ADDIS ABABA UNIVERSITY

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach

BY

Debela Tesfaye

A THESIS SUBMITED TO THE SCHOOL OF GRADUTE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

ADDIS ABABA, ETHIOPIA

June, 2010

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUTE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

Designing a Stemmer for Afaan Oromo Text: A Hybrid Approach

BY

Debela Tesfaye

Signature of the Board of Examiners for Approval

	Name	Signature
1.	_____	_____
2.	_____	_____
3.	_____	_____
4.	_____	_____

DECLARATION

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

_____

Debela Tesfaye

July, 2010

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

_____

Ermias Abebe

July, 2010

Acknowledgment

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

My deepest gratitude goes to my advisor Ato Ermias Abebe for his critical comments on my work.

I would also thank Ato Amanuel Raga, Hika Fekede and Derartu Fikadu (my wife) who has a great input in my study in providing relevant resources and helping me in analyzing Afan Oromo Morphology and grammer.

I am also grateful to my colleague Ato Getachew Mamo who have supported me in moral and providing me good working environment and materials.

Dedication

This work is dedicated to my son, Kenna Debela.

Table of contents

Symbols

( ) What is inside is the meaning of Afan Oromo words in English except those used to indicate cite (source) of documents used and stemmed word.

‘’ what is inside is the suffixes, prepositions and conjuctions in English

Abbreviations and Definitions

C	Consonant
V	Vowel
R1	is the region after the first non-vowel following a vowel. If the word starts in vowel it is the region before the next consonant.
R2	The region after the first non-vowel following a vowel in R1.
M	Measure, (Measure is the number vowel consonant sequence appearing in a word)
NLP	Natural Language Processing
IR	Information Retrival
Sg. 1.p.	1st person singular
Sg. 2.p.	2nd person singular
Sg. 3.p.m.	3rd person singular masculine
Sg. 3.p. f.	3rd person singular feminine
pl. 1.p.	1st person plural
pl. 2.p.	2nd person plural
pl. 3.p.	3rdperson plural

List of Tables

Afan Oromo Consonants

		Bilabial/ Labiodental	Alveolar/ Retroflex	Palato-alveolar/ Palatal	Velar/Glottal			
Stops	Voiceless	(p)	t	k	'			
	Voiced	b	d	g				
	Ejective	ph	x	q				
	Implosive	dh						
Affricates	Voiceless	ch						
	Voiced	j						
	Ejective	c						
Fricatives	Voiceless	f	s	sh	h			
	Voiced	(v)	-	Nasals		m	n	ny
Approximants		w	l	y				
Flap/Trill		R						

Afan Oromo Vowels

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

Abstract

Most natural language processing systems use stemmer as a separate module in their architecture. Specially, it is very significant for developing, machine translator, speech recognizer and search engines. In linguistic morphology, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form.

In this thesis work, a stemming system for Afan Oromo is presented. This system takes as input a word and removes its affixes according to a rule based algorithm. This stemmer is not enough to define every rule applied in Afan Oromo word formation. Therefore, N-gram is integrated with the rule to handle cases that are not covered by rule in the hybrid version of this stemmer. The algorithm follows the known Porter algorithm for the English language and it is developed according to the grammatical rules of the Afan Oromo, as they are described in a Grammatical sketch of Written Oromo (Mewis, 2001) and Caasluga Afaan Oromoo, Jildii-1 (Oromo, 1995). Afan Oromo morphology was studied and described in order to model the language and develop an automatic procedure for conflation. The inflectional and derivational morphologies of the language are discussed.

The result of the study is a prototype context sensitive iterative stemmer for Afan Oromo. Error counting technique was employed to evaluate the performance of this stemmer. For testing purpose 198 sentences (with a total of 2458 words) is collected from different public Afaan Oromo newspapers and bulletins to make the testing set address variety of issues. An evaluation of the system shows that the algorithms accuracy works with better performance than other past stemming algorithms for Afan Oromo giving 95.73 percent correct results. Finally, possible extensions of the proposed system and further evaluation methods are briefly reviewed.

Chapter one: Introduction

1.1 Background of the Research

Nowadays many tools are provided for information retrieval. There are some interesting categories of the available information retrieval software. We can find a variety of internet search engines with advanced search parameters, specialized search engines for retrieving documents in a document collection, data mining and clustering tools as well as other classification tools. During their development we can notice an ongoing specialization on the searching features. These engines are becoming more and more sophisticated trying to cover user's demands to access specific information.

One of the attempts to make the search engines more effective in information retrieval was the usage of word stemming. A stemming algorithm is a procedure that reduces all words with the same stem to a common form by stripping of its derivational and inflectional suffixes (Lovins, 1968). The main objective of the stemming process is to remove all possible affixes and thus reduce the word to its stem (Dawson 1974). Using stemming, many contemporary search engines associate words with prefixes and suffixes to their word stem, to make the search broader in the meaning that it can ensure that the greatest number of relevant matches is included in search results.

Stemming improves IR performance generally by bringing variant forms of a word which share a common meaning under one heading. Significant, although sometimes small, improvements on retrieval systems across a range of test collection is discovered

(Krovetz, 1995). What Krovetz (1995) discovered is that the degree of improvement varies considerably between different collections. These tests were however done on collections in English, and the reasonable assumption of Information Retrival researchers has always been that for languages that are more highly inflected than English (and nearly all are), greater improvements will be observed when stemming is applied. Stemmers are common elements in query systems such as Web search engines, since a user who runs a query on *barataa* which means student for example would probably also be interested in documents that contain the word *barattoota*(students). Searching for "fish" would not have returned "fishing" without stemming.

In addition, stemming helps regularize the vocabulary of an IR system, and this leads to advantages that are not easily quantifiable through standard IR experiments. For example, it helps in presenting lists of terms associated with the query back to the IR user in a relevance feedback cycle, which is one of the underlying ideas of the probabilistic model (Porter, 2001). If a list of stems is to be presented back for query expansion, in place of a stem, the user should be shown a single representative from the set *words*, the one of highest frequency perhaps. The user should also be able to choose for the whole query, or at a lower level for each word in a query (Lovins, 1968).

As explained in the above three paragraphs, the process of stemming, often called conflation, is useful in search engines for query expansion or indexing and other natural language processing problems like part of speech tagging, speech recognition systems,

and word processors (Al-Attram, 1990). Stemming has also applications in machine translation, document summarization (Ntais, 2006).

Stemming can greatly decrease storage space by representing every morphological variations of a word in single term.

The stem produced by a stemmer need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

There are several techniques used for word stemming, developed through time. They are: Dictionary-Based approach, production technique, Suffix stripping, statistical approach (Ntais, 2006).

Dictionary-Based approach

It employs a lookup table which contains relations between root forms and inflected forms. To stem a word, the table is queried to find a matching inflection. If a matching inflection is found, the associated root form is returned. It is unrealistic to expect that all word forms can be captured and manually recorded by human action alone given the number of words in a given language. Manual training of the algorithm is overly time-intensive and the ratio between the effort and the increase in accuracy is marginal at best (Fairweather, 2000).

Production technique algorithm

Production algorithm generates the possible list of inflected forms of a given word (Lovins, 1968). One drawback of the production technique is that there is no guarantee the inflection is real. For example, the technique might join together "run" and "ly" to create the word "runly".

Suffix stripping algorithms

This is the most widely applied stemming technique. One of the suffix stripping algorithim is the algorithm introduced by Porter (1980). With specific rules for the English language, this algorithm removes suffixes iteratively from a given word, reducing it to its stem. Even it has its limitations, it is the most commonly accepted algorithim for its high precision and recall. Lovin's (1968) stemmer follows the same rule-based technique but it does not apply its rules iteratively and it is more conservative than Porter's algorithm (Lovin's, 1968).

In Suffix stripping algorithms list of rules is stored which provide a path for the algorithm, given an input word form, to find its root form. Suffix stripping algorithms are sometimes regarded as crude given the poor performance when dealing with exceptional relations (like 'ran' and 'run'). This, however, is a problem, as not all parts of speech have such a well formulated set of rules. Lemmatization attempts to improve upon this challenge (Lovins 1968).

Lemmatization process involves first determining the part of speech of a word, and applying different normalization rules for each part of speech. The part of speech is first

detected prior to attempting to find the root since for some languages, the stemming rules change depending on a word's part of speech. This approach is highly conditional upon obtaining the correct lexical category (part of speech).

Statistical Approach

Statistical stemming algorithms involve using probability to identify the root form of a word. Statistical stemming algorithms are trained on a table of root form to inflected form relations to develop a probabilistic model.Some of the methodologies are: frequency counts and n-gram (Mayfield and McNamee, 2003). This approach does not require any linguistic knowledge whatsoever, being totally independent of the morphological structure of the target language.

Hybrid Approaches

Hybrid approaches use two or more of the approaches described above in unison. A simple example is a suffix tree algorithm which first consults a lookup table using brute force algorithim. However, instead of trying to store the entire set of relations between words in a given language, the lookup table is kept small and is only used to store a minute amount of "frequent exceptions" like "ran => run". If the word is not in the exception list, apply suffix stripping or lemmatization and output the result (Popovic & Willett, 1992).

1.2 Statement of the Problem

Afaan Oromo is one of the major languages that is widely spoken and used in Ethiopia (Abara, 1988). Currently it is an official language of Oromia state (which is the largest region in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which accounts to 34.5% of the total population (Census, 2008).

The language has become the official language in the Oromia regional state of Ethiopia and is also instructional language starting from elementary to university level. As a result, text books, references and other governmental documents are compiled using the language.

Today as technology improves, several forms of information are in use every where in the World. Books, journals, articles and other documents can be accessed electronically. For some time now, the issue has become one of storing and accessing this pervasire information in an effective and efficient manner. Information retrival is a field which tries to address these issues.

For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. Stemming

makes families of derivationally and inflectionally related words with similar meanings represented using single term.

Several electronic documents printed in Afaan Oromo are available on the web. Therefore, information retrieval systems that process Afaan Oromo documents need a stemmer for indexing and query expanssion. Afan Oromo word processing softwares may also require stemmer used for spell checking. Other natural language processing systems like machine translation, speech recognition also require stemmer as a single component.

So far few researches are conducted to design a stemmer for the language. To mention, a few, Wakshum (2000) attempted the Development of stemming algorithm for Afaan Oromo language text followed by Kekeba, Varma and Pingali (2007) who also aim to design stemming algorithm for Afaan Oromo language Text.

Wakshum (2000) reported 92 percent accuracy for his stemmer. A semi automatic construction of suffix table is used in combination with rules that strips off suffix from a given word by looking up the longest match suffix in the list of suffix (Wakshum, 2000).To stem a word; the table is queried to find a matching suffix. If a matching is found, the affix is removed and the associated root form is returned. The algorithm is only accurate to the extent that the suffix form already exists in the table.

Wakshum (2000) suggested other stemming algorithms to be tried for Afaan Oromo. Irregular formations of variants from root word, morphems that are formed by

duplication of some characters are not covered by the stemmer. The challenge to increase the number and complexity of the rules is also another problem faced by the stemmer. The rules Wakshum (2000) has designed have no linguistic back ground but very few context sensitive rules are added in the modified version of the stemmer. This necessitates the need for the detailed analysis of the language to use such frameworks to generate a totally context sensitive stemmer.

The stemmer developed by Kekeba, Varma and Pingali (2007) used a rule based suffix-stripping algorithms focusing on very common inflectional suffixes of Afan Oromo. This light stemmer is designed to automatically remove frequent inflectional suffixes attached to headwords (base-word forms) of Afaan Oromo.The stemmer didn`t considered words which are formed by duplication of some characters, to indicates repeatition and plural form, at all. In addition, the stemmer is not context sensitive.

Another alternative to the development of stemmers, which is not tried for Afan Oromo, is the use of statistical approaches.

This research is aimed at developing stemmer that uses both statistical and rule based approaches. This is new approach for the language and utilizes the benefit of the two approaches to increase the accuracy level. The number and complexity of rules can be reduced by using statistics. There fore statistics is used as a complement to the rule based stemmer to conflate Afaan Oromo text.

1.3 Objective of the Research

Details of the general and specific objectives of the research work are the following: -

1.3.1 General Objective

The general objective of this research is to develop hybrid (rule based and n-gram) stemmer for Afaan Oromo text.

1.3.2 Specific Objectives

In order to achieve the general objective, the researcher has the following specific objectives

- Preparing corpus for Afaan Oromo stemmer.
- Develop rule clusters applicable in Afan Oromo word formation.
- Develop N-gram stemmer prototype for Afaan Oromo text.
- Develop rule based stemmer prototype for Afaan Oromo text.
- Integrate the rule based stemmer with n-gram stemmer.
- Evaluate the performance of the stemmer.
- To draw conclusions based on experimental result and recommendation to further research areas.

1.4 Scope

There are different ways in developing stemmer. Namely, dictionary method, rule based and statistical approach. This research is aimed at designing hybrid stemmer for Afaan Oromo text. Two methods are integrated together to develop the stemmer. The two methods are n-gram and rule based approach. The approach integrates both rules and statistics in designing the stemmer. N-gram stemmer is used as a complement to the rule based stemmer. Part which is not handled by rule is covered by statistics. Afan Oromo compound words and *hin*, *ni* when they occour as prefix are not considered by the rule based stemmer.

1.5 Methodology

1.5.1 Literature Review

Research conducted on stemming for Afaan Oromo and other languages are reviewed to adopt some stemming concepts and tackle some obstacles faced by researchers.

Other literatures regarding Afan Oromo morphology and grammar are also reviewed to understand the morphological distribution of Afaan Oromo words and develop rule clusters applicable in Afan Oromo word formation.

1.5.2 Data Preparation

For this particular study, corpus was collected from different popular Afaan Oromo newspapers (Bariisaa, Bakkalcha Oromiyaa and Oromiyaa) and bulletins (Qabee and Oromiyaa) to make the corpus variety. Newspapers, bulletins and public magazines are

considered as consisting different issues of the community: social, economical, technological and political issues. They are a potential source for collecting corpus, which is not biased to specific issue, for natural language processing tasks.

Professionals educated in Afaan Oromo are also consulted to prepare standardized morphological distribution of the Language. In addition, literatures regarding Afaan Oromo morphological structure are also reviewed to understand and prepare the document.

1.5.3 Implementation Procedure

Java programming language is selected to test Afaan Oromo stemmer. The reason is that java programming language has a facility to deal with natural language text processing as compared to programming languages like C++.

1.5.4 Testing Procedure

Error counting technique was employed to evaluate the performance of this stemmer. Quantitative analysis is used to see the result of the stemmers. The result is represented in quantitative measures like percentage of correctly stemmed words, the error rate etc. The figure obtained is used to evaluate the accuracy of the stemmer.

1.7 Significance of the Research

Afan Oromo has become the official language in Oromiya region offices and is also instructional language in schools and collages. As the result text books, references and other governmental files are compiled using the language. So during writing the

documents Afaan Oromo text editors can use the stemmer in order to correct spelling errors.

Apart from these, journals, news papers, articles, books printed in Afaan Oromo are available on the web. So Information retrieval system that process Afaan Oromo documents can also use it for indexing.

Further studies related to Afaan Oromo text processing can use it as input. Translation, part of speech tagging, speech recognition system in Afaan Oromo can use the stemmer.

In an IR system with queries and index stemmed, the user needed no special knowledge of the form of the subject terms to expand the query (Lovins, 1968). The user should not have to see the stemmed form of a word. If a list of stems is to be presented back for query expansion, in place of a stem, the user should be shown a single representative from the set *words*, the one of highest frequency perhaps. The user should also be able to choose for the whole query, or at a lower level for each word in a query. Query expansion with stemming results in a much cleaner vocabulary list than without, and this is a main strength of using a stemming process.

Chapter Two: Literature Review

2.1 Introduction

The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. With the rise of the World Wide Web, the diversity of topics has increased as has the applicability of information retrieval to a number of fields. Despite so much success, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tedious and difficult task. For instance, to satisfy his information need, the user might navigate the space of Web links (i.e., the hyperspace) searching for information of interest. However, since the hyperspace is vast and almost unknown, such a navigation task is usually inefficient. The main obstacle is the absence of a well defined underlying data model for the Web, which implies that information definition and structure is frequently of low quality (Yates and Neto, 1999).

Information retrieval almost exclusively referres to indexing texts and searching for documents (Yates and Neto, 1999). Information Retrieval (IR) can be defined broadly as the study of how to determine and retrieve from a corpus of stored information the portions of which are relevant to particular information needs(Lovins, 1968).

For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and

democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. In information retrieval, the abundance of different word forms and lexical variability may result in a greater likelihood of mismatch between the forms of a keyword in a query and its variant forms found in the document index database(s). One of the attempts to make the search engines more effective in information retrieval was the usage of word stemming.

A stemming algorithm is a procedure that reduces all words with the same stem to a common form, usually by stripping each word of its derivational and inflectional suffixes (Lovins, 1968). Before getting in the details of stemming let us explain the difference between 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 (Manning et. al 2009).

For instance:

am, are, is ⇒be

car, cars, car's, cars' ⇒car

The result of this mapping of text will be something like:

The boy's cars are different colors ⇒ the boy car be differ color

However, Stemming and lemmatization differ in their flavor. *S*temming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma (Manning et. al 2009). If confronted with the token ‘saw’, stemming might return just ‘s’, whereas lemmatization would attempt to return either ‘see’ or ‘saw’ depending on whether the use of the token was as a verb or a noun. Stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma.

Even though particular domains may require special stemming rules, the exact stemmed form does not matter, only the equivalence classes it forms. Rather than using a stemmer, we can use a lemmatizer, a tool from Natural Language Processing which does full morphological analysis to accurately identify the lemma for each word. Doing full morphological analysis produces at most very modest benefits for retrieval.

For an IR system stems are usually sufficient, for a morphological analysis system obviously lemmas are a must. In practice, various stemmers produce a mix of stems and lemmas (Manning et. al 2009). Stemming offers a simpler alternative to lemmatization. Stemming also attempts to reduce a word to a base form by removing affixes, but the resulting stem is not necessarily a proper lemma. Such stems can be useful in information retrieval applications.

The main objective of the stemming process is to remove all possible affixes and thus reduce the word to its stem (Dawson, 1974). Using Stemming, many contemporary search engines associate words with prefixes and suffixes to their word stem, to make the search broader in the meaning that it can ensure that the greatest number of relevant matches is included in search results.

Like a number of other African and Ethiopian languages, Afaan Oromo has a very rich morphology (Oromoo, 1995). It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afaan Oromo, Amharic and Zulu most of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems. Both Afaan Oromo nouns and adjectives are highly inflected for number and gender. In comparison to the English plural marker s (-es), there are more than 12 major and very common plural markers in Afaan Oromo nouns (example: *-oota, -ooli, -wwan, -lee, -an, -een, -oo,* etc.) (Oromo, 1995). Afaan Oromo verbs are also highly inflected for gender, person, number and tenses. Moreover, possessions, cases and article markers are often indicated through affixes in Afaan Oromo. Since Afaan Oromo is morphologically very productive, derivations and word formations in the language involve a number of different linguistic features including affixation, reduplication and compounding (Oromoo, 1995). Obviously, these high inflectional forms and extensive derivational features of the language are presenting various challenges for text processing and information retrieval experiments in Afaan Oromo.

Applications of certain level of morphological (linguistic) analysis and natural language processing tools are often assumed to be very essential in IR experiments of morphologically rich and agglutinative languages like Afaan Oromo. A number of previous research works, have indicated the fact that CLIR applications in morphologically rich languages can benefit from stemming and lemmatization of query terms (Kekeba et al 2007).

An evaluation research about stemming techniques and how these affect the search precision, demonstrate that stemming raises the effectiveness of information retrieval (Lennon, 1981). Stemming can affect retrieval performance, but the studies are equivocal. There is no evidence that stemming will degrade retrieval effectiveness (Frakes, 2010). Stemming has also applications in machine translation, document summarization and text classification (Ntais, 2006).

2.2 Types of stemming

There are three common approaches that are used in stemming: rule based, dictionary based and statistical methods (Bento et al., 2005).

2.2.1 Dictionary-Based Technique

Dictionary-based stemmers match every word with a word on a proper digitalized dictionary; correspond each word to its stem (Wikipedia, 2010). Direct dictionary method seems effective but inadequate to deal with the “unlimited” words and their formation, especially in inflected languages with elevated morphological structure Moreover

“dictionary-based stemmers require dictionary maintenance, to keep up with an ever-changing language, and this is actually quite a problem. It is not only that a dictionary created to assist stemming now adays will probably require major updating in a few years time, but also that a dictionary in use for this purpose today may already be several years out of date” (Krovetz, 1995).

2.2.2 Rule-Based Technique

Most rule-based stemmers currently in use are iterative longest match stemmers (Lovins, 1968). An iterative longest match stemmer removes the longest possible string of characters from a word according to a set of rules. This process is repeated until no more characters can be removed. Even after all characters have been removed, stems may not be correctly conflated. The word "skies," for example, may have been reduced to the stem "ski" which will not match "sky." One of the rule based stemming techinique is the porter`s stemming algorithim.

2.2.2.1 Overview of Porter`s Stemming Algorithim

The porter algorithm is too long and intricate to present here, but we will indicate its general nature. The Stemmer is based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a combination of smaller and simpler suffixes. Specifically it has five steps and applying rules within each step. Within each step, if a suffix rule matched to a word, then the conditions attached to that rule are tested on what would be the resulting stem, if that suffix was removed, in the way defined by the rule. For example such a condition may be, the number of vowel characters, which

are followed be a consonant character in the stem (Measure), must be greater than one for the rule to be applied (Porter , 1980). Within each phase there are various conventions to select rules, such as selecting the rule from each rule group that applies to the longest suffix.

In the first phase, this convention is used with the following rule group:

Rule	**Example**
SSES → SS	caresses → caress
IES → I	ponies → poni
SS → SS	caress → caress
S →	cats → cat

Many of the rules use a concept of the ***measure*** of a word, which loosely checks the number of syllables to see whether a word is long enough that it is reasonable to regard the matching portion of a rule as a suffix rather than as part of the stem of a word. For example, the rule:

($m > 1$) EMENT → would map *replacement* to *replac*, but not *cement* to *c*.

Once a Rule passes its conditions and is accepted the rule fires and the suffix is removed and control moves to the next step. If the rule is not accepted then the next rule in the step is tested, until either a rule from that step fires and control passes to the next step or there are no more rules in that step whence control moves to the next step. This process

continues for all five steps, the resultant stem being returned by the stemmer after control has been passed from step five.

The following shortcomings have been found in the stemming algorithm. The algorithm clearly explains that when a set of rules of the type

(condition)S1 → S2

are presented together, only one rule is applied, the one with the longest matching suffix *S1* for the given word. This is true whether the rule succeeds or fails (i.e. whether or not *S2* replaces *1*). Despite this, the rules are sometimes simply applied in turn until either one of them succeeds or the list runs out.

This leads to small errors in various places, for example in the Step 4 rules

(m>1)ement →

(m>1)ment →

(m>1)ent →

to remove final *ement*, *ment* and *ent*.

Properly, *argument* stems to *argument*. The longest matching suffix is *-ment*. Then stem *argu-* has measure *m* equal to 1 and so *-ment* will not be removed. But if the three rules are applied in turn, then for suffix *-ent* the stem *argum-* has measure *m* equal to 2, and *-ent* gets removed (Porter, 1980).

Other than English, Kraaij and Pohlman (1994) developed a Porter stemmer for Dutch which uses the implementation presented in (Frakes and Yates, 1992).

But ***over stemming*** mistakes occurred when a word is reduced to a semantically unrelated stem. Another weak spot of the algorithm is that it has no way to handle irregular forms.

All languages contain irregularities that should be accommodated in a stemming algorithm. An English stemmer, for example, can convert regular plurals to singular form without difficulty (boys, girls, hands ...). Should it do the same with irregular plurals (men, children, feet,)? There are irregular cases with i-suffixes, but there are irregularities with d-suffixes, which (Lovins, 1968) calls 'spelling exceptions'. Absorb/absorption and conceive/conception are examples of this.

The Porter stemmer does not handle irregularities at all and false conflations in fact are always problem, for example new and news. i-suffix irregularities in English go with short, old words, that are either in very common use (man/men, woman/women, see/saw ...) or are used only rarely (ox/oxen, louse/lice, forsake/forsook . Conflation of these forms frequently leads to an error like mis-stemming. An algorithmic stemmer really needs holes where the irregular forms can be plugged in as necessary (Porter, 1980).

The problem of Porter stemmer is worse in morphologically rich languages like Afaan Oromo. Afan Oromo suffixes are not quit different from non suffix endings in Afan Oromo. For example –*an* is one of the suffixes that form plural form of noun. But words like *aannan* (milk), *ilkaan* (teeth), *bishaan* (water), *Shan* (five) ends with -*an* which is not suffix.

In this paper, context sensitive rules are used to identify affixes that should be conflated and accepted as part of the root. For example, *Shan* (five) can`t be confilated to "*sh-*" because one of the rules checks whether there is at least single vowel consonant sequence (VC). Additional rules are also required so that the words listed above are left as they are considering the endings as part of the words. It takes time and the rules become very complicated if more and more rules are added to handle every exception. In this paper statistical stemming technique is used as a complement to handle cases that are not covered by rule.

2.2.3 Statistical Stemming

In statistical methods, through a process of inference and based on a corpus, rules are formulated regarding word formation. Some of the methodologies are: frequency counts, n-gram (Mayfield and McNamee, 2003). This approach does not require any linguistic knowledge whatsoever, being totally independent of the morphological structure of the target language.

2.2.3.1 Overview of N-gram Stemming Algorithim

N-gram matching techniques are one of the most common of statistical approaches (Freund & Willett, 1982). It uses similarity measures based on the number of diagrams in common instead of terms, then applying clustering techniques. The measures are based on n-gram similarities, where the n-gram of string is any substring of some fixed length. N-gram stemmers conflate terms based on the number of n-grams that are shared by the terms, and are language independent. Similarity co-efficient between words is calculated

as a factor of the number of shared sub-strings, to then pair words according to the number of n-grams (Adamson and Boreham ,1974). After counting the number of n-grams among the word pairs, the degree of similarity is calculated using the Dice Coefficient (Adamson and Boreham, 1974; Robertson and Willett, 1998) or some other means of determining the degree of association between two binary variables (f-coefficient, odds ratio, or t-score).

Many of the benefits of stemming can be achieved without any knowledge of the target language by indexing using overlapping sequences of n characters. This is because some of the n-grams derived from a word will span only portions of the word that do not exhibit morphological variation. For example, the words *juggle, juggling* and *jugglers* share the common 5-gram *juggl.* N-grams are also entirely language-neutral; no knowledge of a language is required to apply n-gram tokenization to the language beyond selection of a suitable value for n.

The main idea behind this approach is that, similar words will have a high proportion of *n*-grams in common. Typical values for *n* are 2 or 3, these corresponding to the use of digrams or trigrams, respectively. For example, the word *KONKOLAATAA* (Car) results in the generation of the digrams

K, KO, ON, NK, KO, OL, LA, AA, AT, TA, AA, A

and the trigrams

K, *KO, KON, ONK, NKO, KOL, OLA, LAA, AAT, ATA, TAA, AA*, A

where '*' denotes a padding space. There are *n*+1 such digrams and *n*+2 such trigrams in a word containing *n* characters.

In this technique, association measures between pairs of terms are calculated based on shared unique N consecutive letters. Terms that have a similarity above a predefined threshold are clustered and represented with only one term (*Suleiman, and Qasem, 2004)*. The similarity measure (S) based on unique digrams can be computed according to the following Dice formula:

$$S = 2 \times C / (A + B)$$

Where:

A: represents the number of unique digrams in the first word.

B: represents the number of unique digrams in the second word.

C: represents the number of unique digrams shared by both words.

The selected n-gram was chosen based on the relative document frequencies of the n-grams spanning the word under consideration. The least frequent n-gram was used since frequent n-grams are more likely to be part of the morphologically variable part of a word, not the root form. For the words jugglers and juggling using statistics from 110282 newspaper articles McNamee and Mayfield (2003) concluded that the rarest n-gram `jugg' occurs much less often than the suffixes `ers ' and `ing '. Therefore both `ers ' and `ing ' would get replaced with `jugg' which intuitively seems like a good choice.

Other methods for selecting a single n-gram or a small number of n-grams are possible. One could consider variants such as using two disjoint 3-grams or never ending an n-gram on a vowel, or any number of other permutations that might have linguistic merit (McNamee and Mayfield, 2003).

For example, the degree of association of the terms “statistics” (w1) and “statistically” (w2) is described in Table 2.1. Once we have the only bigrams shared by the two words, we apply the Dice coefficient to acess the degree of association of the terms.

Table 2.1: The degree of association of the terms statistics and statistically

Word	2-grams	Unique 2-grams	Share Unique 2-grams
W_1	{*s, st, ta, at, ti, is, st, ti, ic, cs, s*}	{*s, st, ta, at, ti, is, ic, cs, s*}	{*s, st, ta, at, ti, is, ic}
W_2	{*s, st, ta, at, ti, is, st, ti, ic, ca, al, ll, ly, y*}	{*s, st, ta, at, ti, is, ic, ca, al, ll, ly, y*}	

The similarity measure is performed on all pairs of terms in the Infromation Retrival system database or dictionary, giving a matrix of word-word similarities. The words are associated using a technique of grouping and classification, such as clustering. There are a number of variants of clustering which can give different groupings depending on the agglomeration rule applied. The simplest are the single link and the complete link rules, while more complex rules include the widely used Ward method (Michela et al, 2002). Each n-gram is represented as coordinate on a vector and, on the basis of vector similarity

measures, word clustering is effected. Finally, the stem is identified for each word cluster with the same prefix.

The drawback of n-grams is not in retrieval accuracy, but rather in retrieval performance and disk usage. Because each character of a text begins a new n-gram, an n-gram representation of a text contains many more indexing terms than does a word or stem representation. Not only does this produce larger indexes, it also increases the number of disk seeks required to locate all of the postings for a query (Mayfield and McNamee, 2003).

In addition, this technique might perform well with some Latin languages such as English, but it would pose some problems in heavily inflection languages such as Arabic (Goweder et al, 2005). The occurrence of these problems is due to the fact that most textual word variants involve a high rate of infix structure, which affects the computation of similarity measures.

The following example shows the similarity value of words with different meaning:

Word1: *baate* (she has left)

Associated Unique digrams: *ba,aa,at,te*

Word2: *batte* (flate)

Associated Unique digrams: *ba,at,tt,te*

$$S = 2 \times C / (A + B)$$

$$S= 2*3/(4+4)= 0.75$$

Evnthough the two words have different meaning they are stemmed to the same stem as they have large similarity value.

Stemming has been extensively applied to tasks related to IR, such as query expansion (Adamson and Boreham, 1974; Lennon et al., 1981; Cavnar, 1994; Damashek, 1995). At the same time, n-grams have been used in the automatic spelling correction (Angell et al., 1983; Kosinov, 2001) on the assumption that the problems of morphological variants and spelling variants are similar.

2.2.3 Hybrid approach

Hybrid approaches use one or more of the approaches described above in union. A simple example is a suffix tree algorithm which first consults a lookup table using brute force. However, instead of trying to store the entire set of relations between words in a given language, the lookup table is kept small and is only used to store a minute amount of "frequent exceptions" like "ran => run". If the word is not in the exception list, apply suffix stripping or lemmatisation and output the result. It is tempting to provide a middle-ground solution joining the two worlds. Such a stemmer could use a dictionary for all known words and fall back to a heuristic when no stem is available (Wikipedia, 2010).

A hybrid method which incorporates three different techniques for Arabic stemming overcomes the problems associated with the steming algorithms. The three techniques are: affix removal, dictionaries, and morphological analysis. Each technique is individually adapted to resolve the practical problems associated with it. According to the evaluation of the experiments, it can be concluded that an overall accuracy is good which shows stemming can be performed with low error rates in high inflected languages such as Arabic (Goweder et al, 2005).

This paper usesd a hybrid approach that incorporates rule based and n-gram together to develop stemmer for Afan Oromo text. Porter algorithm which is rule based is adopted to develop the stemmer. In this stemmer every rule that works for every morpheme formation is not exhaustively identified because of the nature of the language. Therefor statistical stemming algorithm takes over when there is no rule that applies for a given word.

2.3 Related works

Very limited works have been done in the past in the areas of stemming in relation to Afaan Oromo. One of the stemmer is developed by (Kekeba et al, 2007) and it used to develop an Oromo-English CLIR system that enable user to access and retrieve online information sources that are available in English by using Afan Oromo queries.

The stemmer used a rule based suffix-stripping algorithms focusing on very common inflectional suffixes of Oromo language. This light stemmer is designed to automatically

remove frequent inflectional suffixes attached to headwords (base-word forms) of Afaan Oromo. Some of the common suffixes that have been considered in their light stemmer include gender (masculine, feminine), number (singular or plural), case (nominative, dative), possession morphemes and other related morphological features in Afaan Oromo.

According to Kekeba et al (2007) it is possible to categorize suffixes in Afaan Oromo into three basic groups: derivational, inflectional, and attached suffixes. Attached suffixes are particles or postpositions like *arra, -bira, -irra, -itti, -dha, -f,* etc. that are attached to stem/root words. For instance the word *adunyaarratti* (in the world) is formed from a stem, i.e. *adunyaa* and two attachment suffixes, i.e. *-irra* + *-itti*. Inflectional suffixes are a combination of word stem with grammatical/syntactic morphemes, usually resulting in a word of the same class as the original stem. These suffixes include plural noun markers such as *-oota,* (e.g. *nama* + *-oota* = *namoota* i.e. person + *-s* = persons); *-lee* (e.g. *jabbi* + *-lee* = *jabbilee*, i.e. calf + -es = calves) and *-wwan* (e.g. *indaaqqo* + *-wwan* = *indaaqqowwan*, i.e. chicken + -s = chickens). Derivational suffixes enable a new word, often with a different grammatical category, to be built from stem/root other words. For example, the stem verb *qabuu* + *-eenyaa* becomes *qabeenyaa* which is noun while the adjective *gowwa* + *-ummaa* becomes *gowwummaa*, which is also another Afan Oromo noun.

Kekeba et al (2007) argue that the most common order/sequence of Afaan Oromo suffixes (right to left) is derivational, inflectional and attached suffixes. Thus, the stemmer is expected to remove from the right end first all the possible attached suffixes,

then inflectional suffixes and finally derivational suffixes. They identified and built three different suffixes clusters/lists with respect to the above three major types of suffixes in Afaan Oromo.

The stemmer first starts with consulting the attached suffix list and try to remove all possible such suffixes if any. Then it consults the inflectional suffix list and strips any inflectional suffixes. The following simple stemming example illustrates some of the major steps performed by the stemmer.

This kind of stemming techniques can have several shortcomings when applied to heavily inflection language such as Afaan Oromo. These shortcomings can include:

1. Improper removal of some affixes (part of a word might appear to be a prefix or suffix). The above stemmer can perform in some cases like any other light stemmer but encounters a number of problems. Since Afaan Oromo is one of the morphologically rich languages the process of stemming is often complex and requires detail algorithms. But the above stemmer is designed and developed a rule-based light stemmer for Afaan Oromo focusing only on its major inflectional and attached affixes. It simply strips of any end of a word that matches one of the affixes in a list without any condition to be tested. The only rule that is followed by the stemmer is stripping the longest ending of a word that matches one of the suffixes within the list of suffixes. For example if we take the word *killee* (an egg), *beenyaa* (asset), will be conflated to "*kil-*" and "*be-*" respectively. The two words are nouns that are the root form of their category that do not need to be

conflated according to Afaan Oromo morphological analysis. There fore additional rules has to be incorporated to handle cases like the above one. If we take *saawwan* (cows) is conflated to *"saa-"* because the suffix -*wwan* matches one of the suffixes listed in the stemmer and this is called under stemming. The proper stem has to be "*saaww-*". The problem here is both -*an* and -*wwan* are possible suffixes in Afaan Oromo. Unlike other language like English, Afaan Oromoo suffixes are not quite different from non suffix endings. Suffixes like -*aa* as in the case of *maqaa,mucaa;-ee,-lee* as in the case of *killee,eelee,itilee,mee,kee,ree*; -*an,-n* in the case of *aannan,ilkaan,Afaan,shan,kan;* -*tuu,-tu* as in the case of *utuu, hatuu, nyaat*u, *baatu, kaatu* are part of a root word as indicated in the above words that should not be conflated. Therefore most word endings can be part of a word and suffix as well. So simple removal of endings can create invalid stems. But the above stemmer conflates the words since the endings of the words matches' one of the suffixes.

2. The stemmer didn`t consider words formed by duplication of some characters at all. But Afaan Oromo is rich in this kind of word formation. Most of the adjectives form the plural by reduplication of the first syllable. For example words like, *jajjabaa (stron-plural), gaggabaabaa (short-plural)* are formed from -*jabaa* and -*gabaabaa* by duplicating the first syllabus, respectively.

Another stemmer is developed by Wakshum (2000) which use suffix table in combination with rules that strips off suffix from a given word by looking up the longest

match suffix in the list (Wakshum, 2000). List of suffixes are compiled aoutomatically by counting the most frequent endings and other linguistically valid suffixes are also included manually. The stemmer fined the longest suffixes that matchs the end of a given word and remove. He suggested other stemming algorithms to be tried for Afaan Oromo. The problems he faced are similar with that of (Kekeba et al, 2007). Some of these are: irregular formation of variants from root word, the challenge to increase the number and complexity of the rules and words formed by duplication of some characters. This necessitates the need for the detailed knowledge of a language to use such frameworks to generate a stemmer that handles words formed by duplication of some characters and other under/over stemming problems.

Another problem of the stemmer by Wakshum (2000) is that the lists of the suffixes are not linguistically valid. Out of 342 suffixes compiled by the stemmer only 70 are linguistically valid. This is because the lists of suffixes are compiled statistically and not based on the analysis of the language. The suffixes are compiled by counting and sorting the most frequent word endings. One great problem occurred with this kind of compilation of suffixes is that during conflation frequently occurring endings which are part of root word is considered as suffixes and removed. The only condition considered to conflate a word is the matching of suffixes with the ending of word and also the number of character must be at least 3 letters.

In Wakshum`s (2000) modified stemmer very few contexts are added in order to reduce the error rate. But the contexts are based on the observed result of the stemmer running

on the test data. And this kind of context is not based on the analysis of the morphology of the language. Therefore the added context are not general and produced errors that didn`t appear in the first version of the stemmer even though some errors are corrected (Wakshum 2000).

The compilation of stop words is also done statistically and frequently occurring content bearing words are also included. For example,*barannoo,barattoo,barnoota* are varieties of the root *barat*(to learn) ,*duree*(rich),*fayyadam*(to use),*dhiyeess*(to approach),*barsiisu*(to teach), *barsiisa* (teacher), *agarsiis* (to show) and the like are included as stop word. This indicates that frequently occurring content bearing words of Afan Oromo are not considered by the stemmer. More than 96 content bearing words that occur frequently are included as stop word.

Generally, Stemming increases recall while harming precision. As an example of what can go wrong, note that the Porter (1980) stemmer stems all of the following words: *operate, operating, operates, operation, operative, operatives* and *operational* to oper. However, since *operate* in its various forms is a common verb, we would expect to lose considerable precision on queries such as the following with Porter stemming: operational AND research, operating AND system, operative AND dentistry For a case like this, even moving to using a lemmatizer would not completely fix the problem because particular inflectional forms are used in particular collocations: a sentence with the words *operate* and *system* is not a good match for the query operating AND system.

Getting better value from term normalization depends more on pragmatic issues of word use than on formal issues of linguistic morphology (Manning et al, 2009).

Chapter Three: Morphology of Afan Oromo

3.1 Introductions

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. Currently, it is an official language of Oromia state (which is the largest Regional State among the current Federal States in Ethiopia). It is used by Oromo people, who are the largest ethnic group in Ethiopia, which amounts to 34.5% of the total population (Census, 2008). With regard to the writing system, Qubee (a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991.

3.2 Morphology

Morphology is a branch of linguistic that studies and describes how words are formed in a language (Hull, 1995). There are two kinds 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. Derivational morphology deals with those changes that result in changing classes of words (changes in the part of speech). For instance, noun or an adjective may be derived from a verb.

3.3 Types of Morphemes in Afan Oromo

There are two categories of morphemes: free and bound morphemes. Free morpheme can stand as a word on its own where as bound morpheme does not occur as a word on its

own (Schiffman, 1999). In Afaan Oromo roots are bound as they can not occur on their own like "*dhug-*" (drink) and "*beek-*" (know), which are pronounceable only when other completing affixes are added to them. In other words these roots serve as base stems in Afaan Oromo since they posse's non-verbalized glosses (Oromoo 1995).

Like the root, an affix is also a morpheme that can not occur independently. It is attached in some manner to the root, which servers 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 form a word. In *beekumsa* (knowlwdge), for instance, *-umsa* is a suffix and *beek-* (know) is a stem. An infix is a morpheme that is inserted within another morpheme. Like English, Afan Oromo does not have infixes as far as I could ascertain from the existing literature.

There are a wide range of word formation processes in Afaan Oromo. In this research, the morphological analysis of the language is organized in to 6 catagories. The catagories are: nouns, pronouns and determinants, case and relational concepts, functional words, verb and adverbs. Almost all 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. Like wise, determinants have number, gender, adjectives, and quantifier markers similar to Afan Oromo nouns. Afaan Oromo verbs are also highly inflected for gender, person, number, tenses, voice and transitivity. Furthermore, prepositions, postpositions and article markers are often indicated through affixes in Afaan Oromo. Since Afaan Oromo is morphologically very productive, derivation,

reduplication and compounding are also common in the language (Oromoo 1995).The detail of the catagories is described in the following sections.

1. Noun

1.1 Gender

Limited group of noun differ by using different suffixes for mascular and faminiene form. The language use *-ssa* for masculine and *-ttii* for feminine. For example *obboleessa* (brother) takes

-ssa and *obboleettii* (sister) takes *-ttii*. Natural female gender corresponds to grammatical feminine as in the case of sun, moon etc. names of towns, countries, rivers are also feminine. 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 (qajeelcha, 1998). Biiftuun *baate* (the sun rose). The word *baate* takes *-te* to show feminine gender. We can see that *-tii* can also show feminie gender in the following statement. *Adurreen maal ariitii?* (What does the cat run after?) (Mewis, 2001).

1.2 number

Afaan oromoo has different suffixes to form the plural of a noun. The use of different suffixes differs from dialects to dialects. Majority of noun plural froms were formed by using the suffix *-(o)ota,* followed by *-lee,-wwan,-een,-olii,-olee* and *-a(n)* (Mewis, 2001).

-o(ota)	hiriyoota	-aan	ilmaan
-wwan	hojiiwwan	-olii/olee	Jaarsolii/jaarsolee
-lee	gaaffilee	-een	fardeen

1.3 Definiteness

Demonstrative pronouns like *kun* (this), *sun* (that) are used to express definiteness. In some Afan Oromo dialects the suffix *-icha* for male and *-ittii(n)* for female and for undermining usually has a singularize function is used where other languages would use a definite article. For example

Afaanichi	afaanicha	Jaarttiin	
jaarttii			
Jaarsichi	jaarsicha	Re`ettiin	re`ettii

1.4 Derived noun forms

Afan Oromo is very productive in word formation by different means. One method is the use of different derivational suffixes. The other method is the formation of compounds (Mewis, 2001).

1.4.1 Derivational 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. The following suffixes play an important role in oromoo word derivation. They are *-eenya, -ina, -ummaa, -annoo, -ii, -ee, -a, -iinsa, -aa,-i(tii), -umsa, -oota, -aata,* and *–ooma.*Examples:

-eenya	jabeenya(strength)	*-(o)oma*	firooma(friendship)
			Fakkeenya(example)
-ina	guddina(growth)	*-tuu*	furtuu(key)
	Dheerina(length)		barattuu(student)
-ummaa	haxxummaa(haxxummaa)	*-oota*	barnoota(education)
-annoo	yaadannoo(rememberance)	*-umsa*	barumsa(science of)
-ii	hawwii(ambition)		Beekumsa(knowledge)

	Butii(hijacking)	*-(t)tii*	fakkaattii(image)
-ee	dhibee(problem)		Abbaltii(intension)
-a	yaada(thought)	*-aa*	amantaa(belief)
-iinsa	bulchiinsa(administration)		

1.4.2 Compound Words

Genitive construction is a method of forming compound nouns.

Example:

Abbaa gadaa (traditional Oromo president)

In addition, two nouns, noun and verb, adverbs or prepositions and nouns can be combined to form a compound. Example: *durataa`aa (*chairman)

2. Pronouns and Determiners

2.1 Adjectives

2.1.1 Gender

We can divide the Oromo adjectives in to four groups with respect to gender making adjectives (Mewis, 2001).

i. In the first group the masculine form terminates in *-aa* and the feminine form in *-oo*. example:

Masculine	feminine
Guddaa (big)	*guddoo*(big)

ii. In the second group the masculine form terminates in *-aa*, the feminine from in *-tuu,-du.*

For example

Dheeraa (tall)	*dheertuu* (tall)

iii. Adjectives that terminates in *-eessa* or *-(a)acha* have a feminine form in *-eettii* or *-aattii*. Example:

Nama dureessa (a rich man)	*dureettii* (rich woman)

iv. Adjectives whose masculine form terminates in a long vowel other than -*aa* as in i) or ii) or in short a (but not of the suffix -*eessa/aacha*) are not differentiated with respect to their gender.

2.1.2 number

i. Most of the adjectives form the plural by reduplication of the first syllable. Masculine and feminine adjectives differ in plural as they do in singular.

Table 3.1: Examples of plural adjectives formed by reduplication

singular		plural	
m	f	m	f
Gudda xinnaa	Guddoo xinnoo	Gugudda xixinnaa	Guguddoo xixinnoo

They take the following form

"Ciccir-"	Cvc+cvc	*"U`utaalu"*	vcvcvvc
"Caccab-"	cvc+cvc	*"Dadammaq-"*	cv+cvc-cvc
"Buburraaq-"	cvc+cvc-cvvc		

ii. There is a further plural form which is gender neutral for adjectives of the second group beside a special masculine and feminine plural. This plural form terminates in -*oo* and is sometimes used with reduplication and sometimes without.

Table 3.2: Examples of plural adjectives formed by reduplication which are gender neutral

singular		plural		plural
M	f	m	f	Gender neutral
dheeraa jabaa	Dheertuu Jabduu	Dhedheeraa Jajabaa	Dhedheertuu Jajjabduu	Dhedheertuu Jajjaboo

iii. Adjectives which may function as nouns as well form the plural only by using noun plural suffixes.

Table 3.3 examples of plural adjectives formed using noun plural suffixes

singular		plural	
m	f	m	f
dureessa	dureettii	Dureeyyii/dureessota	dureettiwwan

iii. Adjectives of the fourth group form the plural without marking the gender, very often by reduplication of the first syllable. Sometimes adjectives of this group form the plural by using a noun plural suffix (Mewis, 2001).

Table 3.4: Examples of plural adjectives formed by reduplication of the first syllable or using noun plural suffixes

singular	plural	English
Adii	A`adii/adaadii	white
collee	colleewwan	active

2.1.3 Definiteness

The demonstrative pronoun that express definiteness in Afan Oromo follows the adjective if the noun is qualified by an adjective and demonstrative pronoun as well.

Example:

Namicha dheeraa sana argitee? (Did you see that tall man?)

The suffix ichi that sometimes has a definite function normally is suffixed to nouns, but it can be suffixed to adjectives or numerals, too.

Example

Lagni guddichi (the big river) *namichi tokkichi* (a single man)

2.2 Numerals and other Quantifiers

2.2.1 Ordinal Numbers

The written sources use the form *-ffaa*.

Tokkoffaa (the first) *Ja`affaa*(the second)

2.2.2 Cardinal Numbers

Examples

Tokko/takka (one) *lama* (two)

2.3 Pronouns

Pronouns are categorized in to six catagories. Namely: personal, demonistrative, possessive, reflexive, reciprocal and introgative pronouns.

3. Case and Relational Concepts

Nouns, pronouns and adjectives are subject to the following case and relational concepts (Oromoo 1995).

3.1 Base Form

The base form is a form of a noun, pronoun or adjective which is used for isolated citation, a direct object, a predicate nominal, to express the different oblique cases (in connection with the corresponding suffixes), for subject in focus.

The base form is usually given in dictionaries because it is that form of a noun, adjective or pronoun that does not have any case ending or suffix.

Examples

Nama (man, person)

Namicha (the man, person)

3.2 Subject Form

The subject form is a form of a noun, pronoun, or adjective which is used for a subject if it is not in focus. It is produced by adding the suffix *–n,-ni* or *–i* to the word.

Examples

Basic form	subject form
Nama(man)	*nam<u>ni</u>*(man)
Harka(hand)	*hark<u>i</u>*(hand)

3.3 Possessive

3.3.1 Genitive Construction

In Afan Oromo the genitive is characterized by the sequence:

Possessed-possessor

Table 3.5: Examples of genitive formation

Base Form		Subject Form		English
Possessed	possessor	Possessed	possessor	
Mana	namaa	Manni	namaa	Some body`s house
Mana	namichaa	Manni	namichaa	The house of the man

Afan Oromo does not have a special genitive marker. The two parts of a genitive construction may be connected by the relative particle *kan (m)* and *tan(f)* but the use of *kan/tan* is not compulsory(Mewis, 2001). The following sequence and examples illustrates the use of *kan* and *tan* as a genitive markers.

Possessed + kan/tan + possessor (+lengthening of a short vowel at the end of the last word.)

Possessed + kan/tan + possessor+ii (for nouns terminating in n).

Example

Farda galmoon arge (I saw Galmoo`s horse). *Farda kan Galmoon arge* (I saw the horse Galmoo).

The genitive is usually formed by lengthening a final short vowel, by adding -*ii* to a final consonant, and by leaving a final long vowel unchanged. The possessor noun follows the possessed noun in a genitive phrase.

Example

Obboleetti (sister), *namichaa* (the man), *obboleetti namichaa* (the man's sister)

Nouns and pronouns terminating in -*n* have to suffix -*ii* if they are the last part of a genitive construction.

Example

Kanneen(those) *kanneennii*(of these)

If the noun has more than one qualifier normally only the last part of the noun phrase has the genitive marker.

Example

Manni nama san<u>aa</u> gaarii dha.(the house of the man is good). In this sentence there are two markers *nama* and *sanaa*, there fore only *sanaa* has genitive marker.

In double genitive constriction only the last part gets a short vowel at the end lengthened.

For example

Seenaa barreessi. (Write your autobiography!)

The sentences has two genitive markers *jireenya* *keetii* and only the second marker`s *keetii* last vowel is lengthened.

3.3.2 Names of Person

The different parts of the name of persons are treated as possessive. We can explain this with the fact that the Afan Oromo does not have family names. The first name in Afaan Oromo is usually the person name; second one the name of the father (Mewis, 2001).

Examples

Maqaan barsiisaa kootii Tulluu dha. (The name of my teacher is Tulluu.)

3.4 Dative

The dative is used for nouns that represent the recipient ‘to’ or the benefactor ‘for’ of an event. The dative form of a verb infinitive (which acts like a noun in Afan Oromo) indicates purpose. The dative can be expressed by lengthening of a short final vowel, lengthening of a short final vowel + adding of a suffix *-f*, adding of the suffix *-f*, *-dhaa* or *-dhaaf* (to nouns with final long vowel), adding of the suffix *-ii* to nouns terminating in consonant, adding of the suffix *-(tii)f* to a genitive construction, and adding of the suffix *-tti*(irrespective of the spelling of the noun)

The following Examples show each of the types:

namich<u>aa</u> buna fidi! (Bring coffee for the man!)

namicha<u>af</u> buna fidi! (Bring coffee for the man!)

Saree<u>dhaa(f)</u> foon kenni!

mee loon<u>ii</u> okaa kenni! (Please give fodder to the cattle!)

mana barumsaa<u>(tii)f</u> (for the school)

3.5 Instrumental

The instrumental is used for nouns that represent the instrument ‘with’, the means ‘by’, the agent ‘by’, the reason, or the time of an event. The formation of the instrumental parallels that of the dative to some extent:

Means of formation of instrumental are

i. noun + n(with lengthening of a short final vowel)
ii. noun + (dhaan)n(to nouns with a long final vowel)
iii. noun + iin(to nouns terminating in a consonant)
iv. noun + tiin(for genitive construction)

-*n* following a long vowel or a lengthened short vowel; -*iin* following a consonant as indicated below.

harka (hand), *harkaan* (by hand), (with a hand)

halkan (night), *halkaniin* (at night)

-*tiin* following a long vowel or a lengthened short vowel

Afaan Oromoo (Oromo language), *Afaan Oromootiin* (in Oromo)

-*dhaan* following a long vowel

yeroo (time), *yeroodhaan* (on time)

3.6 Ablative

The ablative is to represent the source of an event; it corresponds closely to English '*from'*. It is formed in the following ways:

When the word ends in a short vowel, this vowel is lengthened (as for the genitive).

Example

biyya (country), *biyyaa* (from country)

When the word ends in a long vowel, *-dhaa* is added (as for one alternative for the dative).

Finfinneedhaa (from Finfinnee)

When the word ends in a consonant, *-ii* is added (as for the genitive).

Hararii (from Harar)

Following a noun in the genitive, *-tii* is added.

Mana (house), *buna* (coffee), mana *bunaatii* (from café)

An alternative to the ablative is the postposition *irraa* 'from' whose initial vowel may be dropped in the process:

Example

Gabaa (market), *gabaa irraa or gabaarraa* (from market)

3.7 Locative

The locative is used for nouns that represent general locations of events or states, roughly '*at'*. The locative is formed with the suffix *–tti.*

Example

Arsiitti (in Arsii)

4. Functional Words

4.1 Post, pre- and paraposition(pre—postpositions)

Afan Oromo languages use prepositions, postpositions and Para positions (Mewis, 2001), (Oromoo 1995).

i. Suffixed postpositions

-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. Since 1992 the postposition *-rraa* and *-rraa* are used very often with a prothetic i- and spelled then as an independent word.

Examples

Adaamaatti yoom deebina? (When shall we go back to Adama?)

Bantiin sireerra ciise. (Banti lay down on bed.)

ii. Prepositions

Akka (like, according to)

Gara (to, in the direction of)

Hanga/hamma(until, up to)

Karaa (along, the way of, through)

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

Example

Namni <u>akka</u> harkaan waa hojjechuuf fayyadamu arbi maalitti fayyadamaa? (As people use hands to work something what does the elephant use?)

iii. Post positions as independent words

Ala(outside)	*Wajjin*(with, together with)
Bira(beside)	*Teellaa*(behind)
Booda (after, behind)	*Ol*(i) (towards the top, up)
Bukkee (beside)	*Malee* (without)
Duuba(behind)	*Keessa* (in, inside)
Dura (before)	*Jala* (under)
Fuuldura (in front of)	*Irra* (on, above of)
Gad(i) (under/below)	*Gubbaa* (on, over)

Example

Namoota nu <u>bira</u> jiraniis hin jeeqnu. (We don`t hurt people who are with us.)

iv. Parapositions

Gara…tti (to)

Gara…tiin(from, from the direction of)

Hanga…tti/hamma…tti (up to,until)

Example

Lukkichi rifatee jeedaloo dheesuuf gara manaatti garagale. (The cock was scared and went home to take refuge from the fox.)

v. Combinations of postpositions with certain verbs

Jala fiiguu (to run after), *Keessa galuu* (to enter, interfere), *Ittuma dhiisuu* (to *refrain* from, to stop, to leave)

Example

…*man sana* keessa galee… (…he entered that house and …)

4.2 Focus Marker

Focus marker do not have a lexical meaning but only grammatical function. Afan Oromo language has two kinds of focus markers beside several emphatic or focus particles: one for the predicate and one for the subject. The focus marker for the predictate is a pre-verbal particle, while the focus marker for the subject is suffixed to the last constituent of a noun phrase.

4.2.1 The focus marker for the subject

As a focus marker for the subject the suffix *-tu(u)* is used in the northern dialects and in writing. The focus marker *-tu(u)* is added to the base form of the last noun of the nominal phrase, no to the subject from. (Mewis, 2001).

Example

Maaltu oddoo keessatti argama? (What is in the garden?)

4.2.2 The focus marker for the predicate

In the past different morphemes were used in Afan Oromo as predicate marker, example. *hin, in*(both with high tone and stress), *ni*- and in the southern dialects *hin*- for imperfect and *ha(a), ya(a),la ,layu or yayu* for perfect. Since Afan Oromo has been used as a written language and as a medium of instruction from 1992 on we can observe that the particle ni, which was only used in Harar-Oromo previously, is used as the standard from now.

Example

Manni barumsaa wiixata ganama *ni* jalqaba. The school will start on Monday morning

4.3 Conjuctions

Conjuctions are unchanging words which coordinate sentences or single parts of sentence. The main task of conjuctions is to be a syntactical formative element that establishes grammatical and logical relations between the coordinated constituents.examples:

Ammo(but,however),*-f/-fi*(and,that),*Kanaaf jecha*(there fore, because)

5. Verb

Afan Oromo has base stems and four derived stems at its disposal (Mewis, 2001) (Mewis, 2001), (Oromoo 1995).

5.1 Derived Stems

The four derived stems the formation of which is still productive in Afan Oromo are:

Autobenefactive	(AS)
Passive	(PS)
Causative	(CS)
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. 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. The derived stems may be formed from all verbs the meaning of which permits it (Mewis, 2001).

ii. Autobenefactive

The Afan 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.

Example

bitachuu (to buy for oneself).

The conjugation of a middle verb is irregular in the third person singular masculine of the present and past (*-dh* in the stem changes to *-t*) and in the singular imperative (the suffix is *-u* rather than *-i*).

Examples

bit- (buy)	*qabanna*(we hold)
*bitadh-(*buy for oneself)	*qabadh-* (seize, hold (for oneself))

Infinitive and participles are always formed with *-(a)ch,* while the imperative forms have *-(a)(a)dh* instead of *-(a)ch.*

Infinitive	imperative sg.	Imperative pl.	English
Argachuu	*argadhu*	*argadhaa*	(to find/get)

Table 3.6 Examples conjugated forms that have -dh only in the first person singular

Argachuu to find /get	waammachuu (to call up on)
Sg. 1.p. n argadha	n waammadha
Sg. 2.p. argatta	waammatta
Sg. 3.p.m. argata	waammata
Sg. 3.p. f. argatti	waammatti
Pl. 1.p. arganna	waammanna
pl. 2.p. argattani	waammattani
pl. 3.p. argatani	waammatani

iii. Passive

The Oromo passive corresponds closely to the English passive in function. It is formed by adding *-am* to the verb root. The resulting stem is conjugated regularly.

Example

beek- (know) *beekam-* (be known)

iv. Causative

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

Example

Deemuu (to go) *deemsisuu* (to cause to go)

A second causative of an intransitive verb would create a real causative.

Base stem	causative I	causative II
Agarsiisuu	to show	waamsiisuu (to cause to call)
Sg. 1.p.	agarsiisa	waamsiisa
Sg. 2.p.	agarsiifta	waamsiifta
Sg. 3.p.m.	agarsiisa	waamsiisa
Sg. 3.p. f.	agarsiifti	waamsiifti
Pl. 1.p.	agarsiifna	waamsiifna
pl. 2.p.	agarsiifti	waamsiiftu
pl. 3.p.	agarsiisu	waamsiisu

A base stem terminating in *l* will get a causative stem formed by means of *–ch.*

Example

Galuu (to enter, return home) g*alchuu* (to take home, let enter)

Verbs whose roots end in ′ drop this consonant and may lengthen the preceding vowel before adding -*s*.

Example

Ka`uu (to rise /get up) *kaasuu* (to lift up/arouse)

For the conjunction of causative some assimilation occurs as in the following:

s+n ->fn ***S+t ->ft***

Example

Haati harma hoosifte. (The mother breast-feeds.)

v. Intensive

It is formed by duplication of the initial consonant and the following vowel, geminating the consonant.

Example

Waamuu (to call, invite) *wawwaamuu* (to call intensively)

vi. Complex derived stems

The derived stems can be combined with each other in different sequences.

Example

Arguu(to see) -argamuu(to be seen) -argamsiisuu (making to be seen)

Lafti gabbate immoo oomisha gaarii argamsiisa. (*Fertile land brings about a good harvest)*

ba'uu - baasuu - baafachuu

galuu - galchuu - galfachuu

Afuura baafachuu fi galfachuudhaaf bishaan qaama keessa jiraachuun barbaachisaa dha.

(There has to be water in the body in order to breath out and to breath in)

Before the passive and autobenefactive affixes -*am* and –*adhi/at* the affix –*s* becomes-*f* as in *deebi+s+am+uu =deebifamuu*.

Examples

deebiSuu *(to return, repeat)* deddebisuu *(to repeat)*

deebisiisuu *(to return, answer)* deddeebi'uu *(to go back and forth keep repeating)*

5.2 Simple Tenses

5.2.1 Infinite forms

5.2.1.1 Infinitive

The infinitive form of verbs terminates in -uu.

Examples

arguu (*to see)* *deemuu* (*to* go)

The infinitive form of autobenefactive verbs terminates in -chuu.

Example

jiraachuu (*to live)* *bitachuu* (*to buy for oneself)*

5.2.1.2 Participle/ gerund

An active participle is formed by adding -aa to the verb stem (Mewis, 2001).

Example

deemaa (*going)* *jiraachaa* (*living)*

According to the meaning of the verb these forms may serve as agent nouns.

Barsiisaa (*teacher)* gaafatamaa (*responsible person)*

For these agent nouns feminine forms are used according to the pattern of feminine adjective formation.

Barsiiftuu (teacher) *gaafatamtuu (responsible person)*

A gerund is formed by adding -naan to the verb stem.

Deemnaan(*after having gone)* *nyaannaan* (*after having eaten)*

5.2.2 Imperative

Imperative singular of base stems and all derived stems beside autobenefactive stems is formed by means of the suffix -*i*.

Example

Deemi!(go!) *argi!(look!)*

The imperative singular of autobenefactive stems is formed by means of the suffix -*u*.

Example

Jiraadhu! *(live!)*

Imperative plural of all stems is formed by means of -*aa*.

Example

deemaa! (*go!)* *Argaa!* (*see!)*

Negative imperatives are formed by means of -*(i)in* for singular and -*(i)inaa* for plural.

Example

Qubaan jechoota irra hin deemiin. (Don't point on the words with your finger.)

5.2.3 Finite Forms

The Oromo language uses different conjugations for the verbs in main clauses and in subordinated clauses for actions in present or near future. The first person singular is differentiated from the third person masculine by means of an *-n* that normally is suffixed to the word preceding the verb (Oromoo 1995).

5. 2.3.1 Present tense main clause conjugation

The present tense main clause conjugation is characterized by the vowel *-a*:

Deemuu (*to go)*

sg. 1.p.	*deema*
2.p.	*deemta*
3.p.m	*deema*
3.p.f	*deemti*
pI. l.p.	*deemna*
2.p. andpoliteform	*deemtu/deemtan(i)*
3.p. andpoliteform	*deemu/deeman(i)*

Examples

Gara mana yaalaan deema. (*I go to the laboratory.)*

5.2.3.2 Past tense conjugation

The past tense conjugation is characterized by the vowel *-e:*

Deemuu (*to go)*

sg. 1.p.	*deeme*
2.p	*deemte*
3.p.m	*deeme*
3.p.f	*deemte*
pI. l.p.	*deemne*
2.p. andpoliteform	*deemtani*
3.p. andpoliteform	*deemani*

Example

Kumsaan gara mana barumsaa deeme. (Kumsaa *went to the school.)*

5.2.3.3 Subordinate Conjugation

The subordinate conjugation is used in affirmative subordinated clauses and in connection with the particle haa for the jussive. Beside this the subordinate conjugation is used to negate present tense actions.

Deemuu (*to go)*

sg. l.p	*akkan deemu*
2.p.	*akka deemtu*
3.p.m.	*akka deemu*
3.p.f.	*akka deemtu*
pI. l.p.	*akka deemnu*
2.p. andpoliteform	*akka deemtani*
3.p.andpoliteform	*akka deemani*

Examples

Akkan yaadutti biqiltootni guutaniiru. (*As I thought there are many plants.)*

5.2.3.4 Contemporary verb conjugation

The contemporary verb conjugation is used only in connection with the temporal conjunction *-odoo,-otoo,-osoo,-otuu* or *-utuu* that being connected with this conjugation means '*while'* .The contemporary verb conjugation is a kind of subordinated conjugation with lengthened final vowels(Mewis, 2001).

Example

"Otuun isin waamuu maaliif deemta ?" jedhe.(*"While I was calling you (pI.)why do you go?" he said.)*

5.2.3.5 Jussive

To form the jussive in Afan Oromo the particle haa has to be used in connection with the subordinate conjugation.

Example

Isaan haa deemani (they shall go)

5.2.4 Negation

Present tense main clause actions are negated by means of the negative particle hin and the verb in subordinate conjugation.

Example

Maannaaloon hin jiru. (*Maannaaloo is not present.)*

Present tense actions in subordinated clauses are negated by means of the negative particle hin and a suffix -ne that is used for all persons. Past tense actions are negated in the same way using the particle hin and the suffix *-ne.*

Example

Sinbirroon halkanii bakka namni arguu hin dandeenve jiraatu. (*Bats live in places that people cannot see.)*

5.3 Compound tenses

5.3.1 Perfect

The perfect tense is used for actions that have happened in the past and are not lasting to the present time.

The perfect tense is formed by means of the paradigms for consecutive actions and the conjugated present tense forms of the verb jiruu *'to be' (in a place).* Beside the full forms contracted forms are in wider use.

Examaple

Waan haaraan tokko akka uumame tilmaamee jira.(it was guessed that something new was created)

5.3.2 Past Perfect

The past perfect tense is used for actions that have happened in a remote past and are not lasting till the present. The past perfect is formed by means of the paradigm for consecutive actions and the conjugated past tense forms of turuu '*to have been'*. It has the following general form.

V (CA) + turuu (past)

Example

*Callise bira dabruu yaalee ture.(*he was trying to pass by silent)

5.3.3 Present Progressive

The present progressive is used for actions that are not yet completed, that are in progress in the moment described by the speaker or writer. In Afan Oromo a present progressive is formed by means of the present participle of the main verb and the conjugated present tense forms of the verb jiruu '*to be'* or by means of the main verb in infinitive with the suffix -*tti* and the verb jiruu in present tense:

V (participle) + jiruu (present)

V (infinitive) +	**tti + .iiruu (present)**

Example

Ishiin deemuutti jirti.(she is going)

5.3.4 Past Progressive

The past progressive is used for actions that were described by the speaker or writer as having been in progress in the past. The past progressive is formed in Oromo by means of the present participle of the main verb and the conjugated past tense forms of the verb turuu '*to have been'*. (Mewis, 2001).

V (participle) + turuu (past)

Example

Hoggaa kormaan Lukkuu iyyaa ture jeedo araddaa seente.(when the cock was crying the fox entered the garden)

5.3.5. Past perfect progressive

The past perfect progressive is formed by means of the participle of the main verb, the past tense of the verb turuu and the present tense of the verb jiru. It is used for actions that happened in a remote past (Oromoo 1995).

V (participle) + turuu (past tense) + jiruu (present tense)

5.3.6. Past habitual

The past habitual is used to express a habitual, lasting or frequently repeated action that took place in the past. The past habitual is formed with the infinitive of the main verb and the conjugated past forms of the verb turuu '*to have been'*.

V (infinitive) + turuu (past)

5.3.7 Future Definite

Verbs in future tense do not occur frequently. In most cases the present tense is used to express future actions.

Example

Ni rooba. (It will rain.)

The structure of a definite future is:

V (infinitive) + fi

Example

Waan nuti jenne ta'uufi. (*What we've said will happen.)*

5.3.8 Future Perfect

A future perfect would be formed by:

V (infinitive) + -f + turuu (past tense)

5.3.9 Future Indefinite

The future indefinite is used to express actions that may happen in the future. For some examples the English translation was given by the informants in future tense, sometimes in indefinite future. The decision seems to depend on the context. (Mewis, 2001).

V (present tense) + ta'a/ taha

Example

Waan nuti jenne ni ta'a. (*What we've said will/may happen.)*

5.3.10 General past

The general past is used to express an action or a state that took place in the past, and is reported of with respect to the state or action irrespective of beginning and end of them. It has some similarity with the past progressive, but seems that it is not identical with it (Oromo, 1995).

V (present tense) + turuu (past)

5.4 Verb Derivation

Some Oromo verbs are derived from nouns or adjectives by means of an affix *-oom*. These verbs usually express the process of reaching the state or quality that is expressed by the corresponding noun or adjective. From these process verbs causative and autobenefactive stems may be formed. Examples

danuu (*much, many, a lot)*	*guraacha* (*black)*
danoomuu (*to become much)*	*gurraachomuu* (*to become black)*

Causative verbs, however, can also be derived directly from adjectives or nouns by suffixing a causative affix *-eess* to the stem of the noun or adjective, example:

danuu (*much)*	*daneessuu* (*to increase, multiply)*

Another means to derive process verbs from adjectives in Oromo is to form an autobenefactive stem

Example

Adii(*white)*	*addaachuu* (*to become white)*

5.5 Compound Verbs

In addition to the above discussed derived verbs, compound verbs can be formed by means of pre-/postpositions, pronouns and adverbs in Afan Oromo such as *ol* (*above), gad* (*below), wal, waliin, walitti, wajjin* (*together), itti(in)* (*to), keessa* (*in), (into), jala* (*under);* they precede different verbs and express a broad variety of meanings(Qajeelcha, 1998).

Examples

gadi dhiisuu (*to let go of)* *gaddhiisuu* (*to let* go *of)*

Compound verbs can also be formed with *jechuu* or *gochuu*.

Example

With jechuu with gochuu

caljechuu (*to be quiet, silent) cal gochuu* (*to make quiet silent)*

5.6 'To be' and 'to have'

Afan Oromo has different means to express *'to be'*. One of them are copulas, other means are the verbs ta'uu, jiruu and turuu (Mewis, 2001).

The morphemes *(-)dha* and *(-)ti* (suffixed or used as independent words) serve as affirmative copulas as well as the vowel *-i* that is added to nouns terminating in a consonant. The copula *dha* is used only after nouns terminating in a long vowel. Negative copula is *miti*, irrespective of the termination of the noun.

Examples

Present tense:

Atis jabaa dha. (*You are strong, too.)*

Nouns terminating in a short vowel do not take any copula.

Example

Isheen durba. (*She is a girl.)*

Nouns and pronouns terminating in a consonant are combined with the copula *-i*.

Example

Kuni bisbaani. (*This is water.)*

In all utterances related to possession only the copula *-ti* may be used.

Example

Hojiin hundee guddinaa ti! (*Work is the basis of development.)*

Present progressive

Waa'een jarreen Axaballaa warra isaaniitiif qofa otuu hin taane uummata naannoofiyyuu hibboo ta'aa iira. (*The life of Axaballaa is like a mystery not only, for his family, but also for the people around him.)*

Past tense

Sangaan kan eenvuu ture? (*Whose ox was it?)*

The forms of the verb qabuu '*to have'* are overlapping with the forms of the verb *qabuu* '*to grasp', 'keep'*.

The verb *qabuu* appears with the meaning '*to have'* only in the present tense and one past tense form. In present tense conjugation both verbs have the same form.

5.7. The Cases Governed by Verbs

Verbs in Afan Oromo usually govern certain cases. The majority of Afan Oromo verbs seem to be used with the base form of nouns (Mewis, 2001).

Verbs governing a dative with *-f*:

Example

Jaarsichis "Waggaa saddeettama" jedhee deebiseef. (*The old man said "Eighty years", and answered)*

Verbs governing a dative without postposition or suffix:

Example

erguu (*to send)*

ergisuu (*to lend)*

Verbs governing the postposition *-tti:*

Akka abbootiin keenya nutti himanitti, ... (*According to what our fathers told us, ...*)

6. Adverbs

Adverbs have the function to express different adverbial relations such as relations of time, place, manner or measure (Oromoo 1995).

Example of Adverbs of time

Amma (now)

Example of Adverbs of place

achi(tti) (there)

Example of Adverbs of manner

dansatti (fine, properly)

Example of Adverbs of measure:

baay'ee, danuu (much, many, very)

Chapter Four: Development of Stemmer for Afan Oromo Text

4.1 Introduction

The core of every suffix stripper is a set of rules which test whether a word ends with certain character sequence and subsequently delete this sequence. However some strippers are a bit more sophisticated. Instead of deleting a suffix, they can also replace it by another (shorter) suffix or modify the stem itself.

This paper describes the development and evaluation of a stemmer for Afan Oromo. We have chosen to adopt some concepts from the stemming algorithm developed by Porter (Porter, 1980) because it is well known and is frequently used in experimental IR systems.

4.2 The Test Set

A corpus is a collection of texts or speech stored in an electronic machine-readable format (Sandipan et al 2004). Balanced corpus is needed to process natural language processing tasks like stemming. Balanced corpus is a corpus that represents the words that are used in a language. As indicated in (Sandipan, Sarkar and Basu, 2004), texts collected from a unique source, say from scientific magazines, will probably be biased toward some specific words that do not appear in everyday life. Such types of corpora are not balanced, therefore they are not appropriate for many natural languages processing tasks in general and stemming in particular except in special cases.

However, developing a balanced corpus is one of the difficult tasks in NLP research because it requires collecting data from a wide range of sources: fiction, newspapers, technical, and popular literatures etc which demands much time and human effort.

For this particular study, the used corpus was compiled from different popular Afaan Oromo newspapers (Bariisaa, Bakkalcha Oromiyaa and Oromiyaa) and bulletins (Qabee and Oromiyaa) to make the corpus address variety of issues that reduces the biasedness of the corpus. Newspapers, bulletins and public magazines are considered as addressing different issues of the community: social, economical, technological, political etc. Therefore it is believed that they are a potential source for collecting corpus that addresses different issues of the community for natural language processing tasks. This corpus is used for evaluating the performance of the stemmer. The corpus consists of 198 sentences (the total of 2458 tokens).

4.3. Compilation of Stop Word List

Stop word list are a list of words that should not be stemmed by the stemmer as they are non content bearing words.As can be seen from the sample in table 4.1, the stop word list consists of prepositions, conjunctions, articles, and particles. The stop word list is collected and compiled based on information in the books: *A Grammatical sketch of Written Oromo* (Mewis, 2001) and *Caasluga Afaan Oromoo, Jildi I* (Oromoo 1995). The list of linguistically valid Afan Oromo prepositions, conjunctions, articles, particles are available on the above mentioned books.

Table 4.1 Afan Oromo stop word list examples

Number	word
1	kan
2	sun
3	ani
4	ini
5	isaan
6	iseen
7	isaa
8	akka

There are no any content-bearing words in the list. The complete list of the stop word list compiled is given in Appendix I.

4.4 Afan Oromo Stemmer

It is not possible to apply the stemmer developed for English or other languages like Porter's (Porter, 1980) to Afan Oromo due to differences in the patterns of word formations and differences in their morphologies. Some of the concepts from the Porter stemmer`s (Porter, 1980) are however adopted to develop a stemmer for Afan Oromo. Specifically, concepts about measure, arranging the rules in clusters ,analyzing word formation based on the nature of their endings(for example words that attaches *–de* suffixes ends with b/g/d in Afan Oromo) are taken from Porter algorithm.

The Afan Oromo stemmer is based on a series of steps that each removes a certain type of affix by way of substitution rules. These rules only apply when certain conditions hold,

for example, the resulting stem must have a certain minimal length. Most rules have a condition based on the so-called ***measure***. The measure is the number of vowel-consonant sequences (where consecutive vowels or consonants are counted as one) which are present in the resulting stem. This condition must prevent that letters which look like a suffix but are just part of the stem will be removed. Other simple conditions on the stem are:

- ✓ *Does the stem end with a vowel?*
- ✓ *Does the stem end with a consonant?*
- ✓ *Does the stem end with specific character?*
- ✓ *Does the 1st syllabus of the stem repeated?*

4.4.1 Extensions to the Porter Implementation

As described above most concepts are taken from Porter`s stemmer (Porter, 1980). In addition to concept taken from Porter`s stemmer, Afan Oromo words that form repeatition by duplicating some of the starting characters are covered by this stemmer. Because this affix exhibits certain pattern that can be recognized, the algorithm has been extended to handle them. The original Porter stemmer only treats suffixes.

4.4.2 Affix-rules for Afan Oromo Stemmer

The affix-rules for Afan Oromo were written based on information in books: *A Grammatical sketch of Written Oromo* (Mewis, 2001) and *Caasluga Afaan Oromoo, Jildi I* (Oromoo 1995).

Several criteria were taken into consideration while defining the coverage of the rule clusters, the following being the most important ones:

- Inflectional morphology should be covered as fully as possible. Most Inflectional affixes (e.g. plural endings, verbal inflection etc.) are believed to not affect the basic meaning of the underlying stem and can therefore be removed without risk of losing too much information.
- The most frequent affixes and Derivational affixes should be covered by the stemming algorithm. Rarely occurring affixes are not considered to reduce the complexity of the rule (they are assumed to be handled by the statistics).

Taking these considerations into account, 7 rule clusters were created for the stemmer. Each cluster represents a particular class of affixes and the rules within a class are ordered and mutually exclusive, i.e. if the first rule that matches is applied, no other rules in the same cluster are tried in a particular iteration. The affix-clusters are defined by the similarity of their pattern in word formation, the level at which the affixes occur in the word formation process and the length of the affixes. For instance, the most common order/sequence of Afaan Oromo suffixes (within a given word) is: <stem> <derivational suffixes> <inflectional suffixes> <attached suffixes> (Kekeba et al, 2007).

Thus, their stemmer removes (from the right end of a given word) first all the possible attached suffixes, then inflectional suffixes and finally derivational suffixes step by step. This is done to reduce computational time. Affixes that are fromed out of this sequence can also be removed though it takes additional time.Complex affixes are thus removed in consecutive steps. For example, *barattootarratti* (on the students) has four suffixes: *-itti, -rra, -oota* and *-at*. Therefore firs *-tti*, then *-rra*, then *-oota* and finally *-att* is removed to get the root "*bar-*".

In addition to the affix-rules, a number of special conditions had to be designed to cover some specific phenomena. Examples of these conditions are, for instance, Ends with V/C, i.e. when remaining stem ends in a vowel or consonant as discussed in section 4.4.

Two versions of the algorithm were developed. The first version is totally rule based. As this stemmer is not exhaustive enough to include every rule, in the second version, statistics is used to complement the rule so as to handle cases that couldn`t be cought by the earlier.

The affix-rules have the following general form:

Affix-----→substitution measure-condition <additional conditions>

Where:

Affix: *either prefix or suffix to be removed or subsitituted with another one.*

Substitution: *Affix that is subsistituded with the affixed attached to a given word.*

Measure-condition: *the number of vowel consonant sequence occurring in a given word.*

Additional condition: *condition in addition to measure condition to be tested.* ***For*** *example, whether the word ends with vowel or consonant.*

The first version of Afan Oromo stemming algorithm

Following is the algorithm developed to conflate word variants for Afan Oromo text:
1. READ the next word to be stemmed
2. OPEN stop word file
Read a word from the file until match occurs or End of File reached
IF word exists in the stop word list
Go to 5
Else
Go to 3
3. If word matches with one of the rules
Remove the suffix and do the necessary adjustments
Go back to 3

ELSE
Go to 6
4. Return the word and RECORD it in stem dictionary
5. IF end of file not reached

Go to 1

ELSE

Stop processing

6. IF there is no applicable condition and action exist

Remove vowel and return the result
Go to 4

Algorithim 1.

4.4.2 .1 Definitions of Afan Oromo Stemmer

Define a vowel as one of

a e i o u

Define consonants as one of

` b c d f g h j k l m
n p q r s t v w x y z

Define a valid **de, du, di, do, dan**-ending as one of

b g d

R1 is the region after the first non-vowel following a vowel. If the word strts in vowel it is the region before the next consonant.

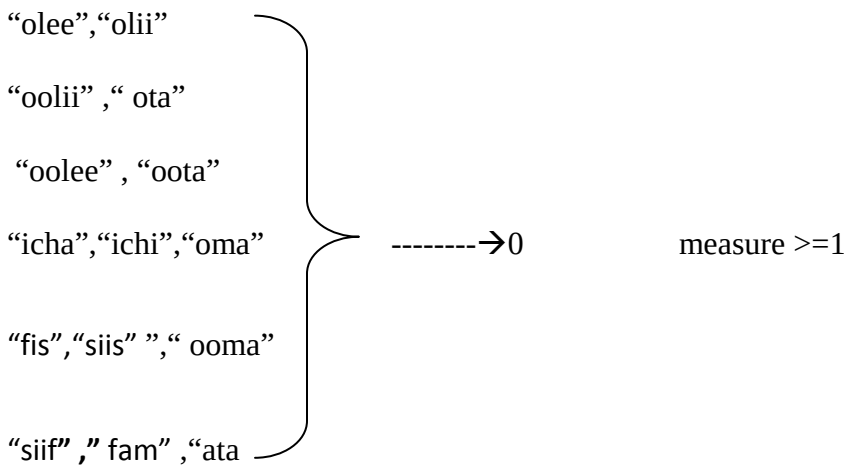
R2 is the region after the first non-vowel following a vowel in R1

C1 is the firs character of a word.

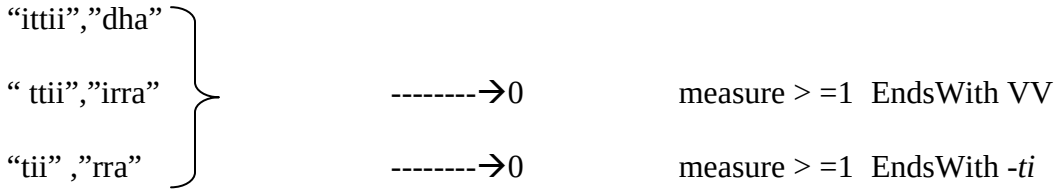
4.4.2.2 The rule clusters

The clusters of the rules are described as follows:

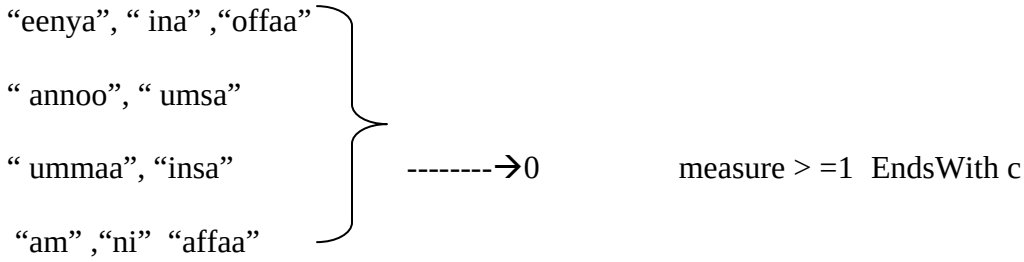
The first cluster of the rules covers suffixes that are deleted if measure is greater or equal to 1.



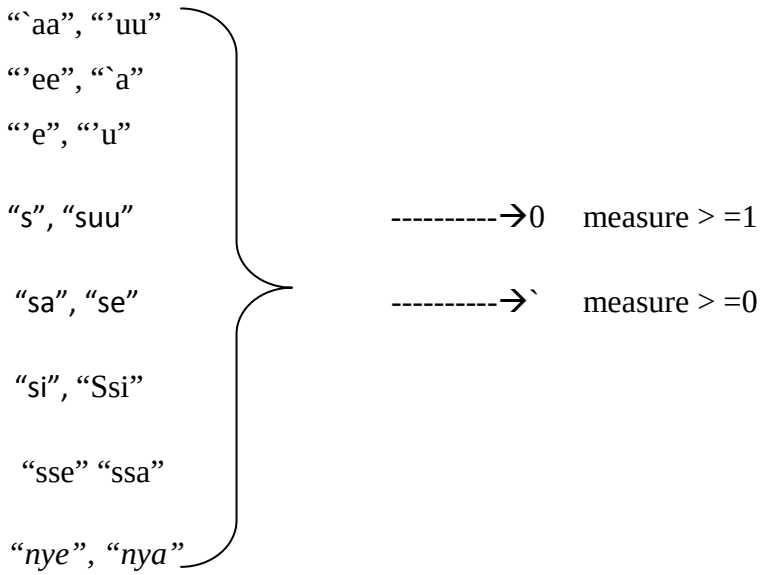
The second cluster suffixes that are deleted if ends with double vowels or the suffix *-ti*



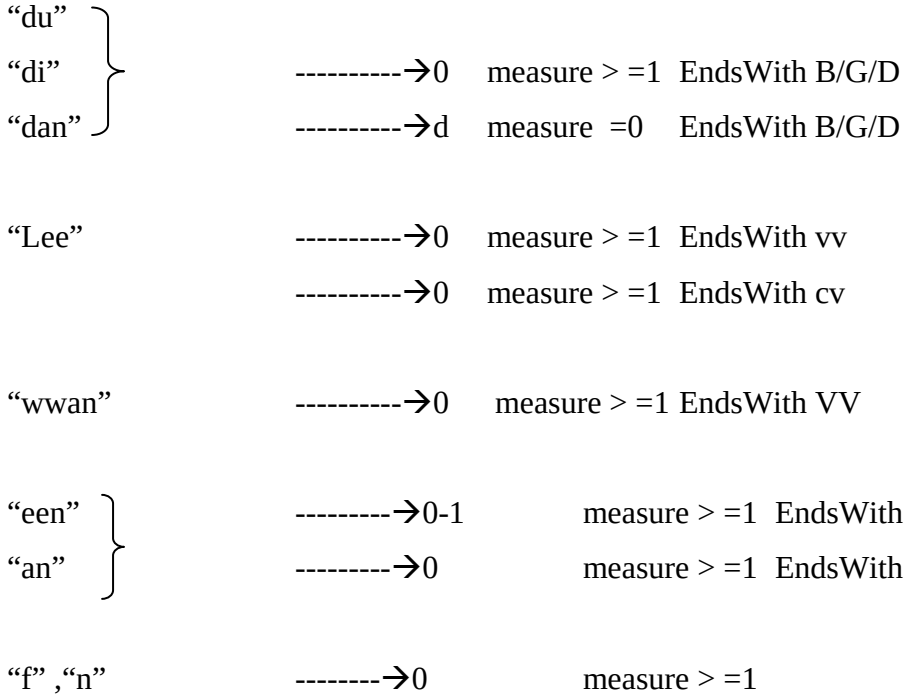
The third cluster covers suffixes that are deleted if measure is greater or equal to 1 and ends with consonant.



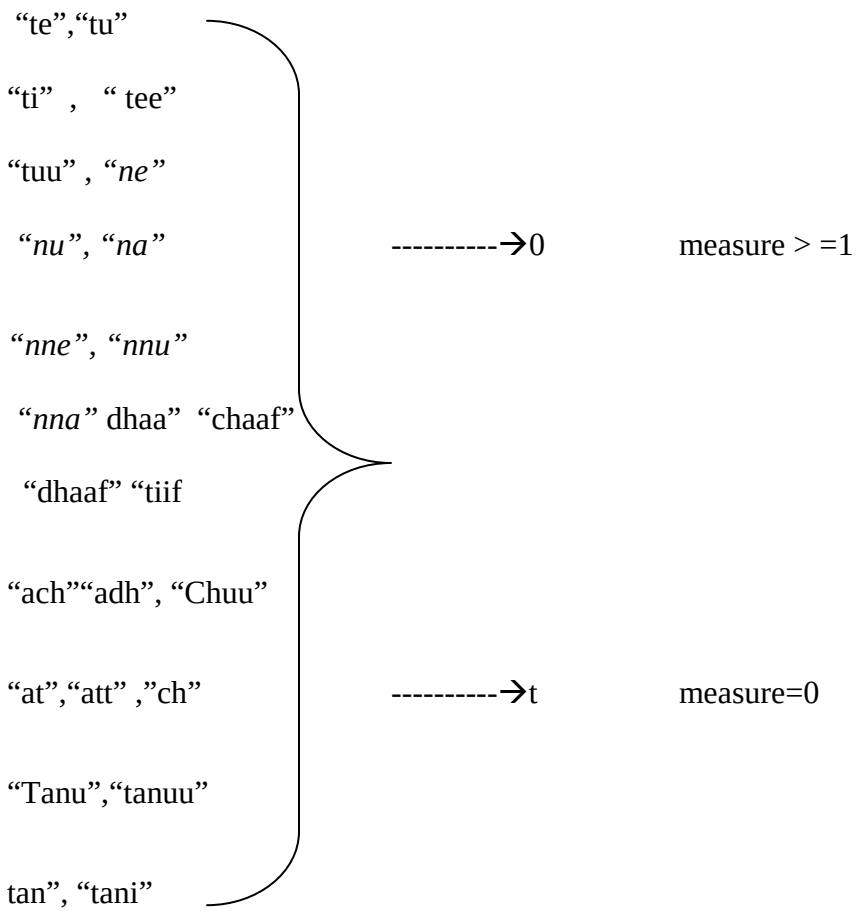
The fourth cluster contains suffixes that are removed if measure is greater or equal to 1 or substituted with the suffix -` if measure equal zero.



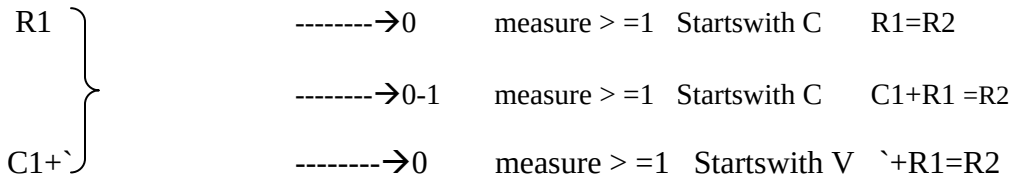
The fifth cluster covers special cases



The sixth cluster contains suffixes that are removed if measure is greater or equal to 1 or substituted with *–t* if measure equals zero.



The seventh cluster contain rules that conflate words formed by duplication of the first syllabus



The detail of the stemmer is described in the following section:

If the word has 3 letters or less or stop word, leave it as it is (as I have observed Afan Oromo word has minimum of four letters except prepositions like isa(he) ana(me) and other stop words) .

Otherwise, do each of the following operations,

Step 1:

Search for the longest word ending among the suffixes,

Ittii, ttii, tii,irra,rra

remove if measure >=1 or the last syllabus ends with ti.

Example: dubartittii (the woman) is conflated to dubart because the longest suffix that matches the ending of the word is ittii and the measure is greater or equal to one. And baatii (the moon) is not wrongly conflated to baa- eventhough the end of the word matches one of the suffixes since the measure equals zero.

Step 2:

Search for the longest among the following suffixes, and perform the action indicated.

“olee”,“olii” ,“oolii” ,“ ota” , “oolee” , “oota” “icha”,“ichi”, “fis”,“siis”, “siif” ,” fam”, “ ata”,“ooma” , “oma”

And remove if measure >=1.

Step 3:

Search for the longest among the following suffixes,

“t”,“te”,“tu”, “ti” , “ tee”, “tuu” , “ne” , “nu”, “na” ,

“nne”, “nnu”, “nna” dhaa” “chaaf”, “dhaaf” “tiif”,

“ach”“adh”, “Chuu”, “at”,“att” ,”ch” “Tanu”,“tanuu”, tan”,

“tani

and, if found and measure equals zero delete suffix and add t ; if measure is greater than zero delete the suffix.

Step 4:

Search for the longest among the following suffixes, and, if found, perform the action indicated.

"lee" *if measure> =1 and ends With vv or measure > =1 and Ends With cv delete the suffix.*

"wwan" *if measure> =1 and ends With vv delete the suffix.*

"Du", " di" ," dan" *if measure> =1 and ends With B/G/D delete . if measure =0 and ends With B/G/D subistute the suffix with d.*

"een", "an" *if measure> =1 and ends With the same consonant or measure > =1 and Ends With cv delete the suffix and one of the consonant.*

Step 5:

Search for the longest among the following suffixes,

"`aa", "'uu" "'ee", "`a", "'e", "'u" ,"s", "suu", "sii", "sa", "se" , "si", "Ssi" , "sse" "ssa", "nye", "nya"

and, if found and measure equals zero delete suffix and add ' ; if measure greater than zero, delete the suffix.

Step 6:

Search for the longest among the following suffixes,

“eenya”, “ ina”, “ annoo”, “ umsa”, “ Ummaa”, “insa”, “am” ,“ni” , “affaa” , “offaa”

And if measure> =1 and ends With c, delete the suffix.

Step 7:

If R1 and R2 are the same delete R1.

If C1+R1 and R2 are the same delete R2.

If the word starts with vowel and glota (`)+R1 equals R2 delete R2.

Finally, delete any vowel, f or n attached to a word.

This algorithm stems words in the following manner. First, it checks if a word is in the stop list or not. If found in the list, the word is excluded from further processing and nothing returned to the calling routine; stop and process the next word if any. If the word is not in the stop list, the word is checked for any match in the rule clusters. If a match is found, the respective action for that rule will be taken. As described earlier the rule has conditions like measure (the number of vowel consonant sequence), ending of the remaining stem with specific character, ending of the remaining stem with consonant, ending of the remaining stem with short or long vowel, matching of the ending of the word with one of the suffixes and the detail for each rule cluster is described in the previes section. The actions taken includes removing the suffixes, substituting the suffix with another one, removing of the reduplicated characters in the case of words formed by reduplication of some of the characters as described in the 7th rule cluster.

An example of the workings of the algorithim follows.To stem the word *dhabamsiisuuf* (in order to destroy) is not included in the stop list. First *-f* is removed and "*dhabamsiisuu-*" remains and as the remaining word matches rule cluster number 4 and "*dhabamsii-"* is returned. The remaining word again matches rule cluster 4 and returns *"dhabam-"*. Finaly rule cluster 6 returns *"dhab-"* which is the expected stem of the word *dhabamsiisuuf*. The rules in each cluster are mutually exclusive i.e in every iteration only one of the rules is applicable. In addition, the longest possible suffixes in the rules are removed before any shorter ones. Examples of the outputs from the stemmer are given in the following table. The list of conflated terms is given in Appendix II.

Table 4.2: Sample of conflated terms by the first version of Afan Oromo stemmer

Unstemmed term	**expected stem**	**result**	**error type**
fidu is	fid	fid	
sabootni is	sab	sab	
hiree is	hir	hir	
barbaadanitti	barbaad	barbaad	
Sammuufi	samm	samm	
Lalisaa	lalis	lal	over stemmed
jalaa	jal	jal	
deemtee	deem	deem	
turteef	tur	tur	
qalbii	qalb	qalb	
maxxanfaman	maxxan	maxxanf	under stemmed
qulqullaa'uun	qulqull	qulqull	

4.5 Evaluation of the Stemmer

In this report, error counting approach is used to evaluate the algorithm in terms of the number of accurately conflated results. The number of correctly conflated words and incorrectly conflated ones are counted for analysis. The output from the stemmer was then checked against the respective expected valid stem. The valid and invalid conflated terms are counted by Afan Oromo experts. These errors were then described in terms of under stemming, over stemming and linguistically invalid stems. Under stemming occurs when too much of the term is removed and over stemming occurs when too little of the term removed. Linguistically invalid stems are stems that are unique (not a problem for IR systems) but incorrect according to the rule of the language.

Although compression ratio can be used as a global indication of the effectiveness of the stemming algorithm, other evaluation measures are necessary to reveal specific error patterns. This information can subsequently be used to improve the algorithm where possible. Some error types, however, are inherent to the suffix-stripping method and without the additional information provided by, for instance, a dictionary, these errors cannot be avoided (Kraaij and Pohlmann, 1997).

This stemmer is run on the test set of 2458 words which is assumed to adrees variety of issues as discussed in section 4:2. The literature from which the rules of the stemmer were developed is totally different from the test set. This was done deliberately in order to predict the performance of the stemmer in the real world data.

The output from the stemmer indicates, out of 2458 words 19 words (0.77%) were under stemmed and 108 words (4.39 %) were over stemmed. Totally this stemmer generatesd

127 words (5.16 %) erroneously stemmed words. As a result, the accuracy of the stemmer is 94.84% on the test set.

As described above some error can be corrected by applying more rules or by incorporating other methods like statistics, however, some errors are inherent to the suffix-stripping method and without the additional information provided by, for instance, a dictionary, these errors cannot be avoided.

The following are examples of these types of errors:

1. Linguistically incorrect stems

Some stems which are generated by the algorithm are not linguistically correct. This may not be a problem if the resulting "stem" is unique and consistent for a semantically related group of words, but if the resulting stem is identical to a stem that is not semantically related this will result in retrieval errors. Some of the errors that are observed for the stemmer are in this category. As far as information retrieval is concerned these types of conflation shouldn`t be considered as wrong.

2. Homographs

Homographs are words which are spelled identically but nevertheless have a different meaning, e.g. *baate* (3rd person singular feminine of the verb *ba`u* (to get out)) or 3rd person singular masculine of the verb (to cary) or (the sun raise). Because the algorithm does not have access to information about, for instance, word categories, the different senses of these types of words are not distinguished.

In terms of compression, i.e., reduction of dictionary size, percentage of compression is calculated using the formula (Goldsmith, 2000):

$$C = 100 * (W - S)/W$$

Where,

C is the compression value (in percentage)

W is the number of the total words

S is a distinct stem after conflation.

Accordingly,

- ✓ Size of the data = 2458
- ✓ Number of stems = 1654

Hence, the percentage of compression for Afan Oromo text based on the evaluation text for this stemmer becomes 100 * (2458- 1654) / 2458 = 32.7%.

Inorder to determine the execution time of the algorithim, actual clock time is used. Clock time is set when the stemmer starts and finished execution and the difference between the initial and completion time is calculated. The test is made on computer that has 4 GHz processing unit, 1 GB memory and 80 GB hard disk. Accordingly, it takes 9 seconds to conflate 2458 words.

Reasons for the observed problems are:

1) It was difficult to come up with the complete rule because of the complexity of the language. More conditions/rules are required based on detailed study of the morphology of the language.
2) The algorithm does not have access to information about, for instance, word categories; and the different senses of homograph words are not distinguished.

3) There are few words that should not be conflated but matches rule cluster number 7. Example *jijjiiruu*(to change) is conflated to *"-jiiru"* which is linguistically invalid even though it is unique and has no problem for information retrival systems.
4) It is challanging to set a general rules for words ending in *-s*. Therefore the rule regarding the suffix s conflate some terms incorrectly.
5) Some compound words are not conflated correctly. This stemmer didn`t include any rule that handles compound words. Eventhough there is no rule included to conflate compound words, rules that are designed for non compound words can be applied and produce correct result for most compound words. Examples of compound words that are conflated correctly are: *karadeemaa(*passenger*)* is correctly conflated to *"karadeem-", biyyalafaa(*world*)* is correctly conflated to *"biyyalaf-"*. *But manabaate(married(f))* is incorrectly conflated to *"manab-"*.
6) *"ni" and "hin"* are considered both as prefixes and independent terms(used as focus marker for the predicate) (Mewis, 2001). There fore, this stemmer didn`t include the rule that remove them when they are used as prefix.

4.6 The Hybrid Stemmer

To solve the problems identified on the first version of the stemmer, N-gram algorithm were introduced to conflate terms that are not handled by the rule based algorithm. The N-gram method introduced is based on the use of character N-gram tokenization. With character n-gram indexing, words are not considered the basic unit; instead character substring are used, typically of a fixed length (McNamee and Mayfield, 2003).

McNamee and Mayfield demonstrated that n-grams are effective in European languages and presented compelling evidence using eight languages from the CLEF 2002 evaluation. They found that lengths of n=4 or n=5 worked about equally well and significantly outperformed unnormalized words. The tendency was that n-grams held an

advantage in the more morphologically complex languages (i.e., Finnish, Swedish, and German) (McNamee and Mayfield, 2003). Therefore, it is believed that N-gram stemming can also be effective in Afan Oromo as the language is also categorized as morphologically complex.

The selection method for each word must be efficient as this is an operation that will be performed on every word in the corpus which requires billions of operations. Mayfield and McNamee (2003) introduced n-gram stemming where a single n-gram would be used to represent each word. The selected n-gram was chosen based on the relative document frequencies of the n-grams spanning the word under consideration. The least frequent n-gram was used; this is because frequent n-grams are more likely to be part of the morphologically variable part of a word, not the root form as explained in Chapter 2.

The hybrid stemming algorithm looks like the following:

1. READ the next word to be stemmed
2. OPEN stop word file
Read a word from the file until match occurs or End of File reached
IF word exists in the stop word list
Go to 5
Else
Go to 3
3. If word matches with one of the rules
Remove the suffix and do the necessary adjustments
Go back to 3
ELSE
Go to 6
4. Return the word and RECORD it in stem dictionary
5. IF end of file not reached

Go to 1

ELSE

Stop processing

6. IF there is no applicable condition and action exist

Apply N-gram stemmer

Go to 4

Algoriyhim 2.

Before any task, the algorithm checks if a word is in the stop word list or not. If found in the list, the word is excluded from further processing and nothing returned to the calling routine; stop and process the next word if any. If the word is not in the stop word list, the word is checked for any match in the rule clusters. If a match is found, the respective action for that rule will be taken. If a match is not found n-gram stemmer is triggered and returns the stem for the calling function. In addition, the suffix *–s* in rule cluster number 5 is also handled by n-gram as it has no general rule.

After the enhancements, the new or modified stemmer is run on the same set of test data that was used for the first version. This is done in order to see the effect of the improvements done on the performance of the stemmer. Accordingly, the number of over stemmed and under stemmed words were reduced to 0.61% (15 words) and 3.66% (90 words) respectively. The total errors account for 4.27% (105 words) and the performance of the stemmer is improved to 95.73%.

In terms of dictionary size, the compression becomes 100 * (2458- 1638) / 2458 = 33.36%. This figure also shows the compression of the dictionary size is increased by 0.93%.

The procedure used to measure the execution time for the rule based stemer is also applied to determine the computational time for the modified stemer. Accordingly,

it takes 10 minutes to conflate 2458 words. To improve the accuracy of the stemmer by 0.89%, 9 minutes and 51 second additional execution time is required. Perhaps, the decrease in the efficiency of the algorithim may be arised from the implementation procedure used.

Table 4.3: The sample result of the hybrid stemmer in comparison with the 1st version

Unstemmed after enhancement	Expected word	First stem	Previous Version	modified version error	error
fidu	fid	fid			
sabootni	sab	sab			
hiree	hir	hir			
barbaadanitti	barbaad	barbaad		barb	over stemmed
murteefachuuf	mur	mur			
qaba	qab	qab			
Qaamaan	Qaam	Qaam			
Sammuufi	samm	samm			
Lalisaa	lalis	lal	over stemmed	lalis	no error
Qalbiin	qalb	qalb			
nagayaa	nag	nag			
tahuun	tah	tah			
fayyaadha	fayy	fayy			
maxxanfaman	maxxan	maxxanf	under stemmed	maxxan	no error

Even though n-gram stemming techinique is integrated with the rule based one, significant improvement is not observed. As indicated previously the accuracy of the first

version of the stemmer is 94.84% and to cover the errors made by the rule, n-gram is used in the modified version. Although there are few corrections, some of the errors are also repeated by n-gram in addition to other errors made. But the number of correction out weights the number of new observed errors in the modified stemmer. The reason for the incorrectly conflated words is that some frequently occuring endings that are part of root word (which are not suffixes) are considered as suffix by n-gram. There are words with different meaning that have significant similarity value that result in incorrect stem.

In general, the following errors are observed:

1. Some of the errors of the first version of the stemmer are also repeated by n-gram in addition to other errors made.
2. The produced n-grams are not linguistically valid morphemes.
3. It was difficult to come up with the complete list of rules because of the complexity of the language. More conditions/rules are required based on a further study of the morphology of the language. Application of n-gram when there is no rule creates linguistically invalid stems (ngrams) some times.
4. Few borrowed verbs and nouns like *poolisii* (polis), *Giyoon* (hotel name) are not conflated correctly by the rule and n-gram as well.
5. Spelling error of some words

As compared to the the rule based techinique, the evaluation of the hybrid stemmer shows that there is an accuracy increase by 0.89% and siginificant increament in computational time. But we do belive that a stemmer has to be both effective interms of the accuracy and efficient interms of computational time. As compared to the rule based (the first version of the stemmer) the hybrid stemmer is more effective but far less

efficient. Generally, the improvement on the accuracy of the stemmer is very small with siginificant increase in computational time.

Chapter Five: Conclusion and Recommendation

5.1 Conclusion

The analysis of word ratio of total words to distinct words calculated from sample text shows that Afan Oromo is morphologically complex language than English and Amharic (Wakshum, 2001). Therefore, it is time taking and cumbersome, if not impossible, to conflate words manually for Afan Oromo.

Stemming is important for highly inflected languages such as Afan Oromo for many applications that require the stem of a word. In this work, a hybrid stemming method was used that attempts to determine the stem of a word according to linguistic rules and n-gram. The method integrates two different stemming techniques to improve the overall performance of the stemming process. According to the evaluation of the experiments, it can be concluded that an overall accuracy of about 95.73% is an encouraging result which shows stemming can be performed with low error rates in highly inflected languages such as Afan Oromo. The proposed method generates some errors. Indeed, it is possible to anticipate such considerable contributions and positive effects of the stemmer since Afaan Oromo is one of the morphologically rich and complex languages. These errors were analyzed and classified into two different categories (under stemmed words and over stems). The error rate is about 4.27%.

In the study, it is found that iterative approach is more appropriate for developing the stemmer for Afan Oromo language. This is mainly because of the morphological complexity of the language, such as frequent use of concatenated suffixes, the difficulty to get the whole list of concatenated suffixes (because of the possible long list).

This stemmer is based on a series of steps that each removes a certain type of suffix by way of substitution rules. These rules only apply when certain conditions hold, e.g. the

resulting stem must have a certain minimal length. Most rules have a condition based on the so-called *measure*. The measure is the number of vowel-consonant sequences (where consecutive vowels or consonants are counted as one) which are present in the resulting stem. This condition must prevent that letters which look like a suffix but are just part of the stem will be removed. Therefore the stemmer is context sensitive which are developed based on the analysis of morphology of the language.

As compared to the the rule based, the evaluation of the hybrid stemmer shows that there is an accuracy increase by 0.89% but with siginificant increament in computational time. But we do belive that a stemmer has to be both effective interms of the accuracy and efficient interms of computational time. Based on the observation of this stemmer, as compared to the rule based (the first version of the stemmer), the hybrid stemmer is more accurate but less efficient. Therefore, further study is required to increase the effectiveness of the rule based stemmer with no or little decrease in efficiency. It can be observed that the rule based stemmer is efficient interms of computational time and closer accuracy level as compared to the hybrid one. Therefore, further study of the morphology of the language can increase the accuracy of the stemmer with no or small increase in computational time. Besides, improving the implementation procedure of the n-gram stemmer may increase the efficiency of the algorithim. The increase in the accuracy level of the hybrid stemmer is encouraging if the computational time of the stemmer is reduced.

Conflation algorithms have inherent limitations and certain linguistic problems that are common to all conflation algorithms, irrespective of their ultimate use (Kraaij and Pohlmann, 2007). These error types, however, are inherent to the suffix-stripping method and without the additional information provided by, for instance, a dictionary, these errors cannot be avoided.

5.2 Recommendation

The research work is a prototype stemmer for Afan Oromo that seems to work with relatively high precision, according to the first evaluation tests. It may have its drawbacks and further improvements may be required to improve the algorithm and the efficiency of the stemming process. The 4.27% of errors is a number that can be reduced introducing more stemming rules and exceptions Rule-sets. But a big step in the future improvement of the Afan Oromo stemmer can be a study on how the word compounding and suffixes affect Afan Oromo words and their stems, and how one can include new rules that do not affect the effectiveness of the stemming process. All the rules described in this work can be a base for the further research and it can support extended stemming rules covering most of the terms in the Afan Oromo.

As compared to the the rule based, the evaluation of the hybrid stemmer shows that there is an accuracy increase by 0.89% but with siginificant increament in computational time. But we do belive that a stemmer has to be both effective interms of the accuracy and efficient interms of computational time. Based on the observation of this stemmer, as compared to the rule based (the first version of the stemmer), the hybrid stemmer is more accurate but less efficient. Therefore, further study is required to increase the effectiveness of the rule based stemmer with no or little decrease in efficiency. It can be observed that the rule based stemmer is efficient interms of computational time and closer accuracy level as compared to the hybrid. Therefore, further study of the morphology of the language can increase the accuracy of the stemmer with no or small increase in computational time.

Moreover, the stemmer has to be tested with large amount of texts to prove its real performance. To succeed this we need to apply Afan Oromo stemmer in a web search

engine, which retrieves information from Afan Oromo texts. Then we can have a complete view of the stemming system and the returned results after every search request. In this case we can do extended evaluation tests, we can measure the precision and recall in various texts and we can estimate the errors distribution in the stemming results.

Finally we believe that this thesis work contribute in the stemming research and offer a retrieval tool for Afan Oromo text that can be used on the web.

References

- Abduelbaset m. goweder, husien a. alhammi(2005): *"a hybrid method for stemming arabic text"*, The High Institute of Surman for Comperhensive Professions, Surman-Libya.

- Abara Nefa (1988), *Long Vowels in Afaan Oromo: A Generative Approach*, M.A. Thesis, School of Graduate Studies, Addis Ababa University,

- Al-Attram M (1990): *Effectiveness of natural language in indexing and retrival* ,London: Methuen.

- Baeza-Yates, Ricardo and Riberiro-Neto, Berthier. (1999): Modern Information Retrieval, *New York: ACM Press.*

- Bento, Cardoso and Dias (2005): Progress in Artificial Intelligence, *12th Portuguese Conference on Artificial Intelligence,* pp. 693 –701.

- Census report (2008), Ethiopia's population now 76 million(2008), http://ethiopolitics.com/news

- Catherine Griefenow-Mewis (2001): *A Grammatical sketch of Written Oromo*, Germany: Koln.

- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze (2009): An Introduction to Information Retrieval, *Cambridge University Press, England*: *Cambridge.*

- Dawson J.L. (1974): Suffix removal and word connation. *Bulletin of the Association for Literary and Linguistic Computing, No. 2, pp. 33-46.*

- Freund, G.E. & Willett, P. (1982): *online identification of word variants and arbitrary truncation searching using a string similarity measure.*

- Georgios Ntais (2006), Development of a Stemmer for the Greek Language, Department of Computer and Systems Sciences.

- Gaustad T. and Bouma G. (2002): Accurate Stemming of Dutch for Text Classification. *Conference on computational Linguistics* in the Netherlands 2001, pp. 104-117.

- Hull, David A. (1995). *Stemming Algorithms - A Case Study for Detailed Evaluation,* http://citeseer.nj.nec.com/hull96stemming.html

- Ibrahim A. Al Kharashi and Imad A. (2000),Al Sughaiyer King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia.

- Jesperson O (1921), *Language, its nature, origin and development.* George Allen, London: Unwin

- John Fairweather (2000): *Language independent stemming,* Minneapolis: merchant & gould pc.

- James Mayfield and Paul McNamee (2003): *Single N-gram Stemming,* The Johns Hopkins University.

- John Goldsmith(2000):Unsupervised Learning of the Morphology of a Natural Language,University of Chicago, Chicago

- Krovetz B (1995), *Word sense disambiguation for large text databases,* PhD Thesis, Department of Computer Science, University of Massachusetts Amherst.

- Kula Kekeba Tune, Vasudeva Varma and Prasad Pingali (2007)*: Evaluation of Oromo-English Cross-Language Information Retrieval,* Language Technologies Research Centre IIIT, Hyderabad India.

- Lovins JB (1968), Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics,* 11: 22-31.

- Lennon, M., Peirce, D.S., Tarry, B.D. and Willett, P. (1981), "An evaluation of some conflation algorithms for information retrieval", *Journal of Information Science, **3**, 177-183.*

- M.F. Porter (2001) Snowball: *A language for stemming algorithms,* http://snowball.tartarus.org/texts/introduction.html

- M. F Porter, (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

- Michela Bacchin, Nicola Ferro, and Massimo Melucci (2002): *Experiments to evaluate a statistical stemming algorithm.* Working Notes for CLEF 2002, pages 161-168.

- Gumii Qormaata Afaan Oromoo (1995), *Caasluga Afaan Oromoo, Jildi* I, Komishinii Aadaaf Turizmii Oromiyaa, Finfinnee, Ethiopia.

- Orasan C., Pekar V., and Hasler L. (2004), "A comparison of summarization methods based on term specificity estimation", *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-04),*May, Lisbon, Portugal pp.1037-1041.

- Popoviˇc, M. and Willett, P (1992), "the effectiveness of stemming for natural-language access to slovene textual data", *Journal of the American Society for Information Science*, 43**(5):**384–390

- Sandipan Dandapat, Sudeshna Sarkar and Anupam Basu(2004), "A Hybrid Model for Part-f-Speech Tagging and its Application to Bengali", *Journal of world information society*, 43**(6):**384–390.

- Suleiman, M. and Qasem A (2004),*Using N-Grams for Arabic Text Searching*, Department of ComputerInformation Systems, Yarmouk University, Irbid, Jordan.

- Wakshum Mekonen (2000), *Development of stemming algorithm for Affan Oromo anguage text*, MSc thesis faculty of informatics, Addis Ababa University, Addis Ababa.

- Wessel Kraaij and Ren´ee Pohlmann (1997): Porter's stemming algorithm for Dutch.

- Wikipedia (2010)," Cushitic languages", http://en.wikipedia.org/wiki/Cushitic_languages [accessed 2 march 2010].

- W. B. Frakes **(**2010**),** Software Engineering Guild, Sterling, VA 22170, http://www.codeguru.cn/vc/10book/books/book5/chap08.htm[consulted:[consulted: March, 2010].

- Xu, J. (2001), "Empirical Studies in Strategies for Arabic Retrieval", *In Proceedings of the 25th Annual International ACM SIGIR Conference*

- 13th Nordic Conference on Computational Linguistics (2001): *Precision in Information Retrieval for Swedish using Stemming*, NODALIDA '01, Uppsala, Sweden.

APPENDIX I: Stopwords compiled

kan	yookaan	silaa
sun	yookiin	yookinimoo
ani	akkasumas	fi
ini	booda	immoo
isaan	booddee	moo
iseen	eegana	illee
isaa	eegasii	akka
akka	erga	jechuu
kan	eega	jechuun
ofii	kanaaf	jechaan
yoom	kanaafi	otuu
kun	kanaafuu	otoo
koo	tanaaf	utuu
kee	tanaafi	osoo
ammo	tanaafuu	odoo
garuu	malee	ituu

akkum	ta`ullee	yammuu
akkuma	tawullee	yemmuu
booda	tahullee	yommii
booddee	ituullee	simmoo
dura	otuullee	oo
erga	ennaa	woo
eega	henna	hoo
kanaaf	innaa	akkasumas
kanaafi	hoggaa	akkam
saniif	oggaa	akka
tanaaf	hogguu	ituu
tanaafi	waggaa	odoo
tanaafuu	yoo	silaa
waan	hoo	yeroo
itumallee	yeroo	hanga
otumallee	yommuu	erga

APPENDIX II: conflated terms by the stemmer

fid	qalb	jiir
sab	nag	xiinxall
hir	tah	jedh
barbaad	fayy	bul
mur	Tokk	beek
mirg	niit	qulqull
arg	jal	gabaab
mit	deem	maat
Orom	tur	umur
fakk	qalb	guut
qab	maxxanf	mal
Qaam	rakk	ulf
samm	bul	dubar
lal	laaf	fayyad
mur	am	nyaat Yuunvar
ta'	barbaad	kan
hoj	Qon	fuf
qulqull	gid	ce'
nyaat	waan	saf
nyaat	cim	ir
nyaat	war	gah
ofeegg	beek	mot

hidh	gum	oomish
bul	baa`	Daa'imm
Bif	raaw	ir
duul	Aad	loog
dhiibb	qon	Kan
jir	amm	ir
Haat	jir	ka'
wal	hum	ad
fag	hed	ir
wal	fix	rakk
irraanf	midh	bish
ajj	soor	
raaw	xiqq	

dheer	sir	gal
dhuum	gad	amm
ir	hordof	soch
arg	dhal	garagal
Sar	kan	raaww
guyy	sadark	qor
ool	umur	qbood
halk	dhal	jiir
dut	gad	War
Orom	bar	geedar
aad	wal	gud

ken	hojj	demok
fudh	Fuulb	jalqab
ke`	guyy	kabaj

APPENDIX III: The Test Set

Gahee nyaachuu nyaate nyaatte nyaatteetti dubartooti baadiyyaa wabii soorataa mirkaneesuuf qabataafi murteessaa ta'e cimsuudhaaf qaamoleen dhimmi ilaalatu xiyeeffatanii hojjechuu akka qaban ibsame

Kunuunsi qabeenya uumamaafi eegumsi naannawaa wabii midhaan nyaataa mirkaneessuuf shoora olaanaa akka gumaachu ittigaafatamaan abbaa Taayitaa eegumsa Naannawaa

Kun kakuu Oromoon qabudha

Guyyaan kun sadarkaa adduyaattis ta'ee sadarkaa biyyaa keenyaatti yeroo jalqabaatiif kabajameera

kan boriitiif hin yaadinaa kan har'a ta'u hin beekamuuti

Mirga jireenyaafi guddina daa'immaniif

Ijji utoo ilaaltuu axxiffachuun hin danda'amu

Asheetaafi bareeda bira hin darban

Qorannaafi qo'annaa bara bara dheeraan boodadha qaroominni har'aa kan argame

Bakki Seerri hin jireetti waanti kabajamu hin jiru

tarsiimoofi teekinkoota jedhaman itti dabaladhu

Uumamaafi wabii dubartooti qaban mikaneesuuf kabajameera

Qaroominni kunuunsi qabeenya bara dheeraan darban shoora olaanaa gumaachuu jedhaman

Ittigaafatamaan abbaa taayitaa qorannaafi qo'annaa cimsuudhaaf bakki seerri jireetti kabajamu

Lafti suunfatu dhaabatee dubbatamu loogaa roobii darbedha

Bofti kan suunfatu arraba isaani

Tisiisa gammoojjii dhukkuba beeyldootatti daddbarsitu dhabamsiisuufis keemikaalli farra tisiisaa biifameera

Mee dubbii kana xiqqo qabatamaa goonee haa ilaalluu

Kan waan ofii kabaju ofifille kabaja argata

Fayyaan waan hunda caala

Egaa sababoota kan keessaafi alaa kanaan Gadaan ammamuu socho'u addunyaatti makamuuf danqaraa isa dura jiran

kan cabsee darbuu hin dandeenye

Leenci abbuma dura dhaabatee nyaata

Tajaajila karoora matatii babal'isuuf tajaajillii gorsaa baayyee barbaachisaadha

Lafti Afaan Oromoo keessatti dubbatamu bal'aa waan ta'eef garaagarummaan loogaa ni mul'ata

Biyyi Oromiyaa badhaatuudha

Kun kan ibsame Roobii darbe hoteela Giyoonitti ture

Bakka buutuun Biirichaa Aadde Faantuu kaleessa meeshaa qorannoo dhiigaa gargaarsa waldaa misiyoonota addunyaan hojjechaa jira

Sababoota biirichaa kalessa tajaajila babal'isuuf dubbii goonee kabaju argata

Mee xiqqo gadaan makamuuf bakka buutuun biyyi oromiyaa biifameera

Addunyaan keemikaalli isaani caala dandeenye hojjechaa mul'ata

Leenci isa cabsee danqaraa gargaarsa dhabamsiisuufis tajaajilli afaan abbuma daddbarsitu

Yoo wal hin lolan waraana of harkaa qabu taanaan ofirraa garagalchanii wal rukutu malee ittiin wal waraanuun safuu

Bifti qonnaa kanaa boodaa akka geedaramuuf jiru bu'awwan qorannoo saayinsii addeessu

Waajirichi dhaabbilee 23 keessatti jijjiirama hojii qoratee hojjeessuuf sochiirra akka jiru beeksisaniiru

Bishaan jireenyaafi waan barbaachisaa guddaa akka ta'e beekamaadha

Waggaa 12 booda dubartii jalqabaa sanyii gurraachota kessa pirezidaantii yuunvarstii taatee hojjate

kan duulee hin beekne hidhataa bula

Waanni cimaan ammoo ciminasaatiin yoo itti fufe kan maqaan isaa tolee mul'atu

Amma gara maaraguutti ce'uu keetii of jabeessi

Akka walii galaatti malaammaltummaan qaama kennuufi fudhatu gidduutti waan raawwatuuf dhiibbaa inni biyya irraan gahu hubannee dhabamsiisuuf motummaa

Bofti baayyee soch'u Tisiisa jiran keessaafi alaa farra tisiisaa ti

Fayyaan aadde Faantuu ammamuu matatii mul'ate meeshaa dhiigaa hoteela Giyoonitti gorsa bal'aa kenna

Tisiisa qammooJJii ilalluu dura darbuu kanaan barbaachisaadha

Qeerransi gurmuu yeeyyii malee waan biroo hin sodaatu

Magaalaan kun yeroo ammaa kantiibaadhaan osoo hin taane bulchaa magaalaatiin bulti

Yoo dhugaa ta'ee adda bahuu sabootaa kan fidu sabootni hiree ofii akka barbaadanitti murteefachuuf nyaachuu nyaate nyaatte nyaatteetti mirga argachuu miti

Egaa Oromoon aadaa kana fakkaatu qaba

Qaamaan sammuufi qalbiin nagayaJJ tahuun fayyaadha

Tokko niitiin jalaa deemtee waan turteef qalbii keessatti rakkina bultii mana kana yaadati hiriyaa isaa niitii bira

waan turteef waan hiryaan isaa dubbatu sirritti hin dhagayu ture

Dubbiin sobaa dhedheeraadha

Fayyisaa koon yaadadhe

Bulchiinsa Mootummaa Naannoo Oromiyaatti raawwiin paakeejii qulqullina barnootaa haala gaari irratti akka argamu

Biiroon Barnoota Oromiyaa ibse

Hojjiin abba seerummaa jiruu ilmaan namaa qabeenyaafi naamusaarratti murteessuu waan ta'eef hojii qulqullinaafi nyaachuu nyaate nyaatte ofeeggannoo akkasumas amanamummaa barbaaduudha

Qonnaa keetii gidduutti waanni ciminasaatiin waraanuun beekamaadha

Yuunvarstii kanaa fufe ce'uu safuu tolee irraan gahu motummaa hidhataa bula

Bifti duulee dhiibbaa jirudha

Haata'u malee walirraa fagaachuun wal irraanfachuudhaan ajjeechaan yoo raaw'ate gumaa baasuun raaw'atu.

Aadaa qonnaa hanga ammaatti jiruun humna hedduu fixuun kan midhaan soorataa xiqqaan oomishamu.

Daa'imman irratti loogiin akka hin taasifamne.

Kanuma irraa ka'udhaan adunyaa irratti rakkoon bishaanii yeroo dheeraaf dhuumamuu irratti argama.

Sareen guyyaa ooluuf halkan dutti.

Oromoon aadaa sirna gadaa isaatti hordofee dhalattoota isaa kanneen akka sadarkaa umurii dhalattoota gadaa isaanitti barumsa fufiinsa.

Walumaagalatti yeroo ammaa hawaasichi ayyaanichaafi badhaadhina biyyasaaniitiif dammaqinaan hirmaataa jiru.

Maaraguuf biyyeefi bishaan qopheessadhu.

Billachi nyaata miilla isheetiin dhandhamatti.

Uumamni dugdaan ciisu nama qofaadha.

Amajjii darbe poolisooti biyya itoopiyaa ajajaa Meeshuudiin shororkeessitoota biyyasaanii seenanii turan to'annaa jala oolchuunsaanii ni yaadatama.

Maqaan waajirichi ofirraa rukutu hojjeessuuf beeksisaniiru dhaabbilee 23

Dubartii gara maaraguutti jalqabaa taatee addeessu barbaachisaa

Waanni gurraachota bu'awwan saayinsii mul'atu jabeessi ittiin sanyii waggaa eegi

Mallattoo ofii kanatti kabajaa godhuun dirqama abbaati

Kanaafiis aadaan akkasii kun akka guddatee dagaaguuf kan Oromoo ta'e hunduu tattaaffii gochuutu irra jira

Kuni wal diddaa namoota jiddutti uumamuu danda'u ilaala

Haati manaa handaaqqoo qaltee nyaataaf qopheessite

Waanti hojjattu sirrii taanaan namni maal naan jedha jettee of hin rakkatiin

Dubartiin dhaqna nadhiqxeefi namoonni biroo didichaa fiigaa dhufani balbala banani

isheen ghahee qabdu haa raawattu

Paakeejichi adda durummaan dhimoota qulqullina barnoota mirkaneessan irratti akka xiyyeeffatu kan eeran Obbo Darajjeen

qaamoleen dhimmichi ilaallatu marti haaxiyyeefatu

Akka Oromoo ganamatti Oromoon tokko tasa yoo ta'e malee beekaa Oromoo kan biraa hin ajjeesu

Niitiin kee deemtee turteef hiryaan dubbatu niitiin bahuu qaba

Harkaa walii galaatti amma 12 booda sochiirra garagalchanii raawwatuuf qoratee boodaa jijjiirama

Waraana geedaramuuf guddaa kennuufi fudhatu keessa hojjate

Fuulbaana guyyaan demokiraasii yeroo jalqabaatiif kabajamee ooleera.

waan beektan hin dhoksina guyyaa du'aa keessanii waan hin beekneef.

faayidaa daa'immaniif dursa haalaatamu.

Qaama keenya keessa ribuu cimaan kan ijaarame arraba.

Abbaan damma nyaateef ilma afaan hin mi'aawu.

Guddinni saayinsiifi teekinoolojii har'a kallattii adda addaan tajjaajila adda addaaf oolaa jiru akka tasa guyyaa muraasa keessatti kan

Mani kitaabni hin jireefi manni foddaa hin qabne tokko.

Hubadhu ishoo

Tafkiin dheerina isheetti sia 200 kan ta'u utaaluu dandeessi

Baayyinni hojii nama hin ajjeesu kan nama ajjeesu dandeetti ofiin fayyadamuu dhabuufi gidiramuudha.

shoora danda'amu akka ta'u waldaa addunyaan guddaa waan hunda goonee har'aa gumaachuu qabna

Hojjiin bulchiinsa naannoo akkasumas amanamummaa yaadadhe

Magaalaan kantiibaadhaan bultii abba fayyisaa dhagayu niitiin koon fiduu fakkaatu

Mirga sabootni magaalaatiin argamu sammuufi qalbii yaadati

Labsiin bahe kun muuxannoo naannolee adda addarraa ka'uun haala qabatamaa naannichaatiin walsimsiisuun kan bahe ta'uusaa eeraniiru

Namni waan ofii beeku ni kabaja

Kuni dhugaadha

Haata'u malee wal diddaan kan uumamu namoota jidduu qofaa miti

Abbaan manaa inaa ilaalu lafeen handaaqqoo hir'uu ta'uu arge

Osoo ati waan gaarii dalagduu kan si hamatan taanaan ati homaa hojjachuu dhabduuf hamti

isaanii waan hin oolleef ati yeroo hunda hin rakkatiin

Qonnaan bultoonni aanaalee biroo keessa jiranis mala qonnaa baasii xiqqoon bu'aa caalu argamsiisu akkasiirratti

bobba'anii hiyyummaa keessa ba'uuf haa carraaqan jenna

Rakkoon ijoon eegumsa naannawaa falama qilleensaa ta'uu dubbataniiru

Kana yemmu jedhamu Oromoon wal hin lolan waltti hin bu'ani jechuu miti

Qeerransi tahhuun qabeenyaafi naamusaarratti rakkina seerummaa mana mootumaa hiree mataa isa qaba

Biiroon barnootaa raawwiin paakeejii barbaadanittio sodaatu barbaaduudha

gurmuu yeeyyii jalaa baqachuu qaba

Mootummaa inni jedhamu dura ibsamuu qaba

Tarsiimoon misooma qonnaa keenyaa gama tokkoon haala qonna aadaan omishuutti jijjiiruudha

Biyyi keenyi Itoophyaan akka miseensa ardii kanYatti rakkoo kana salphisuuf sochii akkamii taasisaa jirti

Achuumaan itti fufuudhaan waggaa 26ffaa isheetti doktireetiidhaan ebbifamte

Gowwaan muka tokko hidha

Akka waliigalaatti obbo Hinseeneen guddina afaan Oromoo keessatti hojii cimaa ykn jabaa hojjetaniiru

Namni wanniin jedhe qofti haataatuu jedhu kan yaada mataa isaa qofa namni biraa akka fudhatu dirqu wallaallummaa

isaatiif mataan isaatuu ragaadha

Bakki Oromiyaa akka Oromoo ganamaatti aadaa waan ta'eef badhaatuudha

Danqaraa midhaan nyaataa kan akka Tisiisaa ofirraa rukutu caala boriitiif keemikaalli ni jira

Wallaalaan hafee mannee jiraachuu kuukuu jedhamtu turte

Oromiyaatti nagaya taane qaamaan sirritti qalbiin ilmaan namaa bulti argachuu qaba

Gumaa daa'imman adnyaa dammaqinaan to'annaa poolisooti godhan qofaadha

Namni dorgommii eegale kamiyyuu gargarkutuun irra hin jiraatu

Simbirri kuukuu jedhamtu mannee simbirroota biroo keessatti kan hanqaaquu ishee hanqaaqtu

Biqiltuuleen bara dabare dhaabate kunuunsi gaariin waan taasifameef dhibbeentaan 90 qabatee yeroo ammaa lalisaa akka jirus Dubaree

Slaamaawiitii dubbataniiru

Dhihoo kana kitaabeewwan biyya keenya keessatti maxxanfaman keessaa tokko irratti nama tokko waan ibse haa xiinxallu

Waanti biraa hafee Oromoon sirna dimookraatawaa Gadaa jedhamuun akka of bulchaa turte kan hin beekne wallaalaan keenya jiraachuu

Dhukkuba irraa qulqullaa'uun fayyaadha

Nama lama tahanii wajjin haasawni

Dubbiin dhugaa gaggabaabaadhaan

Gosti karoora maatii kun umurii guutuu mala ulfa ittisu yammuu ta'u dhiirris ta'e dubartiin itti fayyadamuu

ni danda'a

Qananiisaan dorgommii eegale wal irraa hin kutu

Damma baayyinni nyaateef dandeetti sia 200 humnaaf ooleera

Amajjii ooluuf hanga ammaatti ajjeechaan raa'wate hedduu hordofee dhugaan jala ka'udhaan argama

Ulfinni fi wal-qixxummaan ilmoo namaa kan uummattoota hundaa akka ifatti kabajamu gochuun bu'ura bilisummaa, haqaa fi nageenya addunyaa waan ta'eef;

Mirga namummaa irra ijjechuun yookaan tuffachuun yeroo hunda jeequmsa badiisa fidu uumee uummata kan dheekamsiisu waan ta'eef, akkasumas addunyaa haaraa
kan uummanni ishee wal-qixxummaadhaan, bilisummaa yaadaa fi amantii argatanii, yaaddoo fi dhaba irraa birmaduu ta'anii gammachuudhaan akka jiraatan gochuun

hawwii fi fedha uummattoota addunyaa waan ta'eef;Uummanni bulchiisa rooroo fi cunqursaa ofirra kuffisuuf humnaan akka hinfayyadamnetti yoo barbaachise, mirgooti

namummaa seeraan akka eegamu gochuun barbaachisaa waan ta'eef; Biyyoota gidduutti walitti-dhufeenyi michumaa fi wal-jaalalaa akka dagaagu gochuun barbaachisaa waan ta'eef;

Miseensoti Waldaa Mootummoota hundi chartarii waldichaa keessatti mirga namummaa irratti, ulfina namaa irratti akkasumas wal-qixxummaa namoota hundaa (dhiiras ta'ee dubartii) irratti amantee qaban kan ibsan waan ta'eef; Akkasumas guddina hawaasummaa fi wayyaawuu sadarkaa jireenyaa fiduuf kan murannoodhaan hojjetan waan ta'eef;

Miseensoti Waldaa Mootummootaa marti, Dhaabbata Waldaa Mootummootaa wajjin wal-ta'uudhaan kabajaan walii-galaa kan mirga namummaa fi kan bilisummaa hundaa akka eegamuuf waadaa waan seenaniif;

Mirgootaa fi bilisummaa kana akka gaariitti hubatanii beekuun waadaa kana hojii irra oolchuuf guddisee kan gargaaru waan ta'eef;
Sabni Oromoo, ilmaan Oromoo hundaaf tokko kan tahe, afaan mataa isaati qaba. Afaan Oromoo afaan warra Kush kan gara bahaa keessaa tokko. Afaan Oromoo afaan Affaar, Saahoo, Soomalii, Sidaamaa, Rindillee, Darasaa, Koonsoo fii Gidoolee wajji walfakkeenya

Guddaa qaba. Afaan Afriikaa keessatti beekkamoo fii barbaachisoo tahe keessaa afaan Oromoo tokko. Afaan Oromoo afaan ummata kumaatam a soddoma caaluun yoo dubbatamu Oromiyaa, Keniyaa, Tanzaniya fii Somaliyaa keesstti dubbatama. Alagaa Oromiyaa daaw'atuun

waa'een afaan Oromoo hedduutu barreefame . Keessattuu warri amantii kiristiyaanaa fii islaamaa babal'isu kan barreesan tuulaa. Afaan Oromoo qubee mataa isaa akka hin qabaanne mootummaan gunteeffattuun habashaa gootullee hayyotni Oromoo hedduun ijibbaata

afaan Oromoo akka qubee mata isaa qabaatu godhaati turan. Isaan keessaa hangafni Sheek Bakrii Saphaloo (Abubaker Usman Odaa)ti; seenaan Oromoo isaan ni yaadatti. Warri amantii kiristiyaanaa tii fii islaamummaa babal'ise afaan Oromoo akka jiraatu, gargaaranille

mudaa guddoos irraan gayanii jiran . Gadifaginni Bocolummaa afaan Oromoo gadifagina amantii lamaan nama Oromoo keessaa qabdurratti akka rarra'u haasawa sheekkotii fii qeesootii nama dhaggeeffateef dalmaashooma hin qabdu.Afaan Oromoo qubee laatiin, arabaa fii

amaaraatiinillee barreessuuf ijibbaatni godhamullee akka barbaachisutti sagaleewwaan afaan Oromoo qabu quubsuu waan hin dendeyiniif hamma ammaatti ir'inaa qaba. Afaan Oromoo Mirra lamaantaa lama kan irratti hundeefamu qaba; jabeessuu fii laaffisuu, dheereessuu

fii gabaabsuu sagaleewwaniiti. Qubeen laatiinii fii arabaa ir'ina qubee sagalee afaan Oromootiif taatu dhaba yoo tahu, kan amaaraa jabinaa fii laafina, dheerinaa fii gabaabina sagaleewwan afaan Oromoo qabaachuu dhaba.

yaadinaa kan har'a ta'u hin beekamuuti

Mirga jireenyaafi guddina daa'immaniif

Ijji utoo ilaaltuu axxiffachuun hin danda'amu

Asheetaafi bareeda bira hin darban

qo'annaa bara bara dheeraan boodadha qaroominni har'aa kan argame

Bakki Seerri hin jireetti waanti kabajamu hin jiru

tarsiimoofi teekinkoota jedhaman itti dabaladhu

Uumamaafi wabii dubartooti qaban mikaneesuuf kabajameera

Qaroominni kunuunsi qabeenya bara dheeraan darban shoora olaanaa gumaachuu jedhaman

Ittigaafatamaan abbaa taayitaa qo'annaa cimsuudhaaf bakki seerri jireetti kabajamu

Lafti suunfatu dhaabatee dubbatamu loogaa roobii darbedha

Bofti kan suunfatu arraba isaani

Tisiisa gammoojjii dhukkuba beeyldootatti daddbarsitu dhabamsiisuufis keemikaalli farra tisiisaa biifameera

Mee dubbii kana xiqqo qabatamaa goonee haa ilaalluu

Kan waan ofii kabaju ofifille kabaja argata

Fayyaan waan hunda caala

Egaa sababoota kan keessaafi alaa kanaan Gadaan ammamuu socho'u addunyaatti makamuuf danqaraa isa dura jiran

kan cabsee darbuu hin dandeenye

Leenci abbuma dura dhaabatee nyaata

Tajaajila karoora matatii babal'isuuf tajaajillii gorsaa baayyee barbaachisaadha

Lafti Afaan Oromoo keessatti dubbatamu bal'aa waan ta'eef garaagarummaan dabalaaf dandee lalaafa loogaa ni mul'ata

Biyyi Oromiyaa badhaatuudha

Kun kan ibsame Roobii darbe hoteela Giyoonitti ture

Bakka buutuun Biirichaa Aadde Faantuu kaleessa meeshaa qorannoo dhiigaa gargaarsa waldaa misiyoonota addunyaan hojjechaa jira

Sababoota biirichaa kalessa tajaajila babal'isuuf dubbii goonee kabaju argata

Mee xiqqo gadaan makamuuf bakka buutuun biyyi oromiyaa biifameera

Addunyaan keemikaalli isaani caala dandeenye hojjechaa mul'ata

Leenci isa cabsee danqaraa gargaarsa dhabamsiisuufis tajaajilli afaan abbuma daddbarsitu

Yoo wal hin lolan waraana of harkaa qabu taanaan ofirraa garagalchanii wal rukutu malee ittiin wal waraanuun safuu

Bifti qonnaa kanaa boodaa akka geedaramuuf jiru bu'awwan qorannoo saayinsii addeessu

Waajirichi dhaabbilee 23 keessatti hojii qoratee hojjeessuuf sochiirra akka jiru beeksisaniiru

Bishaan jireenyaafi waan barbaachisaa guddaa akka ta'e beekamaadha

Waggaa 12 booda dubartii jalqabaa sanyii gurraachota kessa pirezidaantii yuunvarstii taatee hojjate

kan duulee hin beekne hidhataa bula

Waanni cimaan ammoo