmatizes on Czech, Finnish and English data. Their algorithm uses two pattern tables to handle suffixes as well as prefixes. Three methods of lemmatization of three kinds of out of vocabulary words (missing full forms, unknown words, and compound words) are introduced. These methods were tested on Czech test data the best result has been achieved by a combination of these methods. The lexicon-free

lemmatizer based on the method of lemmatization of unknown words (lemmatization patterns method) is introduced too.

Tarik El- Shishtawy and Fatima El-Ghannam [29]have been studied on an accurate Arabic Root-Based lemmatizer for information retrieval purposes and they proposed the first non-statistical accurate Arabic lemmatizer algorithm that is suitable for information retrieval system. The proposed lemmatizer makes use of different Arabic language knowledge resources to generate accurate lemma form from and its relevant features that support information retrieval purposes. As a part-of- speech tagger, the experimental results show that, the proposed algorithm achieves accuracy above 90.

A rule based approach proposed by Plisson et al. [15] is one of the most accepted lemmatizing algorithms. It is based on the word endings where the suffix should be removed or added to get the normalized form. It emphasizes on two word lemmatization algorithm which is based on if-then rules and the ripple down approach.

Jongejan&Haltrup [[32]**]** constructed trainable lemmatizer for the lexicographical task of finding lemmas outside the existing dictionary, bootstrapping from a training set of full form– lemma pairs extracted from the existing dictionary. This lemmatizer looked only at the suffix part of the words. Its performance was compared with a stemmer using hand crafted stemming rules, the Euroling Site Seeker stemmer for Swedish, Danish and Norwegian. The results showed that lemmatizer were as good as the stemmer for Swedish, slightly better for Danish and Norwegian but worse for Greek. These results are very dependent on the quality (errors, size) and complexity (diacritics, capitals) of the training data.

2.8.2 Local Works of Text Lemmatization

Much work has been done in developing stemmer for Afan Oromo text. In contrast, no work has been done for the development of lemmatization for Afan Oromo to our best knowledge.

In this work, the researcher proposed the first Afaan Oromo lemmatization and hypothesized it will be more efficient in tokenizing documents than stemming. In addition to the general

stemming benefits, lemmatization can overcome the stemming errors and reduce stemming cost by reducing unnecessary stemming and made words meaningful.

3. AFAAN OROMO MORPHOLOGY

3.1. Introduction

Afaan Oromo is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries like Kenya and Somalia[5]]. It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 34.5% of the total population. Besides first language speakers, a number of members of other ethnicities who are in contact with the Oromo's speak it as a second language, for example, the Omotic speaking Bambassi and the Nilo-Saharan-speaking Kwama in north-western Oromia[12]]. Currently, Afaan Oromo is an official language of Oromia regional state (which is the largest Regional State among the current Federal States in Ethiopia). Being the official language, it has been used as medium of instruction for primary and junior secondary schools of the region. Moreover, the language is offered as a subject from grade one throughout the schools of the region and have their own department in higher education. Few literature works, a number of newspapers, magazines, educational resources, official credentials and religious documents are published and available in the language.

It uses Latin based script called "Qubee" and it has 26 basic characters. Oromo language, literature and folklore delivered as a field of study in many universities located in Ethiopia and other countries. Nowadays journal, magazines, newspapers, news, online education, books, entertainment Medias, videos, pictures, are available in electronic format both on the Internet and on offline sources[13]].

In general, Afaan Oromo is widely used as written and spoken language in Ethiopia and neighboring courtiers like Kenya and Somalia .With regard to the writing system, **"Qubee**" (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991[5]].

The remaining sections of this chapter discusses: Afaan Oromo Alphabet and writing system and usage, Afaan Oromo morphology, word and their formation structure of noun and adjectives.

3.2 Afaan Oromo Alphabets and Writing System

According[7]] Afaan Oromo is a phonetic language, which means that it is spoken in the way it is written. The writing system of the language is straight forward which is designed based on the Latin script. Unlike English or other Latin based languages there are no skipped or unpronounced sounds/alphabets in the language. Every alphabet is to be pronounced in a clear short/quick or long /stretched sounds. In a word where consonant is doubled the sounds are more emphasized. Besides, in a word where the vowels are doubled the sounds are stretched or elongated.

Like in English, Afaan Oromo has vowels and consonants. Afaan Oromo vowels are represented by the five basic letters such as a, e, i, o, u. Besides, it has the typical Eastern Cushitic set of five short and five long vowels by doubling the five vowel letters: 'aa','ee', 'ii', 'oo', 'uu' [7]].

Consonants, on the other hand , do not differ greatly from English, but there are few special combinations such as "**ch**" and "**sh**" (same sound as English),"**dh**" in Afaan Oromo is like an English "*d*" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins. Another Afaan Oromo consonant is "**ph**" made when with a smack of the lips toward the outside "**ny**" closely resembles the English sound of "gn". We commonly use these few special combination letters to form words. For instance, **ch** used in **hubachiisa**'*note'*, **sh** used in **shamarree '***girl'*, **dh** use in **Dhukkuba** '*Sick'* , **ph** used in **buuphaa '***egg',* and **ny** used in **nyaata** '*food'* .

In general, Afaan Oromo has 36 letters (28 consonants and 10 vowels) called "**Qubee".** In general, all letters in English language are also in Afaan Oromo except the way it is written.

Table 1:The consonants of Afaan Oromo and the place they were formed.

Condition of phonemes	Bilabial/ /hidhii/	Alveolar/ irgata/	Palatal /laagee/	Velar /harsaso/	Glottal /kokkee/
Stop/iggata/: Voiced Voiceless	B (p)	D T		G K	?
Ijective/dhiitata	Ph	X	C	Q	
Implosive/kedhowaa		Dh			
Fricative/loyaa/: Voiced Voiceless	(v) F	(z) S,(ts)	J sh		 H
Rigata Voiced Voiceless			Ch		
Nasal /funyataa/	M	N	Ny		
Lateral /maddata/		L,r			
Flap /batoo/	W		Y		

Table 2: Afaan Oromo vowels Alphabet (source [10, 5]

	Front	Central	Back
High	i , ii	u , uu	
Mid	e , ee	o , oo	
Low	a	Aa	

3.3 Afaan Oromo Morphology

Morphology is a branch of linguistics that studies and describes how words are formed in a language[7]]. There are two types of morphology: **inflectional** and **derivational**. **Inflectional morphology** is concerned with the inflectional changes in words where word stems are combined with grammatical markers for things like person, gender, number, tense, case and mode. Inflectional changes do not result in changes of parts of speech. On the other hand, **derivational morphology** deals with those changes that result in changing classes of words (changes in the part of speech). For instance, a noun or an adjective may be derived from a verb. In this thesis we are intended to lemmatize the noun or an adjective that are formed from the verbs and then it becomes a base or root word that have full meaning**.**

3.3.1 Types of morphemes in Afaan Oromo

A morpheme is the smallest semantically meaningful unit in a language. A morpheme is not identical to a word, and the principal difference between the two is that a morpheme may or may not stand alone, whereas a word, by definition, is a free standing unit of meaning. Every word comprises one or more morphemes[33]. In Afaan Oromo, there are two categories of morphemes: free and bound morphemes. Free morpheme can stand as word on its own whereas bound morpheme does not occur as a word on its own[34]].In Afaan Oromo roots (stems) are bound as they cannot occur on their own.

Example: "**dhug-**" (*drink*) and "**beek-**" (*know*), which are pronounceable only when other completing affixes are added to them[11]. But lemmatizing can make bound morpheme to

occur by its own in addition of a remaining characters to the root words like **Dhug + aa+ ‘’Dhugaa’’**(drink or true) and **Beek + aa= Beekaa(know or noun(name of the person Beekaa)).**

Similarly an affix is also a **bounded morpheme** that cannot occur independently. It is attached in some manner to the root, which serves as a base. These affixes are of three types – **prefix**, **suffix** and **infix.** The first and the second types of affixes occur at the beginning and at the end of a root respectively in creating a word whereas the infix occurs in between characters of the word. In’ **dhugaatii’ ‘***drink’,* for example, **-tii** is a suffix and **dhugaa-** is a base root and this studies also stripped the prefix and suffixes that is attached on stem in order to get the full form lemma of the words. Moreover, an infix is a morpheme that is inserted within morpheme. According to many linguistics scholars there is no Afaan Oromo infix like as[7]] it is discovered that Afaan Oromo does not have infixes like English.

3.3.2 Word Formation in Afaan Oromo

There are many ways of word formation in Afaan Oromo. These morphological analyses of the language are organized in six categories[7]]. These categories are: nouns, verbs, adjectives, adverbs, functional words, and conjunctions. Almost all Afaan Oromo nouns in a given text have person, number, gender, and possession markers which are concatenated and affixed to a stem or singular noun form[[34]]. Afaan Oromo verbs are also highly inflected for gender, person, number and tenses[7]]. Adjectives in Afaan Oromo are also inflected for gender and number. Moreover, adverbs can be categorized into: adverb of time, adverb of place, and adverb of manner in which some of the adverbs are affixed.

Furthermore, functional words can be classified as prepositions; postpositions and articles markers which are often indicated through affixes in Afaan Oromo[11]]. Lastly, conjunctions can be separate words (subordinating or coordinating), and some of them are affixed. Since Afaan Oromo is morphologically very productive, derivation, reduplication and compounding are also common in the language[11]].The following is detail descriptions and examples of word formation process of Afaan Oromo based on the works of[7]].

3.3.2.1 Nouns

i. Gender

Afaan Oromo has a two gender system (feminine and masculine). Most nouns are not marked by gender affixes[7]]. Only a limited group of nouns differ by using different suffixes for the masculine and the feminine form. The language use –**ssa** for masculine and -**ttii** for feminine.

Example:

Obboleessa *brother* – **Obboleettii***sister*

Ogeessa*expert* (m.)**- Ogeettii***expert* (f.)

Natural female gender corresponds to grammatical feminine gender. The sun, moon, stars and other astronomic bodies are usually feminine. In[7]] some Afaan Oromo dialects geographical terms such as names of towns, countries, rivers, etc are feminine, in other dialects such terms are treated as masculine nouns.

There are also suffixes like **-a, -e** that indicate present and past form of masculine markers respectively. **–ti**and**–tii** for present feminine marker and –***te*** past tense marker, *-du* for making adjective form[7]].

ii. Number

Afaan Oromo has different suffixes to form the plural of a noun. The use of different suffixes differs from dialects to dialects. In connection with numbers the plural suffix is very often considered unnecessary.

According to[7]] the majority of plural nouns are formed by using the suffixes – **oota**, followed by **–lee**, **-wwan**, **-een**, **-olii**/ **-olee** and –**aan**. Other suffixes like –**iin** in

Sariin *dogs* are found rarely.

Example:

-oota **hiriyyoota***friends* **jechoota***words*

-lee gaafilee *questions* **kitaabilee***books*

-wwan saawwan*cows* **hojiiwwan***works*

-olii/-oleegangoolii*mules* **jarsolii/jarsolee***elders*

-een **fardeen***horses* **mukkeen***trees*

-aan **ilmaan***children*

iii. Definiteness

Afaan Oromo language does not possess a special word class of articles. Instead demonstrative pronouns are used to express definiteness[7].

manni kun *this/ the house (Subject)*

mana kana *this/ the house (Object)*

manni sun *that/ the house (Subject)*

mana sana*that / the house (Object)*

To express indefiniteness emphatically the Oromo speaker my use numerical **tokko** *one*,

Example:

Namni tokko*one / a man*.

In some Afaan Oromo dialects the suffix **-icha**(m.), **-ittii(**n**)**(f.) which usually has a singularize function is used where other languages would use a definite article.

Example:

jaarsichi*the old man (Subject*) **jaarsicha***the old man (Object)*

jaartittiin*the old women (Subject)***jaartittii***the old lady (Object)*

iv. Derived noun formation in Afaan Oromo

Afaan Oromo is very productive in word formation by different means. The most common word formation methods are derivational and compounding[7]].

A. Derivation Suffixes

Derivational suffixes are added to the root or stem of the word. From derived verbal stem and adjectives may be formed by means of derivational suffixes[7]]. The following suffixes play an important role in Afaan Oromo word derivation. They are **-eenya, -ina, -ummaa, -annoo, -ii, -ee, -a, -iinsa, -aa,-i(tii), -umsa, -oota, -aata,** and **–ooma.**

Examples:

jabaa *strong*

jabeenya *strength*

jabina *strength ,hardness*

jabee *intensive*

jabummaa *strength*

jabaachuu *to be strong*

jabaachisuu *to make strong*

jabeessuu*to make strong*

jajabaachuu*to be consoled*

jabeefachuu **t***o make strong for one self*

Our lemmatization was working on these words to strip the suffixes and displayed the head words that all the words were formed from a single word(**jabaa**) and the other suffixes was chopped off to return a lemma of ‘jabaa’.

A.1suffixes that change nouns into abstract nouns are: 'eenya','umma' and 'ooma'.

The above suffixes changed the noun (**'Nagaa'**) into a derived noun('**Nageenya**') by adding suffix –'**eenya**' on the stem 'nag-' but the researcher stripped the suffix that attached to the noun and added two vowels(-**aa**) in lemmatization. When the two nouns are seen (i.e. noun and the formed abstract noun) they are within the same class of words but they are different in the way they are written[11]].

Table 3: shows different suffixes attached on nouns and they formed an abstract noun [10]

Derived noun	suffixes	Stem of noun	Lemma or noun	glossary
nageenya	-eenya	nag-	Nagaa	peaceful
jabeenya	-eenya	jab-	Jabaa	strength
namummaa	-ummaa	nam-	Nama	humanity
sabummaa	-ummaa	sab-	Saba	Nationality
obbolummaa	-ummaa	obb-	Obbolaa	Friendship
garbummaa	-ummaa	gurb-	Garba	Slavery
oromummaa	-ummaa	orom-	Oromoo	Oromian
firooma	-ooma	fir-	Fira	Friendship
namooma	-ooma	nam-	Nama	Humanity
ollooma	-ooma	oll	Ollaa	Neighborhood

The suffix –'umma' formed other derived nouns by suffixing on physically seen nouns and some abstract nouns as depicted on table above.

When the researcher was looking at the derived abstract noun form ('namummaa','subummaa','obbolummaa', 'garbummaa', and 'oromummaa' have the stem (nam-,sab-,obbol-,garb- and orom-) but the these root words were added to the character 'a' or 'aa' to make the noun of ward i.e the lemma or the canonical base words.

The suffix –‘ooma’ formed a new other derived nouns by suffixing on physically seen nouns and some unseen nouns.

From the noun listed under first column of table 3 above are all unseen nouns except **‘nama’** but the other word categories are unseen nouns and [11]] are changed into abstract nouns by taking the suffix –‘ooma’. The formed structure will be as follow:

Root noun + suffix] ⟶ **Abstract noun**

In general, the suffixes that are attached on nouns and formed derived nouns are [-eenya, -umma, -ooma[[11]].

A.2 Suffixes that changed verbs into Nouns (Abstract nouns)

As explained in[11]] the suffix that was attached on nouns and made derived nouns are also suffixed and attached on verbs and formed a derived nouns/ abstract nouns are:‘-eenya’,’-ina’,’-noo’,…. Etc.

Table 4: shows the derived nouns formed from different verbs [10].

Derived noun	Suffixes	Verbs	Stem of verbs	Lemma	Glossary
Qabeenya	-eenya	Qabaachuu	Qab-	Qaba	To Have
Jireenya	-eenya	Jiraachuu	Jir-	Jira	To Live
Rakkina	-ina	Rakkachuu	Rakk-	Rakkoo	Miserable
Guddina	-ina	Guddachuu	Gud-	Gudaa	Large
Baldhina	-ina	Baldhachuu	Baldh-	Baldha	Wide
Hammina	-ina	hammaachuu	Ham-	Hamaa	Dangerous
Qorannoo	-noo	Qorachuu	Qor-	Qora	To examine
Filannoo	-noo	Filachuu	Fil	Fila	To Select
Tiksee	-ee	Tiksuu	Tiks	Tika	To Keep
Falmii-	-ii	Falmuu	Falm	Falmii	To Argue

The words that are listed above under third column of table 4 are verbs; the words listed under fourth column are stem of the verbs of column three and their suffixes that formed the derived nouns are listed under the second column of the tables. The words listed under first column are derived nouns formed by attaching the suffix (‘eenya’,) at the end of the stems and lemma listed under fifth column that the researcher has used in section four of this study. Its structure is

In this example the suffix (‘-ina’) attached on the verbs were created or formed the derived nouns(rakkina, gudina, hammina) and also the formed nouns changed into the lemma i.e. adjective words and formed derived nouns are seen in detail[11]]. This suffix can change their forms and the words class to be identified in the way it has changed the verb’s stem into noun. In general, they have been explained in detail the following form.

Form = stem of the verb + suffixed ⟶ Derived Noun

A.3. Suffixes that change Adjectives to Nouns

The suffixes that change the adjective into the derived noun are ‘**eenya’, ‘ummaa’, ‘ina’ and ‘ooma.** When a word is attached with those suffixes, it changes the word classes and its forms to the derived noun as depicted in table 5 below:

Table 5: Derived nouns that forms from adjective by attaching some suffixes [10]

Derived noun	Suffixes	Stem of adj.	Lemma or adjective
Fageenya	Eenya	Fag-	Fagoo
Hammeenya	Eenya	Ham-	Hamaa
Dhiyeenya	Eenya	Dhiy-	Dhiyoo
Gowwummaa	Ummaa	Gow-	Gowwaa
Gootummaa	Ummaa	Goot-	Goota
Gadhummaa	Ummaa	Gadh-	Gadhee
Diinummaa	Ummaa	Diin-	Diina
Bal'ina	-ina	Bal'-	Bal'aa
Furdina	-ina	Furd-	Furdaa
Jabina	-ina	Jab-	Jabaa
Collooma	-ooma	Coll-	Collee
Argooma	-ooma	Arj-	Arjaa
Gamnooma	-ooma	Gamn	Gamna

The suffixes that are attached on an adjective change the forms and classes of the words in each adjective because it is changed to derived nouns and it has the following forms[11]].

Stems of the Adjective + [suffixes] ⟶ **Derived nouns.**

As[11]] tried to show in the above example, some suffixes are attached on both word classes. For example, the suffixes **[-eenya]** could be attached on nouns, verbs, and adjectives, but the suffix [**ooma]** can also be attached on some nouns and adjective. The suffixes that are attached on nouns can form the derived nouns and it changes the form of the word but not the classes of words. But the suffixes that are attached on verbs and adjectives can also change the forms and classes of the word. In general, except the original nouns in the language, including the nouns form in different classes of words, suffixes can be added to make derived nouns[11]].

On the other hand; the adjectives, that plural maker which were reduplicated by first two or three charcters were simply deleted without adding other characters (for example jajjabaa, dhedheera,diddiima, guggurraacha…etc.). In addition,the adjectives that changed their word

classes also stripped their suffix and either they can added some character like jabaachuu, jabina, jabaatan are stripped suffixes –chuu,-ina and -tan and simply added –aa on the stem of jab and the remaining two word not added any characters and give the headward jabaa[7], [11], [11]]

Table 6: Afan Oromo Adjectives for number example [10]

Singular		**Plural(plural)**	
Male	**Female**	**Male**	**Female**
Diimaa	Diimtuu	Diddiima	Diddiimtu
Gabaabaa	Gabaabdu	Gaggabaaba	Gaggabaabdu
Adii	Adii	A'adii/Adaadii	A'adii/Adaadii*
Dheeraa	Dheertuu	Dhedheera	Dhedheertu
Gurraacha	Gurraatti	Guggurraacha	Guggurraatti
Furda	Furdoo	Fuffurdaa	Fuffurdoo
Guddaa	Guddoo	Gugguddaa	Gugguddoo
Laafaa	Laaftuu	Lallaafaa	Lallaafaa
Jabaa	Jabduu	Jajjabaa	Jajjabduu

3.3.2.1.1 Formation of compound Nouns

Compound nouns are formed by joining two different classes of words or two or more the same classes of nouns. It can be formed by different ways by joining different words.

Example: 'Mata' + 'dura'='Mataduree'

' Abbaa' + 'Lammii' ='Abbalammii'

'Malkaa' + 'Bal'oo' ='Malka bal'oo'

In[11]the above example the compound noun 'mataduree' can be formed from noun 'mata' and the preposition 'dura'. The preposition 'dura' changed its forms to 'duree'. The

compound nouns ‘abbalammii’ is also formed from two nouns ‘abbaa’ and ‘Lammii’. The compound nouns ‘malka bal’oo’ is formed to from the noun ‘malkaa’ and the adjective ‘bal’aa’. The noun ‘mata’,’abbaa’ and ‘malkaa’ ended in long vowels but they are lost the last character or letter when they join to form the compound nouns.

Some compound nouns are formed by joining the words of the first nouns by deleting the last one or two characters and directly attached to the last words.

Example: ‘saree’ + ‘diida’ =’sardiidaa’ or ‘sardiidoo

‘Sab’ + ‘Lammii’ = ‘Sablammii’

Some compound nouns are formed by joining the two words without adding or deleting the any character but they stand for a single purpose.

Example:

‘Abbaa’ + ‘Bokkuu’ = ‘Abbaa bokkuu’

‘Abbaa’ + ‘Alangaa’ = ‘Abbaa Alangaa’

The other compound nouns are formed by joining two different words and add a single character at the end of compound nouns at the last wards.

Example:

‘Galmee’ + ‘jecha’ = ‘Galmee jechootaa’

‘Abbaa’ + ‘duula’ = ‘Abbaa duulaa’

‘Abbaa’ + ‘seera’ =’Abbaa seeraa’

3.3.2.2 Verbs

Verbs are a content word that denotes an action, occurrence, or state of existence[11]].Afaan Oromo has base (stem) verbs and four derived verbs from the stem[34]]. Moreover, verbs in Afaan Oromo are inflected for gender, person, number and tenses.

i. Derived Verbs

There are four derived verbs in the formation which are still productive in Afaan Oromo are: Autobenefactive (AS), Passive (PS), Causative (CS) and Intensive (IS)

Passive, causative, and autobenefactive are formed with addition of a suffix to the root, yielding the stem that the inflectional suffixes are added to[7]]. The personal terminations according to different conjunctions are added to these affixes. The intensive stem is formed by reduplicating the first consonant and vowel of the first syllable of the stem. The derived stems may be formed from all verbs the meaning of which permits it[34]].

a. Autobenefactive

The Afaan Oromo autobenefactive (or "middle" or "reflexive-middle") is formed by adding **-a)adh**, **-(a)ach** or **-(a)at** or sometimes **-edh**, **-ech** *or* –**et** to the verb root. This stem has the function to express an action done for the benefit of the agent himself and the lemma of these verbs changed in to nouns in order to get their headwards.

Example:

Bitachuu to buy for oneself the base root verb in this case is **bittaa-**

Qoodachuu to share oneself, its lemma was **qooda.**

b. Passive

The Oromo passive corresponds are close to the English passive in function. It is formed by adding **-am** to the verb root. The resulting base root is removing **m** conjugated regularly.

Example:

‘beek- know ‘beekam- *‘be known*

*‘*Argi*’* see ‘argamuu*’ be seen*

c. Causative

The Afaan Oromo causative of a verb corresponds to English expressions such as 'cause ','make ', 'let '.With intransitive verbs, it has a transitive function. It is formed by adding **-s, -sis, or –siis** to the verb root example:

Deemuu to go **deemsisuu**to cause to go **Deema** go is the lemma.

d. Intensive

It is the derived verbal aspects (frequentative) or "intensive," formed by copying the first consonant and vowel of the verb root and geminating the second occurrence of the initial consonant[33]]. It is formed by duplication of the initial consonant and the following vowel, geminating the consonant and sometimes it may not geminate the consonants. The resulting verb indicates the repetition or intensive performance of the action of the verb[7]].

Example:

Verb	gloss	intensive	gloss	lemma
Cir	to turn	ciccira	to turn intensively	cira
Cab	to break	caccaba	to break in pieces	caba

The form of intensive verbs that have not geminated syllabi is: cir,cab, ….have Cvc + verb's stem forms and the verbs that have geminated syllabi have the form of cv + stems of that verb[7], [11]]. Therefore; it has the form of cv(c) or cv + stems of their verbs.

3.3.2.5 Pre-, Post, and Para-positions

Afaan Oromo language uses prepositions, postpositions and para-positions[34]]:

i. Prepositions

In Afaan Oromo prepositions can come before and after the word they are coordinating in their formation.

akka*like, according to*

eegathen

eeggasii then after

gara *to, in the direction of*

gad(i) under

gama in direction to

hanga/hamma *until, up to*

karaa*along, the way of , through*

oliabove

waa'ee case ,about, matter of

The prepositions **gara**, **hanga**, and **waa'ee/waayee** are still treated as nouns and therefore they are used in a genitive construction with other noun they belong to, expression[7]: the direction to, the matter of, etc.

ii. **Postpositions**

Postpositions can be grouped into suffixed and independent words.

a. Suffixed postpositions

-ala out

-bira side of

-tti *in, at, to*

-rra/irra *on*

-rraa/irraa *out of, from*

The post position –**tti** is used to form the locative. The postposition **-rraa/irra** may be used to express a meaning similar to ablative.

Example: **Adaamaatti yoom deebina ?***When shall we go back to Adama*?

Gammachuun sireerra ciise .*Gemachu lay down in bed.*

b. Post position as independent words

ala *outside* **wajjiin** *with , together with*

bira *beside***teellaa** *behind*

Example: **Namoota nu bira jiraniis hin jeeqnu.**

We don't hurt people who are with us.

iii. Para-positions

Gara… tti *to*

Gara… tiin*from the direction of*

3.3.2.6 Conjunctions

Conjunctions are unchanging words which coordinate sentences or single ports of a sentence. The main task of conjunctions is to be a syntactical formative element that establishes grammatical and logical relation between the coordinated constituents. According to[7], [11]] the main functions of conjunctions are identified as: the function of coordinating clauses (coordination), the function of coordinating parts of sentence (coordination) and the function of coordinating syntactical unequal clauses(subordination) . On the other hand, with regard to their form we can subdivide the conjunctions of Afaan Oromo into two:

a. Coordinating Conjunctions: Are coordinate or join two or more sentences, main clauses, words, and other parts of speech which are of the same syntactic importance or grammatical equals emphasis to the pair of the main clauses[11]]. Few of the coordinating conjunctions are listed below:

Akkasumas whether

asAmmoo fi and

Garuu, malee *but*

Kana malees In addition

Haa ta'uu malee Nor

Kanaaf, kanaafuu so

Ta'ullee/-llee,-fi So as

Yookan/yookiin or

b. subordinating conjunction

Subordinating conjunction joins a subordinate clause to a main clause. It is always followed by a clause[11]].

E.g**akka***that, as if , as whether*otoo… dura/-yyuu

Erga,eegawhile erga…booddee/booda

Otuu,utuuif yoo…malee/-illee Althought

The complexity of Afaan Oromo like other morphologically reached languages increases the load on professionals working in the field of natural language processes like information retrieval. Morphology adds a burden to information retrieval works. For the purpose of text lemmatization and also other information retrievals, the variant words of a morpheme should be reduced to their headword so that they can be counted as one while calculating term frequency, thereby increase the performance of the lemmatization. Using lemmatization is believed to minimize the difficulty of dealing with different forms of a word. Lemmatizing is the process for reducing inflected or derived words to their base ward or head word. Hierarchical clustering Lemmatization also assigns the candidate lemma to each full form word automatically based on manually annotated. There have not been efforts in developing of lemmatization for Afaan Oromo. For this study the researcher used the Hierarchical clustering algorithm, the inverse of stemmer developed by [11]]for this work.

4. Experimentation and Interpretation of Lemmatization of Afan Oromo

4.1 Introduction

This chapter deals with the lemmatization processes, morphological analysis, and interpretation of lemmatization of Afan Oromo text. We also used the hierarchical clustering for experimentation that was applied on single-link, complete-link and ward algorithm with edit distance to measure their similarity. Hierarchical clustering was also deals with experiment, evaluated and discussed the results with the previous research of stemmer.

4.2 Lemmatization Processes

The tokens that were collected from different sources listed under appendix I were stop words free then applying the lemmatization on the remaining tokens to categorized it into its full form wards and the candidate lemma in using hierarchical clustering. To produce the correct lemma the edit distance were used with the single link algorithm to measure its similarity in local criterion methods.

For example, the Afan Oromo plural makers of suffix (yakkamtan, yakkamtoonni, yakkamtoota, yakkamuu, yakkamtuu, yakkamaniiru) have been analyzed morphologically as depicted in figure 1 below. From this point of view the lemmatization can be regarded as the inverse operation to the inflectional and derivative processes. This approach is advantageous, because it does not need to create the lemmatization from scratch, but[21] can obtain them through the inversion of derivation of stemming. To do the mentioned stripping and adding on lemmatization the edit distance measure has been employed as to it has three operation (deletion, insertion and substitution) of the prefix and the suffix are obligatory to get the candidate lemma or base words from the given full form words. To show its morphological analysis figure 1 below depicted it in detail.

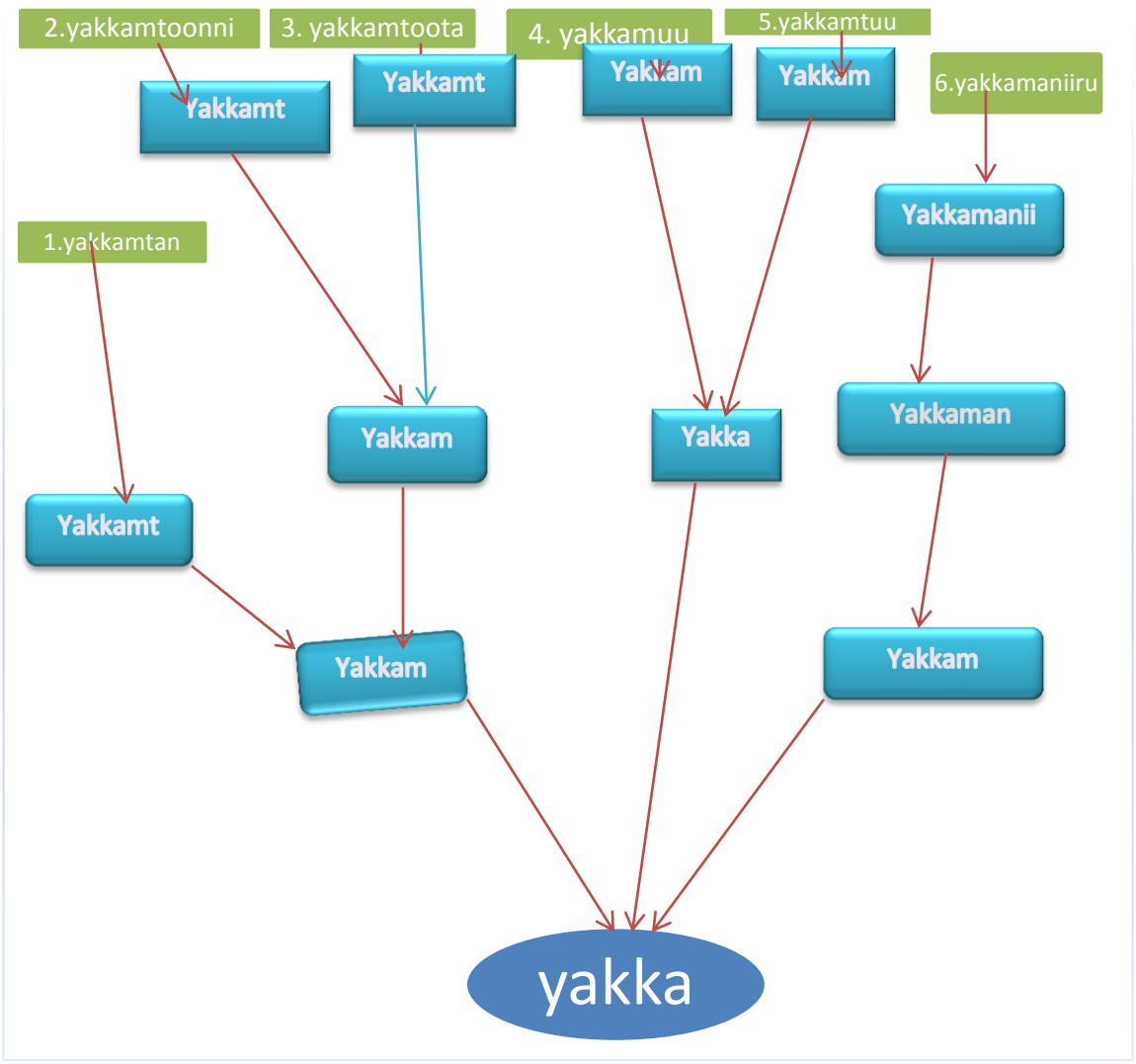

Figure 1: The morphological analysis method [10]

In the above morphological analysis method all the input words have been attached with suffixes and these suffixes have been removed step by step and the un removed characters also changed into zero morphemes (Ø) according to morphological analysis. For instance, the character t-> Ø, m-> Ø as the researcher consulted the linguistics expert. Then all the nodes of the figure have been led to a single output or lemma (**Yakka**) as in the above six words on figure1.

4.3 Experiment

It is very difficult to carry out a systematic study to compare the impact of lemmatization matric on hierarchical cluster quality, because objectively evaluating hierarchical cluster of lemma is difficult in itself. In practice, manually assigned category labels are usually used as baseline criteria for evaluating clusters. As a result, the clusters, which are generated in an unsupervised way, are compared to the pre-defined category structure, which is normally created by human experts of lemma. This kind of evaluation assumes that the objective of hierarchical clustering lemmatization is to replicate human thinking, so the hierarchical clustering solution is good if the clusters are consistent with the manually created categories. However, in practice datasets often come without any manually created categories and this is the exact point where clustering can help. In this case, measures like hierarchical clustering coherence in terms of the similarity words used the maximum closeness of hierarchical cluster distances is conducted in a single-link algorithm with edit distance of the terms between cluster distances can be used for evaluation[35].

In order to make the result of this investigation comparable to previous researches, we choose datasets that have been commonly used for evaluating clustering as the experiment datasets. All of these datasets have appropriately pre-assigned classes of lemma clusters. The rest of this section first describes the characteristics of the datasets, then explains the evaluation measures, and finally presents and analyzes the experiment results.

4.3.1 Datasets

The thousands tokens, gathered from different source like [22]and manual assigned two hundred thirty three candidate lemma for each tokens are used as a datasets Table 1 is chosen for the experiment. These have been widely used for evaluating feature hierarchical clustering on Afaan Oromo text lemmatization. Except few of them have been collected from Afaan Oromo-Deutsch-English online dictionary[36]. These datasets differ in terms of document type, number of categories, average category size, and subjects.

The characteristics and sources of these datasets are summarized and listed on Appendix I. The dataset contains 1000 different words and 231 different candidate lemma or headwords

that are assigned manually to these words. The numbers of cluster classes in each dataset are 1 to 17 classes of cluster instances.

4.4 Analysis of Hierarchical Clustering Lemmatization

Hierarchical clustering (hierarchic clustering) outputs a hierarchy, a structure that is more informative than the unstructured set of clusters returned by flat clustering[37]. Hierarchical clustering does not require us to pre-specify the number of clusters and most hierarchical algorithms that have been used in information retrieval are deterministic[38]. These advantages of hierarchical clustering come at the cost of lower efficiency. The most common hierarchical clustering algorithms have a complexity that is at least quadratic in the number of documents compared to the linear complexity of Single-link[15],[38]

Hierarchical clustering algorithms are either top-down or bottom-up[38]. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative cluster[37]. The following figure shows the analysis of hierarchical clustering is used single-link algorithms and also applied the edit distance to measure their similarity then clusters to classes of clusters.

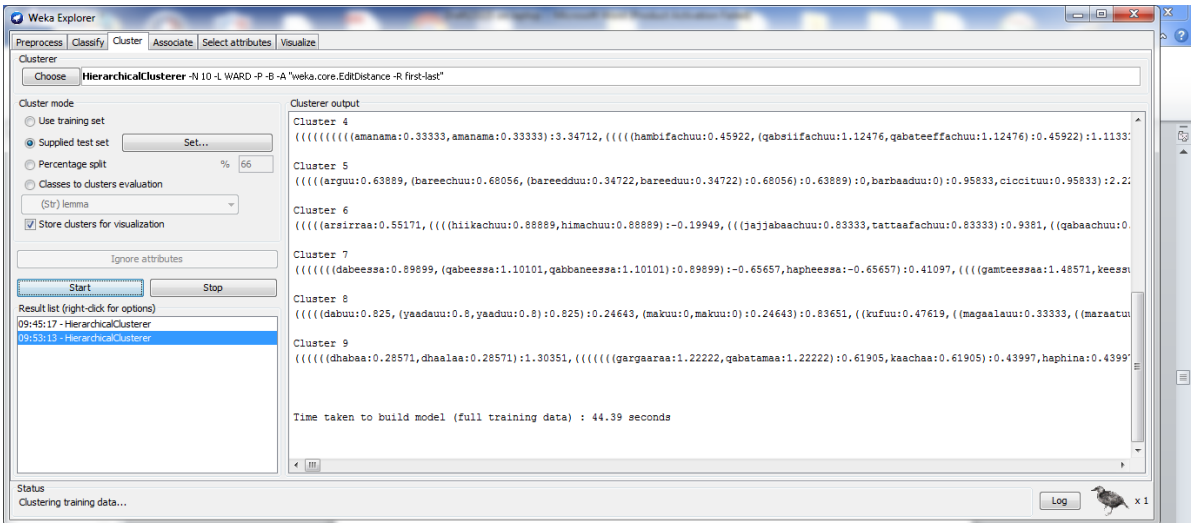

Figure 2: The similarity between words calculated in edit distance and measured by a single-link of hierarchical clusterer output of clusterer.

Figure 2 showed that the full form words with the result of edit distance calculation of their similarity measure by using the single-link algorithm of hierarchical clusterer for twenty two numbers of clusters. It also showed that the time taken to build a model on full training data.

The researcher tested 5 clusters to the classes of cluster and each of them achieved the same result for complete-link and ward algorithms. But when we increased the number of cluster to 10-20 the ward and complete-link algorithm did not produced the classes of cluster lemma but they produced the similarity of the edit distance each words. For example: (dhabaa: 0.28571, dhaalaa: 0.28571):1.30351 and (adduyyaalee: 0.45455, daangaalee: 0.45455): 0.23183, respectively.

But the single-link algorithm could respond for 5-20 numbers of tests of the cluster to the classes of cluster that was used the edit distance similarity measures since it used the maximum closeness of the tokens. In here, the single-link algorithm achieved the best results, because it was the better-merge persistent than the two remaining algorithm of hierarchical clustering for lemmatization. Its merging criterion was local and it was also optimal with respect to the wrong criterion in many documents application.

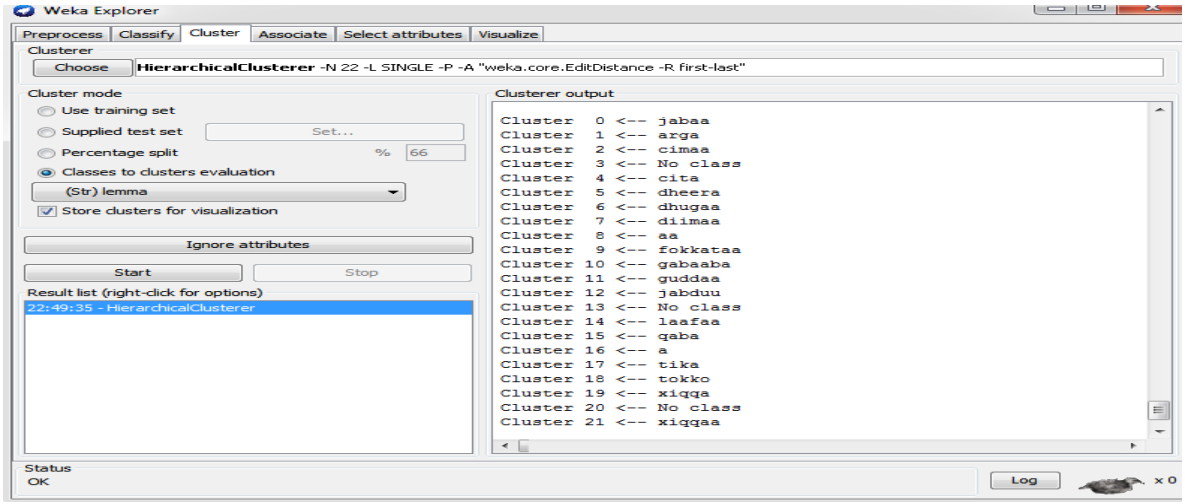

Figure 3: The result of cluster assigned to classes of clusterer out 20 for testing's

Figure 3 shows the classes of cluster that produced the number of classes of clusters out of twenty two testing numbers of clusters. Based on this figure 3, the confusion matrix were derived by counting the number cluster correctly assigned to classes of clusters was seventeen, incorrectly cluster was two and no classes to cluster was three. From this point we could calculated precision, recall and accuracy of lemmatization to measure the performance of hierarchical clustering of single-link algorithm by using edit distance they make a criterion for local similarity of words that are categorized in a single lemma. The tokens that are tested on lemmatization by single-link, complete-link, and wards algorithms with edit distance are used to evaluate the similarity of tokens to cluster with each classes of cluster. For each token, the researcher experimented with different similarity measures (single-link, complete –link and wards) and the researcher has conducted $3 \times 10 = 30$ experiments totally.

Moreover, each experiment was repeatedly run 10 times and the results are the averaged value over 10 runs. Each run has different initial and different times of finishing the run in hierarchical clustering for lemmatization.

4.4.1 Evaluation

For each of the above datasets, we obtained a clustering result from the Hierarchical clustering of single-link, complete link and ward algorithms by conducting edit distance similarity measure[39]. The number of total data elements in clusters is set the same with the number of pre-assigned categories in the data set[40]. The performances of a clustering result was evaluated using three evaluation measures: precision, recall, and accuracy, which are widely used to evaluate the performance of unsupervised learning algorithms[41]]. Before, the three evaluations, we compared the three algorithms. In the first step of the algorithm comparison, five lemmas (initial size) were given to each algorithm and the result is given in table 7 below. At this stage complete link and ward algorithms performed best (100%). For the second step comparison, we doubled the number of lemmas to 10, in the third we tripled to 15 and at fourth step to 20. Starting from the second step, only single link algorithm continued to give result and its performance at the fourth step was 80.0% while the other two algorithms yield 0%. See tables 7 to 10 bellow.

Table 7: Comparison of the three algorithms in five cluster testing

Type of algorithm	**Five cluster of lemma**	**seconds**	**Correctly cluster lemma**	**Incorrectly cluster lemma**
Single-link	5	6.48	3	2
Complete-link	5	4.5	5	0
Ward	5	46.25	5	0

Table 7 compares the three algorithm of hierarchical clustering experiment for five testing classes of cluster using edit distance depicted their time taken to finished running, numbers of classes of clusters correctly and incorrectly clustered lemma for five clusters of lemma.

Table 8: Comparison of the three algorithms for ten cluster lemma testing

Type of algorithm	**Number of cluster lemma**	**seconds**	**Correctly cluster lemma**	**Incorrectly cluster lemma**
Single-link	10	5.26	8	2
Complete-link	10	4.3	0	0
Ward	10	38.22	0	0

Table 9: Comparison of the three algorithms in for fifteen cluster lemma testing

Type of algorithm	Number of cluster lemma	Seconds	Correctly cluster lemma	Incorrectly cluster lemma
Single-link	15	3.48	13	2
Complete-link	15	Too long	0	0
Ward	15	Too long	0	0

Table 10: Comparison of the three algorithms in for twenty cluster lemma testing

Type of algorithm	Number of cluster lemma	Seconds	Correctly cluster lemma	Incorrectly cluster lemma	No class
Single-link	22	4.06	17	2	3
Complete-link	22	Too long	0	0	0
Ward	22	Too long	0	0	0

Table 11 above depicted the two algorithms did not respond the results but the single-link of hierarchical clusters responded sixteen lemma cluster correctly clustered and two were not correctly clustered 2 were no class for twenty cluster tests.

Based on the table 10, the hierarchical clustering of single-link algorithm's that derived confusion matrix for lemmatization clustered classes. This indicated that the hierarchical clustering lemmatization for Afan Oromo is better solution than stemmer, which is consistent with our testing clusterer classes.

Table 11: The confusion matrix of hierarchical clustering of lemmatization

	Correct cluster	**Incorrect cluster**
Correct cluster	TP(16)	FP(2)
Incorrect cluster	FN(2)	TN()

Table 11 was the summary of table 10 of clusterer output that was derived this information(confusion matric) to interpreted the clustering is to view it as a series of decisions, one for each of the pairs of lemma cluster in the dataset[37]. A true-positive decision assigns two similar documents to the same cluster; a true-negative decision assigns two dissimilar documents to different clusters. There are two types of errors we can commit. A false-positive decision assigns two dissimilar documents to the same cluster. A false-negative decision assigns two similar documents to different clusters. The accuracy measures the percentage of decisions that are correct [35].

4.4.2 Discussion

The hierarchical clustering algorithms (complete-link and ward) with edit distance measures was good small numbers of clusters. But as the number of cluster increased from small to large clusters their computation time would be increased or not produced the results of the clusters for these two hierarchical clustering algorithms. On the other hand; the single-link algorithms and edit distance measures[39] are slightly better than the two algorithms in generating more coherent clusters, which means the clusters have higher accuracy scores than these two algorithms. However, the two solutions (precision, recall) have the same value, which is 0.88 based on the confusion matrix. The overall accuracy value for edit distance measure and hierarchical cluster has shown 98.3% and, the overall error rate values are close and sometimes with only 1.7 % difference. When we compere the results that were achieved from single link algorithm with the other (hybrid stemmer of Debela that was achieved the accuracy of 95.73 % and Al-shammari used k-means clustering lemmatization that achieved 70.8% accuracy for Nobel Arabic Lemmatization algorithms). The Edit distance is again proved to be an effective metric for modeling the similarity between text documents for hierarchical clustering in single-link algorithm. Then, hierarchical clustering algorithm of single-link was produced good performance for Afan Oromo text lemmatization.

5. Conclusion and Recommendation

5.1. Conclusion

To conclude, in this research the researcher compared the single-link with complete-link and ward algorithms. Finally, the result obtained from this comparison was as follow:

- For five testing clusters of lemma the complete-link and ward algorithm had the same result but the two were displayed different lemma.
- When the number of cluster of lemma increased the two algorithms did not responded properly.
- When the researcher tested or experimented the single-link with edit distance to find out the correctly clustered lemmas, the result obtained for twenty testing's.
- At the end, based on the above findings the researcher prepared his own confusion matrix table.
- As the result of this, the researcher found out
    - The precision and recall value was obtained 0.88 based on the confusion matric table 11.
- In addition, accuracy was calculated and the result obtained was different i.e. 98.3 because of the addition of true positive and true negative divided by the total data and multiply by hundred brought it.
- When we take the error rate was 1.7 the difference seen here: could be the algorithm itself, or may be the manually annotated data manipulation error
- Lastly, when we compare the values of our hierarchical clustering lemma performs better than stemmer of Debela and lemmatization of al shammari

5.2. Recommendation

- When the number of clusters of lemma increases the single-link algorithm does not give proper response. Therefore; the testing cluster for 25,30 and above cluster does not respond accordingly.
- It needs other standalone computers that have high capacity to run and test to all candidate lemma to minimize its computational time.
- All the word formation was not included in this study it needs further studies.
- We tested only twenty clusters of candidate lemma, but it also needs other algorithm which is best in lemmatization for Afan Oromo text.
- Our data covers small area, but the study needs to gather huge data to compare its accuracy whether it increased or decreased as data increased.

6. References

[1] S. Paul, M. Tandon, N. Joshi, and I. Mathur, “Design of a Rule Based Hindi Lemmatizer,” 2013, pp. 67–74.

[2] N. Abera, “Long Vowels in Afan Oromo: A generic Approach,” Master Thesis School of graduate studies, Addis Ababa Unversity, 1989.

[3] D. Tesfaye and E. Abebe, “Designing a Rule Based Stemmer for Afaan Oromo Text,” *Int. J. Comput. Linguist. IJCL*, vol. 1, no. 2, 2010.

[4] S. Paul, M. Tandon, N. Joshi, and I. Mathur, “Design of a Rule Based Hindi Lemmatizer,” 2013, pp. 67–74.

[5] Abera. Long vowels in Afan Oromo: A generic approach. Master’s thesis,School of graduate studies. Addis Ababa, Univers.” .

[6] B. Daniel, “afaan oromo-english cross-lingual information retrieval (clir): a corpus based approach,” aau, 2011.

[7] D. Tesfaye, “Designing a Stemmer for Afaan Oromo Text,” AAU, 2010.

[8] statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf

[9] “weka manual, “‘http://www.cs.waikato.ac.nz/ml/weka.”

[10]“iasri.res.in/ebook/win_school_aa/notes/WEKA.pdf".

[11] “ Gumii Qormaata Afaan Oromoo;Caasluga Afaan Oromoo Jildii 1,Komishinii Aadaa fi Turizmii Oromiyaa,finfinnee,1995.” .

[12] “Tilahun;QubeAfaan Oromo: Reasons for Choosing the Latin Script for Developing an Oromo Alphabet,1993.”

[13] “Kula K. el at;Evaluation of Oromo-English Cross-Language Information Retrieval,Hyderabad,India,2008.”

[14] “Ayana Danie;Afaan Oromo-English Cross-Lingual Information Retrieval(CLIR): A corpus based approach,Master Thesis, AAU,2011.”

[15] “Christopher D. Manning et al;Introduction to Information Retrieval,cambridge university press,New York,NY 10013-2473,” .

[16] “Plisson et al;A Rule based approach to word lemmatization,Institute Jozef Stefan,Ljubljana,2008.”

[17] “http://www.christianlehmann.eu/ling/ling_meth/ling_description/lexicography/index.html.” .

[18] D. Suhartono, “Lemmatization Technique in Bahasa: Indonesian,” *J. Softw.*, vol. 9, no. 5, p. 1203, 2014.

[2o[21]“Jakub K. et a;Comparison of Different Lemmatization Approaches through the Means of Information Retrieval Performance,2010.”

[22] “Oromia culture and tourism Bureau;Afan Oromo-Amharic-English Dictionary,Oromia culture and tourism Baorse,Finfinnee,2006 E.C or 2014 G.C.” .

[23] “Okan Ozturkmenoglu, Adil Alpkocak,Comparison of different lemmatization approaches for information retrieval on Turkish text collection,Innovations in Intelligent Systems and Applications (INISTA),IEEE, 2012.”

[24] “[Manning_Christopher_D.,_Raghavan_Prabhakar,_Schü(BookZZ.org).pdf.” .

[25] “Jakub K. et al;Comparison of Different Lemmatization Approaches through the Means of Information Retrieval Performance,.”

[26] “Bellet, A., Habrard, A., and Sebban, M. (2013). A Survey on Metric Learning for Feature Vectors and Structured Data.Technical report, arXiv:1306.6709,” .
[27] S.-S. Kang, “Word Similarity Calculation by Using the Edit Distance Metrics with Consonant Normalization.,” *J. Inf. Process. Syst.*, vol. 11, no. 4, 2015.
[28] E. Al-Shammari and J. Lin, “A novel Arabic lemmatization algorithm,” 2008, pp. 113–118.
[29] T. El-Shishtawy and F. El-Ghannam, “An accurate arabic root-based lemmatizer for information retrieval purposes,” *ArXiv Prepr. ArXiv12033584*, 2012.
[30] E. Al-Shammari and J. Lin, “A novel Arabic lemmatization algorithm,” in *Proceedings of the second workshop on Analytics for noisy unstructured text data*, 2008, pp. 113–118.
[31] J. Kanis and L. Müller, “Automatic Lemmatizer Construction with Focus on OOV Words Lemmatization,” in *Text, Speech and Dialogue*, vol. 3658, V. Matoušek, P. Mautner, and T. Pavelka, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 132–139.
[32] “Jongejan, B.et al;constructed trainable lemmatizers for the lexicographical,Suntec,Singapore,2009.”
[33] G. D. Dinegde and M. Y. Tachbelie, “Afan Oromo News Text Summarizer,” *Int. J. Comput. Appl.*, vol. 103, no. 4, 2014.
[34] “Mewis;A Grammatical Sketch of Written Oromo (Language and dialect atlas of Kenya,2001.”
[35] “IRbook ccu press 2008 and edit online in2009,” .
[36] “http://afaan-oromo-German-dictionary.soft112.com/.” .
[37] “Manning, Christopher D.Introduction to information retrieval/ Christopher D. Manning, Prabhakar Raghavan, Hinrich Sch ¨ utze.irbook, Cambridge University Press 2008 16.4, page 364). Or Online edition (c)2009 Cambridge UPcf. Section 16.4, page 364).,” .
[38] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, Reprinted. Cambridge: Cambridge Univ. Press, 2009.
[39] “Edit-Distance of Weighted Automata General Definitions and.pdf.” .
[40] “Clustering of Categorized Text Data Using.pdf.” .
[41] K. Black, E. K. Ringger, P. Felt, K. D. Seppi, K. Heal, and D. Lonsdale, “Evaluating Lemmatization Models for Machine-Assisted Corpus-Dictionary Linkage.,” in *LREC*, 2014, pp. 3798–3805.

Appendix I Afan Oromo words,prefix,suffix and lemma

id	words	prefix	suffix	Lemma
1	abbaadhumma	N/a	dhumma	abbaa
2	abbabbaa	N/a	abbaa	abbaa
3	abbbicha	N/a	icha	abbaa
4	abbichumma	N/a	ichumma	abbaa
5	abboomaa	N/a	oomaa	abbaa
6	abboomamaa	N/a	oomamaa	abbaa
7	abboomamuu	N/a	oomamuu	abbaa
8	abboommachuu	N/a	oommachuu	abbaa
9	abboommii	N/a	oommii	abbaa
10	abboomsisuu	N/a	oomsisuu	abbaa
11	abboomuu	N/a	oomuu	abbaa
12	abbootii	N/a	ootii	abbaa
13	Abbummaa	N/a	ummaa	abbaa
14	abbaanaa	N/a	naa	abbaa
15	abdachiiftuu	N/a	achiiftuu	abdii
16	abdachiisaa	N/a	achiisaa	abdii
17	abdachiisuu	N/a	achiisuu	abdii
18	abdachuu	N/a	achuu	abdii
19	abdansa	N/a	ansa	abdii
20	abdataa	N/a	ataa	abdii
21	Abdate	N/a	ate	abdii
22	Adaamaarra	N/a	irra	adaamaa
23	Adaamaatti	N/a	itti	adaamaa
24	adaammarraa	N/a	irraa	adaamaa
25	adduyyaalee	N/a	lee	addunyaa
26	adduyyaalessumma	N/a	lesummaa	addunyaa
27	adduyyaalessuu	N/a	lessuu	addunyaa
28	adduyyalessa	N/a	lessa	addunyaa
29	amanuu	N/a	uu	amana
30	amanama	N/a	ma	amana
31	Amanama	N/a	ama	amana
32	amanatuu	N/a	tuu	amana
33	Amanamummaa	N/a	ummaa	amana

34	Amanamummaadhaan	N/a	ummaadhaan	amana
35	ammayyeesse	N/a	eesse	ammayyaa
36	Ammayyoomsuu	N/a	ooyyoomsuu	ammayyaa
37	ammayyummaa	N/a	ummaa	ammayyaa
38	Anga?oota	N/a	oota	angoo
39	anga'aa	N/a	a'aa	angoo
40	arge	N/a	e	argaa
41	arguu	N/a	uu	argaa
42	arsichaa	N/a	ichaa	arsii
43	arsirraa	N/a	irraa	arsii
44	arsittii	N/a	ittii	arsii
45	arsirratti	N/a	irratti	arsii
46	baqata	N/a	ta	baqa
47	baqataa	N/a	taa	baqa
48	baqatii	N/a	atii	baqa
49	baqatuu	N/a	tuu	baqa
50	baqachiisuu	N/a	achiisuu	baqa
51	baqachuu	N/a	achuu	baqa
52	baqate	N/a	ate	baqa
53	barataan	N/a	aan	barataa
54	barattoota	N/a	oota	barataa
55	barbaadamuu	N/a	muu	barbaada
56	barbaadachuu	N/a	chuub	barbaada
57	barbaaduu	N/a	uu	barbaada
58	barbaacha	N/a	cha	barbaada
59	bareechaa	N/a	cha	bareeda
60	bareechisuu	N/a	chisuu	bareeda
61	bareechuu	N/a	chuu	bareeda
62	bareedina	N/a	ina	bareeda
63	bareedaa	N/a	a	bareeda
64	bareedduu	N/a	uu	bareeda
65	bareedum	N/a	uma	bareeda
66	bareeduu	N/a	uu	bareeda
67	bareeffachuu	N/a	ffachuu	bareeda
68	babbareeduu	bab	uu	bareedaa
69	babbareedan	bab	n	bareedaa
70	babbareeda	bab	N/a	bareedaa
71	babarfannaa	ba	annaa	barfata

72	babarfataa	ba	a	barfata
73	babarfachiisuu	ba	achiisuu	barfata
74	babarfachuu	ba	achuu	barfata
75	barsiisonni	N/a	onni	barsiisa
76	barsiisota	N/a	ota	barsiisa
77	beekumsa	N/a	umsa	beekaa
78	beekumsa	N/a	umsa	beektu
79	Beerran	N/a	ran	Beera
80	bilisoomte	N/a	oomte	Bilisa
81	Bilisummaa	N/a	ummaa	Bilisa
82	bilisummaadhaaf	N/a	ummaadhaaf	bilisa
83	bilisummaaf	N/a	ummaaf	bilisa
84	Birmadeessaa	N/a	eessaa	Birmada
85	Birmadummaa	N/a	ummaa	Birmada
86	Bulchan	N/a	an	Bulchaa
87	Bulchiinsa	N/a	iinsa	Bulchaa
88	Bulchoota	N/a	oota	Bulchaa
89	Bulchuu	N/a	uu	Bulchaa
90	caalaadhaaf	N/a	dhaaf	caalaa
91	caalaadhaan	N/a	dhaan	caalaa
92	caalaarra	N/a	irra	caalaa
93	caalchisan	N/a	chisan	caalaa
94	caaliinsa	N/a	iinsa	caalaa
95	caalaatti	N/a	tti	caalaa
96	caccaba	cac	N/a	caba
97	caccaban	cac	an	caba
98	caccabe	cac	N/a	caba
99	caccabsan	cac	san	caba
100	caccabse	cac	se	caba
101	caccabsiis	cac	siise	caba
102	caccabsuu	cac	suu	caba
103	ciccimaa	cic	N/a	cimaa
104	ciccimsiisuu	cic	siisuu	cimaa
105	ciccimoo	cic	oo	cimaa
106	ciccimaniiru	cic	aniiru	cimaa
107	ciccitiinsa	cic	iinsa	cita
108	ciccituu	cic	uu	cita
109	ciccite	cic	e	cita

110	ciccita	cic	N/a	cita
111	da'umsa	N/a	umsa	da'a
112	daa'imman	N/a	man	daa'ima
113	dabeessa	N/a	eessa	daba
114	dabiinsa	N/a	iinsa	daba
115	dabina	N/a	ina	daba
116	dabsiisuu	N/a	siisuu	daba
117	dabuu	N/a	uu	daba
118	dabummaa	N/a	ummaa	daba
119	dangaalee	N/a	lee	dangaa
120	dardarummaa	N/a	ummaa	dardara
121	dhabaa	N/a	a	dbaba
122	dhaalaa	N/a	a	dhaala
123	dhaalchisuu	N/a	chisuu	dhaala
124	dhaaliinsa	N/a	iinsa	dhaala
125	dhaaltuu	N/a	tuu	dhaala
126	dhaaluu	N/a	uu	dhaala
127	dhaabachiisuu	N/a	achiisuu	dhaba
128	dhaabachuu	N/a	achuu	dhaba
129	dhaabbii	N/a	ii	dhaba
130	dhaabsisuu	N/a	siisuu	dhaba
131	dhaabuu	N/a	uu	dhaba
132	dhabamsiisuu	N/a	msiisuu	dhaba
133	dhabamuu	N/a	muu	dhaba
134	dhabbata	N/a	ata	dhaba
135	dhabbataa	N/a	ataa	dhaba
136	dhabduu	N/a	duu	dhaba
137	dhabiinsa	N/a	iinsa	dhaba
138	dhabsiisuu	N/a	siisuu	dhaba
139	dhabsuu	N/a	suu	dhaba
140	dhabuu	N/a	uu	dhaba
141	dhalaa	N/a	a	dhala
142	dhalachuu	N/a	achuu	dhala
143	dhalata	N/a	ta	dhala
144	dhaloota	N/a	oota	dhala
145	dhaltii	N/a	tii	dhala
146	dhaltoolee	N/a	toolee	dhala
147	dhaluu	N/a	uu	dhala

148	dhedheera	dhe	N/a	dheera
149	dhedheerachuu	dhe	chuu	dheera
150	dhedheerachuu	dhe	achuu	dheera
151	dhedheeressuu	dhe	essuu	dheera
152	dhedheerina	dhe	ina	dheera
153	dhedheero	dhe	o	dheera
154	dhedheeroo	dhe	oo	dheera
155	dheerachuu	N/a	chuu	dheera
156	dheeressuu	N/a	essuu	dheera
157	dheerina	N/a	ina	dheera
158	dheertoo	N/a	too	dheera
159	dhedheertu	dhe	N/a	dheertu
160	dhedheertuu	dhe	N/a	dheertuu
161	dhiiromuu	N/a	omuu	dhiira
162	dhiirti	N/a	ti	dhiira
163	dhiirummaa	N/a	ummaa	dhiira
164	dhiirummaa	N/a	ummaa	dhiira
165	dhimmamaan	N/a	maan	dhimma
166	dhimmamtoota	N/a	toota	dhimma
167	dhiphachuu	N/a	achuu	dhiphaa
168	dhiphachuu	N/a	achuu	dhiphaa
169	dhiphata	N/a	ata	dhiphaa
170	dhiphataa	N/a	taa	dhiphaa
171	dhiphattuu	N/a	attuu	dhiphaa
172	dhiphattuu	N/a	ttuu	dhiphaa
173	dhiphina	N/a	ina	dhiphaa
174	dhiphina	N/a	ina	dhiphaa
175	dhiphisiisuu	N/a	isiisuu	dhiphaa
176	dhiphisuu	N/a	isuu	dhiphaa
177	dhiphoo	N/a	oo	dhiphaa
178	dhiphoo	N/a	oo	dhiphaa
179	dhiphuu	N/a	uu	dhiphaa
180	dhoksuu	N/a	uu	dhiphaa
181	dhokachoo	N/a	achoo	dhoksaa
182	dhokachuu	N/a	achuu	dhoksaa
183	dhokfachuu	N/a	achuu	dhoksaa
184	dhokfamuu	N/a	amuu	dhoksaa
185	dhoksachuu	N/a	achuu	dhoksaa

186	dhoksamuu	N/a	muu	dhoksaa
187	dhoksiisuu	N/a	iisuu	dhoksaa
188	dhufaatii	N/a	atii	dhufa
189	dhufiinsa	N/a	iinsa	dhufa
190	dhufuu	N/a	uu	dhufa
191	dhugaadhumattii	N/a	dhumattii	dhugaa
192	dhugaati	N/a	ti	dhugaa
193	dhugaatti	N/a	ttii	dhugaa
194	dhugaattiyyuu	N/a	ttiyyuu	dhugaa
195	dhuguma	N/a	uma	dhugaa
196	dhugumaaf	N/a	umaaf	dhugaa
197	dhuunfaaf	N/a	af	dhuunfa
198	dhuunfaan	N/a	an	dhuunfa
199	dhuunfarraa	N/a	rraa	dhuunfa
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
984	xumuruu	N/a	uu	xumura
985	yaadachiisa	N/a	achiisa	yaada
986	yaadachiisaa	N/a	achiisaa	yaada
987	yaadachiisuu	N/a	achiisuu	yaada
988	yaadannoo	N/a	nnoo	yaada
989	yaada'uu	N/a	a'uu	yaada
990	yaaddessaa	N/a	essaa	yaada
991	yaaddessuu	N/a	essuu	yaada
992	yaaddoo	N/a	oo	yaada
993	yaaduu	N/a	uu	yaada
994	yakkamaa	N/a	amaa	yakka
995	yakkamtoota	N/a	amtoota	yakka
996	yakkamtuu	N/a	amtuu	yakka
997	yakkamuu	N/a	amuu	yakka
998	yakkamaniiru	N/a	maniiru	yakka
999	yakkamtan	N/a	tan	yakka
1000	yakkamtoonni	N/a	toonni	yakka