ADDIS ABABA UNIVERSITY

FUCULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

DEVELOPMENT OF STEMMING ALGORITHM FOR WOLAYTTA TEXT

LEMMA LESSA FEREDE

JULY, 2003

ADDIS ABABA UNIVERSITY

FUCULTY OF INFORMATICS

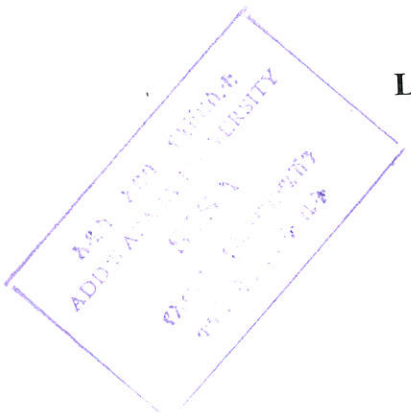
DEPARTMENT OF INFORMATION SCIENCE

DEVELOPMENT OF STEMMING ALGORITHM FOR WOLAYTTA TEXT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

LEMMA LESSA FEREDE



JULY, 2003

DECLARATION

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

Lemma Lessa

July, 2003

The thesis has been submitted for examination with our approval as University advisors.

Mesfin Getachew (Ato)	Atelach Alemu (W/t)	Haile Eyesus Engdashet(Dr)
July, 2003	July, 2003	July, 2003

Dedicated to my son,

Kibruyisfa Lemma,

who born during the first year of my study

at the School of Graduate Studies, Addis Ababa University.

ADDIS ABABA UNIVERSITY

FUCULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

DEVELOPMENT OF STEMMING ALGORITHM FOR WOLAYTTA TEXT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE

BY

LEMMA LESSA FEREDE

(Name and Signature of Members of the Examining board)

_____, Chairman, Examining board _____

Ato Mesfin Getachew, Advisor _____

W/t Atelach Alemu, Advisor _____

Dr Haile Eyesus Engdashet, Advisor _____

_____, External Examiner _____

ACKNOWLEDGEMENTS

First of all, I would like to forward my heartfelt gratitude to my research advisors: Ato Mesfin Getachew, W/t Atelach Alemu and Dr Haile Eyesus Engdashet for their crucial advice since the conception of this research work. Had it not been for their invaluable advice, I would not have surely completed the work.

My deepest gratitude also goes to Dr Nega Alemayehu (University of Sheffield, UK) for his cooperation in providing me professional comments through out the research work. I also wish to express my respect and appreciation to Dr Martin Porter (Author of Porter Stemmer) for his comments on the code I have generated.

Finally, my thanks goes to my family (esp. my wife, W/ro Serkalem Assefa), colleagues, friends like Chuba Chino and others who have helped me in one way or the other through the two hectic years of my study at School of Graduate Studies, Addis Ababa University.

June, 2003
Lemma Lessa
Addis Ababa, Ethiopia

TABLE OF CONTENTS

WOLAYTTA CONSONANTS AND THEIR SABA ALPHABET EQUVALENTS

Be - በ	Bu - ቡ	Bi - ቢ	Ba - ባ	B - ብ	Bo - ቦ
Ce - ጨ	Cu - ጩ	Ci - ጪ	Ca - ጫ	C - ጭ	Co - ጨ
De - ደ	Du - ድ	Di - ዲ	Da - ዳ	D - ዶ	Do - ዶ
Fe - ፈ	Fu - ፉ	Fi - ፊ	Fa - ፋ	E - ኤ	Fo - ፎ
Ge - ገ	Gu - ጉ	Gi - ጊ	Ga - ጋ	F - ፈ	Go - ጎ
Ha - ሀ	Hu - ሁ	Hi - ሂ	Ja - ጃ	G - ግ	Ho - ሆ
Je - ጀ	Ju - ጁ	Ji - ጂ	Ka - ካ	H - ህ	Jo - ጆ
Ke - ከ	Ku - ኩ	Ki - ኪ	La - ላ	I -ኢ	Ko -ኮ
Le - ለ	Lu - ሉ	Li - ሊ	Ma - ማ	J - ጅ	Lo - ሎ
Me - መ	Mu - ሙ	Mi - ሚ	Na - ና	K - ክ	Mo - ሞ
Ne - ነ	Nu - ኑ	Ni - ኒ	Pa - ፓ	L - ል	No - ኖ
Pe - ፐ	Pu - ፐ	Pi - ፒ	Qa - ቃ	ም - M	Po - ፐ
Qe - ቀ	Qu - ቁ	Qi - ቂ	Ra -ራ	N - ን	Qo - ቆ
Re - ረ	Ru -ሩ	Ri -ሪ	Sa - ሳ	O - ኦ	Ro - ሮ
Se - ሰ	Su - ሱ	Si - ሲ	Ta - ታ	P - ፕ	So - ሶ
Te - ተ	Tu - ቱ	Ti - ቲ	Va - ቫ	Q - ቅ	To -ቶ
Ve - ቨ	Vu - ቩ	Vi - ቪ	Wa -ዋ	R - ር	Vo -ቮ
We - ወ	Wu - ዉ	Wi -ዊ	Xa - ፃ/ጣ	S - ስ	Wo - ዎ
Xe - ፀ/ጠ	Xu - ፁ/ጡ	Xi - ፂ/ጢ	Ya - ያ	T - ት	Xo - ፃ/ጦ
Ye - የ	Yu - ዩ	Yu - ዮ	Za - ዛ	U - ኡ	Yo - ዮ
Ze - ዘ	Zu -ዙ	Zi - ዚ	Cha - ቻ	V - ቭ	Zo - ዞ
Che - ቸ	Chu - ቹ	Chi - ቺ	Nya -ኛ	W - ዉ	Chso - ቾ
Nye - ኘ	Nyu - ኙ	Nyi -ኚ	Pha - ፋ	X - ፅ/ፕ	Nyo - ኟ
Phe - ፀ	Phu - ፁ	Phi - ፂ	Sha - ሻ	Y - ይ	Pho - ፆ
She - ሸ	Shu - ሹ	Shi - ሺ	Zha -ዣ	Z -ዝ	Sho - ሾ
Zhe - ዠ	Zhu - ዡ	Zhi - ዢ	Ny - ኝ	Zh - ዠ	Zho - ዦ
Ch - ች	Ph - ፆ	Sh - ሽ			

ABSTRACT

This study describes the design of a stemming algorithm for Wolaytta language. To give a solid background for the thesis, literatures on conflation in general and stemming algorithms in particular were reviewed. Since it is the nature and characteristics of affixation that guide the development of stemmer, the Wolaytta language 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. It is indicated that suffixation is the main word formation process in Wolaytta language. It is also attempted to show that the language is morphologically complex and uses extensive concatenation of suffixes.

The result of the study is a prototype context sensitive iterative stemmer for Wolaytta language. Error counting technique was employed to evaluate the performance of this stemmer. The stemmer was trained on 3537 words (80% of the sample text) and the improved version reveals an accuracy of 90.6% on the training set. The number of over stemmed and understemmed words on the training set were 8.6% (304 words) and 0.8% (28 words) respectively. When the stemmer runs on the unseen sample of 884 words (20% of the sample text), it performed with an accuracy of 86.9%. The percentage of errors recorded as understemmed and overstemmed on this unseen (test set) were 9% and 4.1%, respectively. Moreover, a dictionary reduction of 38.92% was attained on the test set. The major sources of errors are also reported with possible recommendations to further improve the performance of the stemmer and also for further research.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

As we know, the world is changing and changing fast. Some of these changes are social; others are political, and so on. No matter where we plan to live or what we plan to do; constant and rapid change will be part of our life. In this process, the role played by information is high. Everybody needs information in his day to day life. We need information (Sanders, 1981) to make the best decisions possible. In each of our personal activities, decisions are required and information is needed to support the decisions. Information is needed in virtually every field of human thought and action.

Recently, digital information and new forms of information technology have become the focus in our society. Especially, the growth in Information Technology (IT) contributed a lot to the great availability of recorded information. This leads to a further need to devise new ways of gaining access to the increasing volume of information created. Therefore, if it is to be used in the most economical and fruitful way, the information created must be organized. As discussed by Girma (2001), the mass production of electronic information, the development of computerized library collections and increasing awareness of the importance of information by the society in day-to-day activity (business, research etc.) demands storing, maintaining and retrieving information in a

systematic way. This becomes a requirement because it is getting difficult for users to access the large amount of information. Nail (1999) cited by Girma (2001) described information retrieval as the process whereby information (or documents containing it) is stored and made available to users, and retrieving those relevant to a user's information need.

Information retrieval systems (IRs) (Salton and McGill, 1983) are designed to facilitate access to stored information. In addition, they are concerned with the representation, storage and organization of information items. The components of an IR system consists of a set of information items, a set of requests and a mechanism to determine which information items are most likely to meet the requirements of the requests. Basically, users of information state their information needs in the form of queries and submit their queries to the information retrieval system. Based on their queries, the system outputs information items that are believed to be relevant to the user query. In other words, the items that have common entries in both the database and users query are returned to users. This happens when a matching exists between queries and information items. To attain a match between these two components, the items in the documents as well as in the query should be represented in some form.

Frequently, the user specifies a word in a query but variant of this word are present in a relevant document. The variation prevents a perfect match between a query word and a relevant document word. This problem can be overcome with the substitution of the words by their respective stems. A stem is the portion of a word that is left after the removal of its affixes (i.e., prefixes and suffixes). A typical example of a stem can be the word "comput" which is the stem for the variants "computed," "computing," "computation," and "computations." Therefore, in IR,

grouping words having the same root under the same stem (or indexing term) increases the success of matching of documents to a query.

Distinctions between IR systems can be made based on their effectiveness in relation to users query. The effectiveness of an IR system entirely depends on the indexing and searching techniques employed. In a very crude term, the search result from a given query can be an indication of how effective a given information retrieval is (Wakshum, 2000). It can be said that the higher the number of correct responses to users query, the more effective is the IR system and viseversa.

According to Salton, et al (1981), in information retrieval, the stored information items and the incoming search requests are normally represented by sets of content identifiers variously known as keywords, index terms, or simply terms. Stemming is part of the composite process of extracting the words from text and turning them into index terms in an IR system.

Indexing is the process of selecting keywords for representing a document. It can be done manually or automatically. In manual indexing the process of selecting the keywords is done by trained indexers, who are knowledgeable about the subject matter of the database, through scanning of the entire text or selected portions of the text, like titles and abstracts. When indexing is carried out by using computers, it is known as automatic indexing. Automatic indexing assumes the storage of information items in a database in a computer (Salton, 1983; Tesfaye, 1987).

As described by (Rijsbergen, 1979), the process of automatic indexing has three parts: removal of high frequency words, suffix stripping, and detecting equivalent terms. The first part, i.e. removal of f high frequency words, involves the removal of frequently occurring words (stopwords), such as definite articles, which have least significance in representing a document. The second part, i.e. suffix stripping (commonly done by stemming algorithm) process, is a conflation procedure that reduces different variants of the same word to single common term by removing suffixes. The last part, i.e. detecting equivalent terms, deals with selection of index terms.

An important component of any IR system for text searching is the ability to identify accurately the variant word forms that arise from grammatical modifications or alternative spellings of the words in a user's query. Such variants are normally encompassed by means of either right-hand truncation or stemming. Truncation is carried out by the searcher, who removes as many letters from the right-hand side of the word as seems appropriate to achieve a plausible root, and the search then retrieves all words that commence with this root, regardless of their endings. Stemming, conversely, is carried out automatically by reducing all words with the same stem to a common form, typically by removing the inflectional and derivational suffixes (Schinke et al, 1996).

1.2 STATEMENT OF THE PROBLEM AND JUSTIFICATION OF THE STUDY

Most of the languages in Ethiopia are included in the Afro-Asia language family. Of these Ethiopian languages majority are Cushitic and Omotic languages and some are Semitic languages (Bender, 1976). According to Adams (1983), the term "Wolaytta" has always been used by the Wolaytta (people) as the name by which they refer to themselves and their language. Fleming (cited by Adams, 1983) grouped the Wolaytta, Gemu, Goffa, Kucha, and Kullo dialects as one language called Wolaytta. According to the office of Population and housing Census Commission of Ethiopia (1999), there are over 3.3 million speakers of these languages. Of this sum those people who speak the Wolaytta dialect are about 2.3 million. Wolaytta is categorized under Omotic language family (Adams, 1983). The language plays a crucial role for the people of the region in social, political and economic activities.

The Wolaytta language serves as a medium of instruction in the primary schools, and is also offered as a subject in the junior and secondary schools of Wolaytta zone. Two other neighboring zones (Gamo Gofa and Dawro) are also using the language with minor dialect differences. In the primary schools, textbooks and other reference material are available in Wolaytta. The language is widely used in public services (e.g., marketing/shopping) and other events such as religious teachings. As a result, a significant number of people are able to read and write Wolaytta. The Latin script is being used since 1993 to write the language.

Lots of translation works have been started in translating materials from other languages to Wolaytta. The language is being studied at various levels both locally and by researchers from

abroad. For instance, there are different research works conducted in this language and organized by Section of Cushitic and Omotic languages, Department of African Languages and Cultures, Leiden University, the Netherlands (Section of Cushitic and Omotic languages, 2002). Currently there are over 10, 000 documents available in hard copy in Wolaytta (most of these materials are religious ones and some written before 15-25 years ago). A good sum of computers is being used in offices and educational institutions. Moreover, the language has become medium of instruction in schools. Such opportunities open bright future to produce more documents in Wolaytta language.

Wolaytta is morphologically complex language (Adams, 1983; Lamberti & Sottile, 1997). It uses both kinds of morphologies, i.e. inflectional and derivational. As pointed out by Alemayehu and Willet (2002), depending on the morphological complexity of a language, both inflectional and derivational morphologies can result in very large numbers of variants for a single word. As a result, word form variations can have a strong impact on the effectiveness of information retrieval (IR) systems and on morphological analysis tools. As Wolaytta is morphologically complex language, there is thus a need for automated procedures that can reduce the size of a lexicon to a manageable level and improve retrieval performance, and also capture the strong relationships existing between different word forms in the language.

Alemayehu and Willet (2003) further discussed that stemming plays an important role in the identification of a word stem from a full word by removing inflectional and derivational affixes, and there has thus been much interest for algorithms for this purpose. This interest is likely to

increase still further as more and more types of text-processing application become of wide spread importance.

At some point in time, we need a retrieval system for Wolaytta language text. This requires developing a model on how the derivation of Wolaytta words takes place; and how semantically related words can be conflated together algorithmically. Since stemmer is language specific, it is not possible to make use of a stemmer developed for other languages.

To the best of my knowledge, there has never been any effort made to develop a stemmer for use with Wolaytta text. Thus, it makes worth to do a research on developing a stemmer for Wolaytta text.

1.3 OBJECTIVES

1.3.1 GENERAL OBJECTIVE

The general objective of this research is to develop a stemmer for Wolaytta language (a prototype).

1.3.2 SPECIFIC OBJECTIVES

The specific objectives of the research are:

- review different stemming algorithms that have been developed for other languages.

- review properties of the Wolaytta language in order to get familiar with the different aspects of the language (e.g how the separation of words is achieved);
- select Wolaytta text for the experiment and prepare it for processing;
- compile a list of affixes used in Wolaytta;
- compile a stop word list in Wolaytta;
- Develop or adopt stemming algorithm(s) to experiment with Wolaytta text;
- Write (develop) computer program for stemming inflectional and derivational affixes of Wolaytta;
- test the developed algorithm to a selected Wolaytta text in order to see its effectiveness;
- forward recommendations for further study.

1.4 METHODOLOGY

1.4.1 LITERATURE REVIEW

As studying the language's morphology constitute an important component in the research, a literature survey is made to gather information and in understanding the subject. Furthermore, appropriate individuals especially from Omotic language branch of Ethiopian Languages Research Center, Addis Ababa University, are consulted on the morphology of Wolaytta.

1.4.2 DATA SOURCES

A text corpus is one of the resources required in IR researches. A good sized text can show a reasonable language morphological behavior. Selection of text is, therefore, an important component in developing a stemmer. For the purpose of this research, a corpus designed and collected for another purpose was used (Lamberti and Sotille, 1997). The corpus by Lamberti and Sotille (1997) was collected to study the Wolaytta language. Hence, same corpus was used as a representative sample to study the morphological behavior of words and come up with a Wolaytta stemmer.

1.4.3 DEVELOPING AND TRAINING THE STEMMER

After a detailed study of the language's morphology, a stemmer is developed based on 80% the sample corpus (See appendix VI). Compilation of stop words (non-content bearing words) and affixes were considered first. The characteristics of the affixes were then used to guide the development of the stemmer. The approach considered is iterative one and it was selected in due course.

1.4.4 STEMMER TESTING

Testing of the stemmer is made on the rest 20% of the dataset (unseen part of the corpus) (See appendix V). Then, the prototype stemmer developed for Wolaytta language was evaluated using

Wolaytta language morphology is reviewed in chapter three. The inflectional and derivational morphologies of the language are the main concerns of this chapter. Word formation processes for Wolaytta nouns, adjectives, and verbs are also presented in detail in the chapter.

Discussions on the development and experimentation of the stemming algorithm for Wolaytta text is dealt in the fourth chapter. The compilation of stopwords list and suffix list are presented in this chapter. The approach employed to develop the stemmer and the reasons for its selection are also parts of the discussions in the chapter.

The last chapter, chapter five, presents conclusions deduced from the findings and recommendations for future research.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

2.1 INTRODUCTION

This chapter presents the need for term conflation and approaches followed in term conflation for the purpose of free-text retrieval systems. Discussions are made on the advantages and disadvantages of the different approaches for term conflation. Finally, different stemming algorithms are reviewed from the point of view of modes of operation, methodology and performance.

2.2. CONFLATION TECHNIQUES

As discussed by Hull (1995) in information retrieval, the relationship between a query and a document is determined primarily by the frequency of terms that they have in common. But, words that appear in documents and in queries often have many morphological variants. The variations may arise from a range of causes including the requirements of grammar, e.g. "computing" and "computational"; valid alternative spellings, e.g. "recognize" and "recognise"; antonyms, e.g. “ability” and “disability,” and problems arising from misspellings and abbreviations. Lennon et al. (1981) considers these as one of the main problems involved in the use of free text for indexing and retrieval.

truncated to 'computer', rather than 'comput*' (which would also include words such as 'computing' and 'computational') (Tesfaye, 1987). Automatic conflation, on the other hand, is effected by means of a stemming algorithm. Stemming algorithm (*Stemmer* for short) is a technique for reducing words to their grammatical root form (Baeza-Yates, 1999; Croft and Xu, 1995). As stated by Salton, Wu, and Yu (1981), Stemming is part of the composite process of extracting the words from text and turning them into index terms in an IR system. Unlike manual conflation, automatic conflation is applied to words in queries as well as in documents.

Frakes (1992) distinguished four types of stemming strategies: affix removal, table lookup, successor variety, and n-grams. Affix removal algorithms (the strategy that will be followed in this research work) remove suffixes and/or prefixes from terms leaving a stem. Table lookup consists simply of looking for the stem of a word in a table. It is a simple procedure but one which is dependent on data on stems for the whole language. Since such data is not readily available and might require considerable storage space, this type of stemming algorithm might not be practical. Successor variety stemming is based on the determination of morpheme boundaries, i.e., it requires knowledge from linguistics, and is more complex than affix-removal stemming algorithms. It should be applied to large collections to get adequate statistical information, and work with letters instead of phonemes (Virginia Polytechnic Institute, 2003). N-grams stemming, which is not language dependent, is based on the identification of bigrams and trigrams and is more a term clustering procedure than a stemming one.

CHAPTER TWO

REVIEW OF RELATED LITERATURE

2.1 INTRODUCTION

This chapter presents the need for term conflation and approaches followed in term conflation for the purpose of free-text retrieval systems. Discussions are made on the advantages and disadvantages of the different approaches for term conflation. Finally, different stemming algorithms are reviewed from the point of view of modes of operation, methodology and performance.

2.2. CONFLATION TECHNIQUES

As discussed by Hull (1995) in information retrieval, the relationship between a query and a document is determined primarily by the frequency of terms that they have in common. But, words that appear in documents and in queries often have many morphological variants. The variations may arise from a range of causes including the requirements of grammar, e.g. "computing" and "computational"; valid alternative spellings, e.g. "recognize" and "recognise"; antonyms, e.g. “ability” and “disability,” and problems arising from misspellings and abbreviations. Lennon et al. (1981) considers these as one of the main problems involved in the use of free text for indexing and retrieval.

Morphological variants of the same word have similar semantic interpretation and are considered as equivalent terms for the purpose of IR application. But, they are not recognized as equivalent without some form of natural language processing. Hence, such equivalent word forms should be detected and reduced to a single form using conflation procedure thereby improving system performance through improving recall (Tesfaye, 1987; Paice, 1990; Hmeidi, Kanaan, and Evans, 1997). Actually, it doesn't matter whether the stems generated are genuine words or not – thus, "computation" might be stemmed to "comput" provided that different words with the same "base meaning" are conflated to the same form, and words with distinct meanings are kept separate.

According to Frakes (1992) conflation is defined as the process of matching morphological term variants. It maps term variants to a single form, usually a unique well-formed root for each word. Accordingly, the key terms of a query or document are represented by stems rather than by the original words (Al-Kharashi & Evans, 1994).

Conflation can be done either manually using some kind of regular expressions or automatically, via programs called stemmers. Manual conflation is commonly performed by right-hand truncation[1] at search time. It is applied only to words in queries. During the process of truncation two things could happen: over truncation and under truncation. Over truncation occurs when too short a stem remains after truncation and may result in totally unrelated words being truncated to the same stem; under-truncation, on the other hand, arises if too short a suffix is removed and may result in related terms being described by different stems, as with 'computers' being

[1] Truncation is carried out manually by the searcher who removes as many letters from the right-hand side of the word

truncated to 'computer', rather than 'comput*' (which would also include words such as 'computing' and 'computational') (Tesfaye, 1987). Automatic conflation, on the other hand, is effected by means of a stemming algorithm. Stemming algorithm (*Stemmer* for short) is a technique for reducing words to their grammatical root form (Baeza-Yates, 1999; Croft and Xu, 1995). As stated by Salton, Wu, and Yu (1981), Stemming is part of the composite process of extracting the words from text and turning them into index terms in an IR system. Unlike manual conflation, automatic conflation is applied to words in queries as well as in documents.

Frakes (1992) distinguished four types of stemming strategies: affix removal, table lookup, successor variety, and n-grams. Affix removal algorithms (the strategy that will be followed in this research work) remove suffixes and/or prefixes from terms leaving a stem. Table lookup consists simply of looking for the stem of a word in a table. It is a simple procedure but one which is dependent on data on stems for the whole language. Since such data is not readily available and might require considerable storage space, this type of stemming algorithm might not be practical. Successor variety stemming is based on the determination of morpheme boundaries, i.e., it requires knowledge from linguistics, and is more complex than affix-removal stemming algorithms. It should be applied to large collections to get adequate statistical information, and work with letters instead of phonemes (Virginia Polytechnic Institute, 2003). N-grams stemming, which is not language dependent, is based on the identification of bigrams and trigrams and is more a term clustering procedure than a stemming one.

2.3 STEMMING ALGORITHMS

Lovins (1968) defines a stemming algorithm as "a procedure to reduce all words with the same stem to a common form, usually by stripping each word of its derivational and inflectional suffixes." Stemming is generally effected by means of suffix dictionaries that contain list of possible word endings, and this approach has been applied successfully to many different languages (Ekmekcioglu, 2003).

Currently, a number of stemming algorithms have been developed in a notion of enhancing information storage and retrieval performance (Kosnov, 2003). For instance, there are various suffix-stripping algorithms have already been proposed, ranking from a weak stemmer, which removes plural inflections (and also perhaps the past participle "-ed" and the gerund or present participle "-ing"), to more sophisticated schemes designed to remove suffixes and even prefixes (Savoy, 1993).

Savoy (1993) further added that the design of such procedures is based mainly on one of two principles (or manners of operation):

- Longest match - removing the longest match suffix; and
- Iterative - iterating over a set of predefined classes of suffixes

A stemming algorithm can be iterative in character, use the principle of longest-match assignment, or a combination of these two methods. An iterative stemming algorithm is based on

the fact that suffixes are attached to stems one after the other. Such algorithm involves a recursive procedure which removes the suffixes one at a time, starting at the end of a word and working towards its beginning. For instance, a word such as willingness might have "–ness" removed in the first iteration and "-ing" in the second. Such an algorithm requires the construction of a dictionary, which contains a list of all possible word endings (Tesfaye, 1987).

Longest-match algorithms, on the contrary, involve only a single iteration, i.e., if more than one suffix matches the end of the word, the longest one is removed. This requires, however, the compilation of all possible combinations of suffixes. In order to reduce programming complexities, this list of suffixes is sorted in decreasing order of suffix length. The procedure is then to scan through suffix list in order of decreasing length. That is, the longer endings are first scanned, and if a match is not found, then the shorter ones are scanned. Thus the word "relativistic" would be stemmed to "relativ" if both "–istic" and "ic" were included in the suffix listing being used. Longest-match algorithms are often easier to program but require a match longer dictionary since frequent combinations of short suffixes must be included. Lovin's stemmer is one best example of longest match stemming algorithm (Tesfaye, 1987).

Another basic characteristic of stemming algorithm is whether it is context-free or context sensitive, which refers to any attribute of the remaining stem. In context-free algorithm, no restriction is placed on the removal of a suffix and thus any ending, which matches, is accepted for stripping. In context sensitive algorithms, however, various restrictions are placed on the usage of the suffix. Therefore, such kind of algorithm requires the construction of suffix dictionary and the formation of a set of rules defining the morphological context of the suffixes.

The dictionary gives the exact suffix form, while the rules define, for example, the treatment of dictionary suffix when preceded by a double constant (such as "stemming" -> "stem"), or the minimal root size that must be retained (such as the removal of "ual" from "factual" but not from "equal"), and some general rules, for example, do not remove a suffix that begins with "-en" following "–e" (as in "seen"). If the set of rules defining the correct morphological context for the suffix is satisfied, it is replaced by another string, either the null string (if the suffix is to be removed) or specified replacement string (for example, to create nominal forms to adjectival forms). Both dictionary and rules require careful analysis of vocabulary and language behavior, and are thus time-consuming to create. However, generally such techniques are rewarded by high accuracy and speed, and simplicity in implementation (Savoy, 1993).

According to Savoy (1993), to produce "better" stems, and even to find the "right" root, a context-sensitive approach adds constraints to the stripping operation. He proposes three general types of constraints:

- *Quantitative constraints*: the length of remaining stem must exceed a given number (e.g., from the word "ring" and the suffix "-ing", the system cannot derive "r" because a minimum length of two characters is required.)
- *Qualitative constraints*: the stem ending must satisfy a given condition (e.g., remove the suffix "-ize" if the remaining stem does not end with "e"). In this case, removing "-ize" from "seize" is not allowed.
- *Recoding rules*: spelling or adjustment rules must be used to improve the accuracy of conflation of the stems produced by the suffix stripping algorithm.

Context free removal leads to a significant error rate. For example, we may well want "ual" removed from "factual" but not from "equal." To avoid erroneously removing suffixes, context rules are devised so that a suffix will be removed only if the context is right (Rijsbergen, 1979).

Stemming, in general, increases recall at the cost of decreased precision. Studies of the effects of stemming on retrieval effectiveness are equivocal, but in general stemming has either no effect, or a positive effect, on retrieval performance where the measures used include both recall and precision (Nikolaidis and Kalamboukis, 2003). Stemming can also have a marked effect on the size of indexing files, some times decreasing the size of the files as much as 50 percent (Frakes, 1992; Harman, 1991).

The issue of stop words is also worth mentioning in relation to retrieval effectiveness. The removal of stop words from indexing and query, results in effectiveness of retrieval by reducing storage requirement and increasing the matching of a query with index terms of a document (Savoy, 1993). Hence, compiling a stop word list is important in building IR system. List of stop words can be compiled by sorting a vocabulary of a text corpus and making selection based on word frequency (Porter, 2000). Salton, et al (1981) supports this stating that the frequency characteristics of terms in the documents of a collection have been used as indicators of term importance for content analysis and indexing purpose.

So far, stemming algorithms have been developed for different languages such as English (Lovins, 1968; Porter, 1980; Dawson, 1974; Paice/Husk, 1990; Krovetz1993), Arabic (Al-

conflated to the same stem and words with similar meanings are not conflated at all. For example, the Porter stemmer conflates *general, generous, generation,* and *generic* to the same root, while related pairs like *recognize* and *recognition* are not conflated.

The Amharic stemmer uses a context-sensitive iterative procedure that removes both prefixes and suffixes. To measure performances of the stemmer, it was tasted on a sample data of 1221 words. The result of the experiment shows that the stemmer performed at an accuracy of 95.9%.

The Afaan Oromoo stemmer uses the longest-match context-sensitive approach and rules that removes prefix and suffix. The evaluation for the stemmer was by counting stemming errors and also reduction of dictionary size. The stemmer performed at an accuracy of 92.52% based on the sample data of 1061 word.

The Tigrigna Stemmer was developed based on iterative procedure and uses context-sensitive rules that removes prefix, suffix, prefix-suffix pair and reduplication of single and double letters. Similar to that of the Afaan Oromo stemmer, the accuracy of the stemmer is tested based on error counting techniques. The result of the experiment shows that the stemmer performs at accuracy of 84% and brings a dictionary reduction of 32.40%.

Two methods are widely used to evaluate stemmers: information retrieval performance and error counting. In the former case, the stemmer is used in information retrieval system and its performance monitored and evaluated. In the later case, the evaluation is based on error counting

where indices of the stemmer are generated by assessing the understemming and overstemmng errors (O'Neill, 2000).

There are mixed results about the evaluation of English stemmers. The results of most of the previous studies indicate that there is non-significant improvement in retrieval performance (Frakes, 1992). Harman (1991), for instance, reported a weakly positive result on retrieval performance by evaluating the Lovins stemmer, S-stemmer and the Porter stemmer. But, among recent works, a study by Krovetz (cited by Hull, 1995) indicates that there is favorable increase in retrieval performance when stemmers are used for English. Popovic el al. (1992) concluded that the effectiveness of stemming algorithm is determined by the morphological complexity of the language that is designed to process. Accordingly, the stemming research on other morphologically complex languages indicates significant performance difference between stemmed and non-stemmed operations.

As far as the performance in IR environment is concerned, of the three stemmers tried for Ethiopic languages, Amharic (Nega, 1999), Affan Oromo (Wakshum, 2000) and Tigrigna (Girma, 2001), except the Amharic stemmer, the other two were not tested for their effectiveness in IR environment.

2.4 CHAPTER SUMMARY

In this chapter, manual right-hand truncation and stemming as basic approaches to term conflation in free-text retrieval systems are discussed. Accordingly, it is viewed that stemming has got advantages over right-hand truncation. Stemming is a language dependent procedure because of the fact that the affixation rules are different for different languages. Four types of stemming strategies: affix removal, table lookup, successor variety and n-grams are dealt in depth. The two common modes of operations being used to develop stemmers, longest match and iterative, and their advantages and disadvantages are also discussed. It is seen that the performance results of English stemmers are equivocal. Especially, it is tried to indicate that stemming is more effective if a language is morphologically complex as indicated for non-English languages like French (Savoy, 1993) and Amharic (Nega, 1999). In the following chapter, Wolaytta language morphology is discussed.

CHAPTER THREE

MORPHOLOGY OF WOLAYTTA LANGUAGE

3.1 INTRODUCTION

The problem related to research in developing stemming algorithm for a language requires studying and modeling the language phenomenon in terms of word formation. There fore, it is necessary to study Wolaytta morphology in order to model it and develop an automatic procedure for conflation of words for the language. Accordingly, this chapter will be concerned on the inflectional and derivational morphology of the language since it is the nature and characteristics of affixation that guide the development of the stemmer.

3.2 MORPHOLOGY

Morphology is a branch of linguistic that studies and describes how words are formed in language (Hull, 1995; Morphology, 2003). Similarly, Silzer (cited by Wakshum, 2000) defines morphology as the study of the structure of words. As discussed by Brew (1997), 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.

3.2.1 TYPES OF MORPHEMES OCCUR IN WOLAYTTA

As defined by Silzer (cited by Wakshum, 2000) a morpheme is the minimal linguistic unit of a language that carries a meaning and that can not be further decomposed into a meaningful unit. 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). Both types of morphemes occur in Wolaytta (Adams, 1983).

3.2.2 HOW WORDS ARE FORMED IN WOLAYTTA

Affixation and compounding are two basic word formation processes in Wolaytta (Lamberti and Sottile, 1997). Affixation is a process by which affixes are added/attached in some manner to the root, which serve as a base. Affixes are morphemes that can not occur independently. Prefix, suffix, and infix are the three types of affixes. Among Ethiopic languages, for instance, both types of affixes are reported for Amharic (Nega, 1999) and Tigrigna (Girma, 2001). I could ascertain from the existing literature that Wolaytta language does not have prefix and infix. Instead, Suffixation is the basic way of word formation in Wolaytta.

In forming word, adding one suffix to another is common in Wolaytta. This process of adding one suffix to another suffix can result in relatively long word, which often contains an amount of

semantic information equivalent to a whole English phrase, clause or sentence. Due to this complex morphological structure, a single Wolaytta word can give rise to a very large number of variants (Adams, 1983).

The second word formation process in Wolaytta is compounding. According to Wardhaugh (1977), compounding is the joining together of two linguistic forms, which function independently. Although Wolaytta is very rich in compounds, compound morphemes are rare in Wolaytta and their formation process is irregular (Lamberti and Sottile, 1997). As a result, it is difficult to determine the stem of compounds from which the words are made.

3.3 INFLECTIONAL AFFIXES OF WOLAYTTA

3.3.1. NOUNS

The basic phonetic structure of the Wolaytta nouns consists of the character sequence of $C_1V_1C_2V_2$, where C and V represent a consonant and a vowel respectively. This means that most Wolaytta stems are in principle bi-radical[1]. C_1 can also represent a glottal stop. V_1 consists either of a short or of a long vowel or more seldom of a diphthong, while C_2 can represent simple or geminated consonant. Finally V_2 represents either the usual ending in the absolutive case or a thematic vowel by which the endings required by the syntactic function of the noun that are linked to the noun stem (Lamberti and Sottile, 1997). Examples (3.1a), (3.1b), (3.1c) and (3.1d) illustrate the basic phonetic structure of Wolaytta nouns described above:

[1] radical is the number of consonants in a word (stem)

(3.1a)

CVCV:

?asa 'person' ?isha 'brother' kana (dog),

(3.1b)

CVC:V:

?acca 'tooth' matta 'bee' kamma 'night'

(3.1c)

CV:CV:

doona 'tongue' ?aawa 'father' miiza 'cattle'

(3.1d)

CV:C:V:

buucca 'chin' toossa 'god' ?eessa 'honey'

Wolaytta also possesses pluriradical (mostly tri-radical) stems.

(3.2) wozan-a 'heart' ?agen-a 'moon' goofin-iy[a] 'lung'

3.3.1.1 INFLECTION OF NOUNS

Lamberti and Sottile (1997) subdivided wolaytta nouns in four main classes according to the endings they take in their inflection. Thus, the classification of a noun as a member of the one or

3.3.1.1.1 GENDER

The Wolaytta noun system, like most other languages, exhibits two different genders: "masculine" and "feminine" (Lamberti and Sottile, 1997). Accordingly, the nouns belonging to the 4th class are "feminine" while nouns belonging to all others (i.e. those of the 1st, 2nd and 3rd class) are defined "masculines". Formally, feminines differ from masculines by their endings. The former end in *–u* in the absolutive case where as masculine are all characterized by ending in *–a* (in the absolutive case) and, if they are used as subject, additionally by the marker *–y* (see (3.8) for examples). Exceptions are possessive pronouns and some demonstratives. These may show forms which are limited to the one or to the other gender as shown in (3.9.) below.

(3.8)

dorsa ('sheep', masculine)	vs.	dorsu ('sheep', feminine)
desha ('goat', masculine)	vs.	deshu ('goat', feminine)

(3.9)

ha-ge 'this, masculine'	vs.	ha-nna 'this, feminine'
taa-gaa 'he/it is mine'	vs.	taa-ro 'she/it is mine'

3.3.1.1.2 NUMBER

The Wolaytta noun system comprehends two numbers, namely the singular and the plural (Adams, 1983). The singular usually consists of the basic noun form, while the plural is formed by means of a suffix used only for this purpose.

Plurals formed by means of –nta are represented by a few pronouns such as the demonstrative pronoun *hanni-nta* 'these, for both genders' (*ha-nna* 'this, feminine'), reflexive pronoun *ba-nta* 'themselves/each other' (*ba* 'himself/herself') and interrogative pronoun *awunni-nta* 'which ones?, for both gender' (awu-nna 'which one, feminine').

The plural ending *–ta* reflects the inflection of the plurals in their absolutive or object case. If a plural noun is used as subject, then the ending *–ta* changes to *–ti*, while the inflecting stem for the other cases as a rule ends in *–tu-*.

(3.15)

inflecting stem	***absolutive/ object case***	***subject case***	***gloss***
naa-	naa-ta	naa-ti	"children",
?asa-	?asa/ta	?asa-ti	"persons",
gellaaw-	gelaaw-ta	gelaaw-ti	"girls".

3.3.1.1.3 CASE

Wolaytta exhibits a quite complicated noun inflection. This consists of several cases. The inflection takes place by the suffixation of case endings to the noun stem or to the absolutive case form. Accordingly, the absolutive case is characterized, as we have already seen above, by the ending *–a* (1st class and plural), *-iya* (2nd class), *-uwa* (3rd class) and *–(i)yu* (4th class) respectively, while the subject case ends in *–y* (first three classes), *-i* (plural), and *–(i)ya* (4th class). The genitive is represented either by the noun stem alone or is more often characterized by the

lengthening of the final vowel of the absolutive form. The object case of the noun inflection agrees with the respective absolutive case (Lamberti and Sottile, 1997). The other cases are marked by the endings shown in (3.16) below.

(3.16)

case marker (morphemes)	***function***
-ssi, -w or –yoo	dative and benefactive case
-kko or -mati	directive case
-ni or garsani	locative
garsa	inessive
-ppe	ablative
-ppe or garsaappe	exitive
-ni	locative and instrumental
-ra	comitative and sometimes instrumental case
-ow or –ey	vocative

The case markings *–ssi*, *-w* and *–yoo* are free variants and as such they are interchangeable. They show the restriction, however, that *–w* may not be used in connection with plurals.

(3.17)

na?a 'boy' ==➔ na?aa-ssi or na?aa-w or na?a-yoo 'to/for the boy'

naata 'children' ==➔ naatuu-ssi or naatuu-yoo/naataa-yoo 'to/for the children'

Similarly *–kko* and *–mati* seem to be two free variants, but this is noted here with a certain reservation. The morphemes *–mati*, *garsani*, *garsa* and *garsaappe* are no case endings in their proper sense, but they are actually postpositions, the last two (i.e. garsa and garsaappe) always involve a certain motion into and out of a place respectively.

Examples are given in table 3.1, 3.2, 3.3, 3.4 and 3.5 for the case markers given in (3.16) above. In the examples, the nouns *na?a* 'boy', *keetta* 'house' and *tama* 'fire' are used for the 1st class (see table 3.1); *bitanniya* 'man', *satiniya* 'box' and *kaamiya* 'lorry' are used for the 2nd class (see table 3.2);

kaawuwa 'king' and *tohuwa* 'foot/leg' are used for the 3rd class (see table 3.3) and *na?yu* 'girl' and *macciyu* 'wife' are used for the 4th class. Finally, the nouns *naata* 'children', *?asata* 'people' and *keettata* 'houses' are used for the plural (see table 3.5).

Table 3.1 Case markers for 1st class

Case markers	**Examples**
absolutive case	na?-a 'boy', keett-a 'house', tam-a 'fire'
subject case	na?-a-y 'the boy', keett-a-y 'the house', tam-a-y 'the fire'
object	na?-a 'the boy', keeta-a 'the house' tam-a 'the fire'
genitive	na?-aa 'of the boy', keett-aa 'of the house'

genitive	bitann-iy-aa 'of the man'
dative/benefactive	bitann-iaa-ssi/bitann-iya (a) –w/ bitann-iy-oo/bitann-iyaa-yoo 'to/ for the man'
directive	bitann-iya-kko 'to the man'
locative	satin-iya-ni 'in the box'
inessive	satin-iyagarsa 'into the box'
ablaitive	bitann-iyaa-ppe 'from the man'
exitive	kaam-iyaa-ppe/kaamiya garsaappe 'out of the lorry'
instrumental	kaam-iyaa-ni 'by lorry'
comitative	bitann-iya-ra 'with the man'
vocative	bitann-iy-ow 'eh, man!'

Table 3.3 Case markers for 3rd class

Case markers	**Examples**
absolutive case	kaaw-uwa 'king', toh-uwa 'leg/foot'
subject case	kaaw-o-y 'the king', toh-o-y 'the leg/foot'
object case	kaaw-uwa 'the king', toh-uwa 'the leg/foot'
genitive	kaaw-uw-aa 'of the king'
	kaaw-uwaa-ssi/ kaaw-uwaa-yoo

dative/benefactive	'to/for the king'
locative	ees-uwaa-ni 'in a hurry'
instrumental	toh-uwa-ni 'by/on foot'
comitative	toh-o-ra 'with the foot/ on foot' soh-uwa-ra 'on the place/suddenly'

Table 3.4 Case markers for 4th class

Case markers	**Examples**
absolutive case	na?-yu 'girl', macc-iyu 'wife'
subject case	na?-ya 'the girl', macc-iya 'the wife'
object case	na?-yu 'the girl', macc-iyu 'the wife'
genitive	na?-ee 'of the girl', macc-ee 'of the wife'
dative/benefactive	na?-ee-ssi/na?-ee-mati 'to for the girl'
directive	na?-ee-kko/na?ee-yoo 'to the girl' ta macc-ee-kko 'to my wife'
ablaitive	na?-ee-ppe 'from the girl'
comitative	na?-ee-ra 'with the girl'
vocative	na?-ey 'eh, girl!'

Table 3.5 Case markers for plural

Case markers	Examples
absolutive case	naa-t-a 'children', ?asa-t-a 'people' keetta-t-a 'houses'
subject case	naa-t-a 'the children', ?asa-t-i 'the people' keetta-t-i 'the houses'
object case	naa-t-a 'the children', ?asa-t-a 'the people' keetta-t-a 'the house'
genitive	naa-t-u 'of the children', ?asa-t-u 'of the people' kaaw-o-t-u 'of the king'
dative/benefactive	naa-t-uu-ssi/ naa-t-uu-yoo/naa-t-aa-yoo 'to/for the children'
directive	naa-t-u-kko 'to the children', ?asa-t-u-kko 'to the pople'
locative	keetta-t-u-ni 'in the houses'
inessive	?asa-t-u garsa 'among the people'
ablaitive	keetta-t-uu-ppe 'from the houses' laagge-t-u-ppe 'from the friends'
comitative	naa-t-u-ra 'with the children' laagge-t-u-ra 'with the friends'
vocative	?asa-t-oow 'eh, people!' naaggade-t-oow 'eh, merchants'

3.3.1.2. DERIVATION OF NOUNS

As it is discussed by Lamberti and Sottile (1997) Wolaytta exhibits a considerable amalgam of formative suffixes[1]. The class suffixes we have already seen above (-a, -uw^{a}, -(i)y^{u} and less frequently -iy^{a}) (see section 3.3.1.1 for more detail), are also often used to form nouns from a verb stem. Nouns formed by the class suffixes -a and -uw^{a} can refer either to abstract terms or to very concrete objects and they also serve to express action nouns. Examples are given in (3.18) below.

(3.18)

hassay- 'speak'	==➔	haasay-a 'conversation/action of speaking'
wott- 'run'	==➔	wott-a 'action of running'
?er- 'know'	==➔	?er-a 'wisdom'
fayd- 'count'	==➔	fayd-uwa 'number'
oot- 'work'	==➔	oot-uwa 'work/action of working'
be?- 'see'	==➔	be?-uwa 'action of seing'

As it is shown in (3.19) below, the formative suffix of the 4th class *-(i)yu* serves to derive nouns referring to living beings of female sex.

(3.19)

na?-a 'boy/son'	==➔	*na?-yu* 'girl/daughter'
cim-a 'old man'	==➔	cim-iyu 'old woman'
duud-iya 'dumb man'	==➔	duud-iyu 'dumb woman'

[1] Formative suffixes are suffixes that have important influence on the meaning of the stem to which they are attached.

The morpheme *-iya* is also very productive in Wolaytta.

(3.20)

waayy-iya 'trouble'

harg-iya 'sickness'

kussh-iya 'hand'

Two additional Wolaytta formative morphemes are the suffixes *-tta* and *-ta*. They also occur very frequently in Wolaytta.

(3.21)

Suffix *-ta*:

gulba-ta 'knee'

sun-ta 'name'

wurse-tta 'end'

The morpheme *-tta* also serves to form action nouns.

(3.22)

d?- 'live/exits' ==➔ de?e-tta 'the action of living'

y- 'come' ==➔ yee-tta 'the action of coming'

m- 'eat' ==➔ mee-tta 'the action of eating'

g- 'say' ==➔ gee-tta 'the action of saying'

A morphological compound of *te-tta* is the denominative for abstracts.

(3.23)

eeyya-tetta 'stupidity'

kaawo-tetta 'kingdom'

The suffix *-sa* is used to form nouns from monoradical verbs. See the examples given in (3.24) below.

(3.24)

m- 'eat' ==➔ muu-sa 'the action of eating'

b- 'go' ==➔ buu-sa 'the action of going'

y- 'come' ==➔ yuu-sa 'the action of coming'

g- 'say' ==➔ guu-sa 'the action of saying'

Lamberti and Sottile (1997) further pointed that agent nouns are formed by attaching the suffix *aanca.*

(3.25)

?er- 'know' ==➔ ?er-aanca 'wise person'

boog- 'rob' ==➔ boog-aanca 'robber'

word- 'tell lies' ==➔ word-aanca 'liar'

3.3.2 ADJECTIVES

The phonetic structure of adjectives corresponds to that of nouns but in Wolaytta their qualificative function is rather formally accomplished by nouns inflected in genitive and placed before the governing noun which is to be qualified. The words acting as" qualificative modifiers" (i.e. as adjectives.) are as a rule biradical in wolaytta, ending in -a,-iy^{a} or -uw^{a} and characterized by the same formative suffixes seen for nouns (Lamberti and Sottile, 1997). See (3.26), (3.27) and (3.28) below.

(3.26) "adjectives" ending in *-a*:

geessh-a 'clean' cinc-a 'clever' kaant-a 'short'

(3.27) "adjectives" ending in *-iya*:

mal-iya 'sweet' haankett-iya 'violent' yesshiss-iya 'fierce'

(3.28) "adjectives" ending in -uwa:

lo?-uwa 'good/nice' yuush-uwa 'round'

When adjectives are used in the attributive position, they always precede the noun they modify and very often remain unchanged, since they do not have to agree in Wolaytta with their governing noun either in gender or in number or in case (Lamberti and Sottile, 1997). But if used

in the attributive position, most adjectives ending in *-uwa* and a few in *-iya* are replaced by the endings -o and -e respectively.

(3.29)

lo?-uw[a] 'good/nice' ==➔ lo?-o asa 'a good person'

haah-uw[a] 'wide/far' ==➔ haah-o sohuwa 'a far place'

luul-iya 'straight' ==➔ luul-e ?ogiya 'a straight road'

Similarly, Lamberti and Sottile (1997) stated that if adjectives are used in predicative position, they do not usually undergo any change, thus they appear in their basic form. If adjectives are used in predicative position, -uwa will be changed to -o.

(3.30)

lo?-uwa 'nice/good' ==➔ lo?-o:

he bitanney lo?-o 'this man is/was good'

he gelaawiya lo?-o 'that girl is/was nice'

Wolaytta does not make use of any suffixes or prefixes in order to express the comparative (i.e., the English morphemes -er and -est) (Adams, 1983; Lamberti and Sottile, 1997). The second comparison term is inflected in ablative and it is immediately placed before the adjective in the comparative. If the adjective is used attributively, it usually appears in its basic form.

(3.31)

he macc'a-?asee-ppe adussa bitanne-y yiisi[i] 'the man who is taller than that woman came'

Otherwise, if the adjective is used productively, it can appear either in its basic form as in (3.32) or in an inflected form as in (3.33).

(3.32)

he bitanney he macc'a-?asee-ppe adussa 'this man is taller than that woman'

(3.33)

hage hagaa-ppe adukkeese 'this is longer than that'

3.3.3. VERBS

As pointed by Lamberti and Sottile (1997), the verbal system of Wolaytta, like that of most Ethiopian languages, is very complex.

Almost all Wolaytta verbal stems are biradical in their basic structure and they usually consist of a sequence $C_1V_1C_2$, where C stands for a consonant and V for a vowel.

(3.34)

gel- 'enter' moog- 'bury' ker- 'split'

There are also verbal stems which are tri-or, less commonly, pluriradical. They are as a rule either loans (especially from Ethio-semitic) (see 3.35) or secondly derived stems (see 3.36).

(3.35)

?azzaz 'command' nabbab- 'read' kassas- 'accuse'

(3.36)

be?erett- 'see many times' bayzett- 'be sold' wordot- 'tell lies'

Wolaytta also possesses a few verbal stems characterized by the loss of the second stem consonant and thus consisting of only one radical.

(3.37)

m- 'eat' b- 'go" y- 'come' g- 'say'

3.3.3.1. VERB INFLECTION

Wolaytta verbs can be conjugated according to four different inflections: a positive inflection, a negative inflection, an interrogative inflection and an interrogative-negative inflection (Lamberti and Sottile, 1997). These four inflections differ from each other by the fact that they are characterized by different endings. More over, Wolaytta exhibits a very complex inflection system depending on different parameters. In its inflection, the Wolaytta verb depends on mood, tense, kind of action and aspect. Tables 3.6 - 3.13 below (all adopted from Adams, 1983) provide summarized information about Wolaytta verb inflection.

Table 3.6 Inflection of Independent verbs (Punctilliar[1] primary aspect - Past tense)

STATEMENT			QUESTION	
PERSON	**AFFIRMATIVE**	**NEGATIVE**	**AFFIRMATIVE**	**NEGATIVE**
	I gave	I didn't give	Did I give?	Did I not give?
1	?imm-a:si	?imm-abeikke	?imm- idana:	?imm-abeikkina
2	?imm-adasa	?imm-aba:kka	?imm -adi	?imm -abeikki
3*f*	?imm-a:su	?imm-abe:ikku	?imm -ade	?imm -abe:kke
3*m*	?imm-i:si	?imm-ibe:nna	?imm -ide	?imm -ibe:nne
1*pl*	?imm-ida	?imm-ibo:kko	?imm-ido	?imm -ibo:kkoni
2*pl*	?imm-ideta	?imm-ibe:kketa	?imm-ideti	?imm -ibe:kketi
3*pl*	?imm-idosona	?imm-ibo: kkona	?imm-idona	?imm -ibo:kkona

Table 3.7 Inflection of Independent verbs (Punctilliar primary aspect - Future tense)

STATEMENT			QUESTION	
PERSON	**AFFIRMATIVE**	**NEGATIVE**	**AFFIRMATIVE**	**NEGATIVE**
	I will give	I will not give	Will I give?	Will I not give?
1	ta ?imm-ana	?imm-ikke	ta ?imm-ane	?imm-ikkina
2	ne ?imm - ana	?imm -akka	ne ?imm - ane:/?immu:te	?imm -ikki
3*f*	a ?imm - ana	?imm -ukku	a ?imm -ane	?imm -ekke
3*m*	i ?imm - ana	?imm -enna	i ?imm - ane	?imm -enne
1*pl*	nu ?imm - ana	?imm -okko	nu ?imm -ane	?imm -okkoni
2*pl*	?inte ?imm - ana	?imm -ekketa	?inte ?imm- ane:/?immu:teti	?imm -ekketi
3*pl*	?eti ?imm - ana	?imm -okkona	?eti ?imm -ane	?imm -okkona

[1] Punctilliar is to mean "perfect" or "complete" or non-continuous, happening at some point in time.

Table 3.10 Inflection of Subordinate verbs

ENGLISH	POLARITY	PERSONS	PUNCTILLIAR			CONTINUOUS/TENSELESS/	
			PAST		FUTURE	SAME PARTICI-PANT	DIFFERRNT PARTICIPANT
			SAME PARTICIPANT	DIFFERR-NT PARTICIP-ANT			
having	Aff	1,2,3f	?imm-ada	?imm-ini	same as past		
		3m pls	?imm -idi	?imm -ini			
	Neg	all	?imm -ennani			-aidda	?imm -isini
while	Aff	1,2,3f				-i: ddi	?imm - isini
		3m pls					
	Neg	all				-ennani	?imm -enna:sini
during the time	Aff	all		?imm - ida:sini	?imm - ana:sini		?imm -isini
	Neg	all		?imm - enna:sini	?imm - enna:sini		?imm -enna:sini
if	Aff	all			?imm - ikko		
	Neg	all			?imm - ana tayikko		
even though	Aff	1,2,3f	?imm -aiddakka				
		3m pls	?imm -iddakka				
	Neg	all	?imm -ennanka				
	Aff	all	?imm -ikkokka				
		all	?imm -ikkonne				
to	Aff	all			?imm - ana		
just after	Aff	1,2,3f	?imm -anne				
		3m, pls	?imm -inne				
soon after	Aff	all	?imm -obare				

Table 3.11 Inflection of Relative verbs

ASPECT	TENSE	POLARITY	PERSONS	SUBJECT-ORIENTED 'who ...'	NON-SUBJECT-ORIENTED 'which ...'
PUNCTILLIAR	PAST	Aff	all	?imm-ida 'who gave'	?imm-ido 'which (I) gave'
		Neg	1,2,3f	?imm -abe:nna 'who didn't give'	same as subject-oriented
			3m,pls	?imm -ibe:nna 'who will give'	same as subject-oriented
	FUTURE	Aff	all	?imm -ana who will give'	same as subject-oriented
		Neg	all	?imm -enna 'who won't give'	same as subject-oriented
CONTINUOUS	Tenseless	Aff	all	?imm -iya 'who is giving'	?imm-iyo 'which (he) is giving'

Table 3.12 Inflection of verbs (Uncertainty aspect)

ASPECT	TENSE	PERSONS	AFFIRMATIVE	NEGATIVE
PUNCTILLIAR	PAST	1	'did he find or not'	'did he not find of did he?'
			demm-adina:ssa	demm-abeikkina:ssa
		2	demm -adi:ssa	demm-abe:ikki:ssa
		3*f*	demm -ide:ssa	demm-abe:kke:ssa
		3*m*	demm -adi:ssa	demm-ibe:nne:ssa
		1*pl*	demm -idoni:ssa	demm-ibe:kkoni:ssa
		2*pl*	demm -ideti:ssa	demm-ibe:kketi:ssa
		3*pl*	demm -idona:ssa	demm-abo:kkona:ssa
	FUTURE	1	demm -andina:ssa	demm -ikkina:ssa
		2	demm -andi:ssa	demm -ikki:ssa
		3*f*	demm -ande:ssa	demm -ekke:ssa
		3*m*	demm -ande:ssa	demm -enne:ssa
		1*pl*	demm -andoni:ssa	demm -okkoni:ssa
		2*pl*	demm -u:teti:ssa	demm -ekketi:ssa
		3*pl*	demm -andona:ssa	demm -okkona:ssa
CONTINUOUS	TENSELESS	All persons	demm -ane:ssa	demm -enne:ssa

Table 3.13 Inflection verbs (Hypothetical - desiderative aspect)

PERSONS	AFFIRMATIVE	NEGATIVE
1	'Oh if only he would ...!'	'Oh if only he would not have ...!'
	?imm-arkina:ssa	?immenani ?agg-arkina:ssa
2	?imm-arki:ssa	?immenani ?agg-arki:ssa
3*f*	?imm-arke:ssa	?immenani ?agg-arke:ssa
3*m*	?imm-e:renne:ssa	?immenani ?agg-e:ernne:ssa
1*pl*	?imm-orkoni:ssa	?immenani ?agg-orkoni:ssa
2*pl*	?imm-erketi:ssa	?immenani ?agg-erketi:ssa
3*pl*	?imm-orkona:ssa	?immenani ?agg-orkona:ssa

3.3.3.2. VERB DERIVATION

Like other Cushitic and Semitic languages of Ethiopia, Wolaytta makes use of some morphemes in order to derive a further stem from a root or a stem. This derivation procedure is concretely applied in Wolaytta by suffixing one or more morphemes to verbal stem (Lamberti and Sottile, 1997). They further indicated that it is possible to form three different kinds of secondarily derived verbal stems: iteratives (or intensives), causatives and passives (or reflexives).

Iteratives and intensives stems are expressed in Wolaytta by means of the same morpheme erett, which is regularly suffixed to its verbal stem.

(3.38)

ment- 'to break' ==➔ ment-erett- 'to break many times or in many pieces'

shissh- 'to collect' ==➔ shissh-erett- 'to collect many times or in many things'

siy- 'to hear' ==➔ siy-erett- 'to hear many times /hera many things'

The Wolaytta language possesses only productive causative morpheme, i.e –is-s.

(3.39)

k'or- 'pick up' ==➔ k' or-iss- 'let someone pick up'

gel- 'enter' ==➔ gel-iss- 'let someone enter/put into'

be?- 'see' ==➔ be?-iss- 'let someone see'

Verbs which have their primary stem ending in *-y-* or *-y-y-* form their causative by replacing all existence of *-y-* by *-sh-*.

(3.40)

uy-y- 'drink' ==➔ ush-sh- 'let someone drink'

yuuy-y- 'turn, intr' ==➔ yuush-sh- 'turn,tr'

kooy-y- 'ask for/want' ==➔ kiish-sh- 'let someone ask for be necessary'

Passive verb stem serves to confer the following six connotations on the verb (Lamberti and Sottile, 1997) (see 3.41, 3.42, 3.43, 3.44 and 3.45 below for examples):

1. It changes active verbs to passives;
2. It renders transitive verbs intransitive;
3. It renders active or transitive verbs reflexive;
4. It serves to express benefactive relationship or the subject's interest in the action;
5. It can express the reciprocity of the action;
6. It serves to form verbs derived from nouns and adjectives.

Passives are formed in Wolaytta by means of the morpheme *-et-t-*.

(3.41)

kor- 'collect' ==➔ kor-ett- 'be collected'

mat- 'pick up' ==➔ mat-ett- 'be picked up'

siy- 'hear' ==➔ siy-ett- 'be heared'

(3.42)

de- 'live' ==➔ de-ett- 'be lived, said of a life'

gel- 'enter' ==➔ gel-ett- 'be entered, said of a room or a house'

The denominatives derived from nouns belonging to the first class are formed by means of the morpheme *-a(a)t-/ -aat-t-*.

(3.43)

ish-aat- 'become a brother/ become a very good friend'

aaw-aat 'become father'

ful-aatt- 'be beautiful'

Nouns belonging to the second or the fourth class form their denominatives by means of the morpheme *-eet-(t-)*.

(3.44)

dur-eett- 'become rich'

laagg-eett- 'become a friend'

3.5 CHAPTER SUMMARY

This chapter mainly discussed Wolaytta language morphology. Both the inflectional and derivational morphologies involve suffixing It has been observed that the main word formation process in Wolaytta is regular. A single word in the language can have so many variants (See appendix III). It is also shown that the language uses an extensive concatenation of suffixes. This characteristic of the language make a very short stem into a long word even some times equivalent to a sentence in English. As it is evidenced by different authors, such nature of the language indicates the complexity of the language's morphology. The complexity of the language is one of the main reasons for a language to desire a stemmer since stemmer is an automated mechanism that conflates variants of words. It is an important to access documents in natural language text. The next chapter presents the development of a stemming algorithm to conflate variants of Wolaytta words including other matters related to this issue.

CHAPTER FOUR

DEVELOPMENT OF STEMMER FOR WOLAYTTA TEXT

4.1. INTRODUCTION

This chapter and the next present the core or major contribution of the study. This chapter, as a main concern discusses the actual development of the stemmer for Wolaytta text. Prior to this issue, discussions are made on the training and test set collected from the sample text. Discussions also include the suffixes and stopwords compiled in the suffix list and stopword list from the sample text. The procedures followed to come up with the suffix list are presented in this chapter, too.The approach adopted to develop the stemmer and reasons for selection are also discussions made here. Finally, a report is made on how the developed stemmer is trained using set of data and tested on another set of data in order to evaluate its performance.

4.2. SAMPLE TEXT

The text used by Lamberti and Sottile (1997) for studying the morphology of Wolaytta language was used as it is for this research work. The main reason for the selection is that these authors stated that they thought the data is satisfactory to study the language. Since this study is mainly on the morphology of the language, it seems to me also that the same data could be helpful in the current work. In terms of number of words, the sample text consists of a total of 4421 words

(1122 sentences with varieties of words presented in different parts of speech) (For more detail see Lamberti and Sottile, 1997).

4.2.1. TEST SET

To test the performance of the stemmer, 20% (884 words) of the sample text were used as a test set (or data). First the entire words in the sample text are put in alphabetic order and then systematic random sampling technique is employed to select the words that are to be included in the test set. To come up with the test set, the first word was deliberately selected and then every 5^{th} word is selected after the last selected word. The number of words to be passed before getting the next word to be selected was decided by dividing the size of the sample text to the number of words to be included in the test data. The aim in doing this was to get whole test data in one pass through the sample data without coming back for another round in order to protect the chance of a word to be selected again and again.

4.2.2. TRAINING SET

All the 3537 words (80% of the sample test) that were not included in the test set were used for training the stemmer.

4.3. WORD DISTRIBUTION OF WOLAYTTA

Hameidi (1997) cited by Girma (2001) indicated in his work that the morphological complexity of a language can be determined based on the word ratio of the total words to the distinct words

in a given text. Even if no treshold value is set, it is pointed in general that the higher the word ratio, the lesser the morphological complexity of a language. Table 4.1 below shows the word ratio calculated for Wolaytta from the sample text and figures for other local languages and English. The data for Affan Oromo is adopted from Wakshum (2000) and the rest except Wolaytta are adopted from Girma (2001).

Table 4.1. Comparison of word distribution ratio

Language	Text	Total words	Distinct words	Word-ratio
Wolaytta	Text 1	4421	2277	1.941
Tigrigna	Text 1	1632	918	1.777
Afan oromo	Text 1	1555	856	1.816
English	Text 1	1600	621	2.576

From the data in table 4.1 above, although there is considerable difference in the size of words between the text considered for Wolaytta and other languages, Wolaytta seems less complex language than the two local languages considered here. However, the language is more complex as compared to English. Close analysis of the results show that the difference in word ratio is less between the three local languages considered here. In general these figures indicate that Wolaytta is a morphologically complex language.

4.4. COMPLATION OF STOPWORD LIST

Stopwords were compiled from the sample text by collecting the most frequently occurring words. First frequency of each word in the text is identified. Then the words are arranged in descending order based on their frequency. This task is done semi-automatically by MS-Excel office application software. The top 30 words (highly frequent words) are listed in Table 4.2 below.

Table 4.2. Top highly frequent words in the sample text

No	Word	Frequency
1	guutta	107
2	aa	96
3	ne	96
4	loo	93
5	ha	86
6	he	73
7	aw	71
8	aybaa	52
9	asati	51
10	nu	41
11	taw	38
12	neeni	34
13	adussa	33

14	ba	31
15	daroo	30
16	eeti	30
17	eeta	29
18	kuma	27
19	hegati	26
20	taani	26
21	ii	25
22	ubbaappe	25
23	yoota	25
24	hagani	24
25	tani	24
26	hinte	23
27	yiisi	23
28	bitanney	22
28	hagata	22
30	nuuni	21

As can be seen from Table 4.2, the stopword list consists of prepositions, conjunctions, articles and the like. There are also non-function words (content-bearing words) in the list. Function words such as prepositions and articles in Wolaytta exist affixed to words. For this reason, the frequency of function words in Wolaytta does not seem to be as high as in some other languages like English. That is, the frequency of function words in Wolaytta is low. Because of this reason,

the use of frequency alone did not help to generate all the stopwords and many other function words were added manually to the stopword list by consulting relevant literature. The complete list of the stopword list compiled from the sample text is given in Appendix I.

4.5. WOLAYTTA AFFIX LIST COMPILATION

Among local languages for which extensive list of prefixes, suffixes, and prefix-suffix pairs are reported are Amharic (Nega, 1999) and Tigrigna (Girma, 2001). In the case of Affan Oromo language (Wakshum, 2000), only one prefix (hi/ni) is identified and otherwise the language is based on suffixation to create words. Unlike these languages, Wolaytta is a language seems dependent on suffixation (at least from the literature I have reviewed) to form different forms of a given word. Those morphemes that are used to represent a prefix and prefix-suffix pairs in other languages are all represented by a suffix in the case of Wolaytta. For example, in the Amharic word ከ-ኢትዮጵያ 'from Ethiopia', the prefix {ከ-}'from' is prefixed to the word ኢትዮጵያ 'Ethiopia'. Its equivalent in Wolaytta is *Ethiopia-ppe* where *{-ppe}* is the suffix which represents the Amharic prefix {ከ-}. Another example can be seen by considering the Amharic word ከ-ኢትዮጵያ-ም 'also from Ethiopia'. In this example the prefix { ከ } and the suffix{-ም} occurred at the same time creating prefix-suffix pair ከ-ም. Its equivalent for Wolaytta is *Ethiopia-ppe-ka* where the suffix *{-ppe}* is equivalent to the prefix { ከ-} and the suffix *{-kka}* replaces the suffix{ ም}.

As pointed out by Nega (1999), the actual affix compilation is dependent on the nature of a stemming algorithm, *longest match* or *iterative*. The longest match algorithm requires all forms

of the affixes, basic[1] and derived[2] (or concatenated), for successful stemming. Where as, the iterative approach only requires a list of basic affixes and removes them iteratively (Nega, 1999).

Concatenation of suffixes is common in Wolaytta. As a result, two or more suffixes may be concatenated together and attached (or affixed) to a word. In the language, possible list of combination can be very large making difficult to have complete list of combination (concatenations). Besides, concatenation in the language makes suffixes long ones attaching one suffix to another. Hence, iteratively removing each base suffix one by one is considered the best choice. As a result, iterative approach is adopted to develop the stemmer for Wolaytta text.

In this study, a semi-automatic means is employed to compile the possible suffix list (word endings). In the process of compiling the suffix dictionary, the words in the sample text were first written in reverse order. The reversed list of words was then sorted and frequencies of matching sub-strings identified. Finally the sub-strings which occur more than once are selected as suffixes. The whole list of suffixes identified using these procedures is given in Appendix II. Table 4.3 below lists a total of 98 genuine (or linguistically correct) suffixes found in the language that are obtained from whole list of possible suffixes for the selected sample text.

Table 4.3. Linguistically valid suffixes obtained from the sample text

iya	garsa	ga
uwa	ppe	ibo
iyu	garsappe	go
yu	ra	gato

[1] basic suffix is minimum form of suffix that can not be decomposed into more suffixes
[2] derived suffix is a suffix formed by concatenation of two or more base suffixes.

The way how the algorithm works is explained using example as follows. First, a word is inputted to the stemmer. Let the stemmer get the word *zigiribennageta* 'we/you /they who did not slander'. In this word, we get four suffixes *{-ibe}*, *{-nna}*, *{-ge}* and *{-ta}*. The stem is *zigir* 'slander'. After getting this word, the stemmer opens the stopword file and check if the word is available in the list as stopword. If so, the procedure leads to the fifth step where another word is inputted to the stemmer provided that End of File (EOF) is not reached. The stemming process is stopped if end of file reached. The word *zigiribennageta* does not exist in the stopword list. This leads to the third step where suffix file is opened to check whether a suffix from the list is attached to the inputted word. Assume that the four suffixes listed above are kept in the suffix dictionary in the order they are presented above. When reading suffix from the suffix dictionary, the stemmer first gets the suffix *{-ibe}*. Even if this suffix is available in the word *zigiribennageta*, it is not located at the end of the word. With out stripping this suffix, it reads the next suffix to check against the word *zigiribennageta* for match. What has happened for the suffix *{-ibe}* is also true for *-nna* which is the second suffix in the list. The same is true for the suffix *{-ge}*. Since the suffix *-ta* is available at the end of the word under consideration; it will immediately be removed resulting with the word *zigiribennage*. Accordingly, the other three suffixes in the example word are removed one by one and finally results in the stem *zigir*. Then, this leads to the seventh step where the resulting stem *zigir* will be considered as a new word and a recursive process goes on starting from step 3 where suffix file is opened and checked for the existence of a suffix in the inputted word, i.e, *zigir*. Next to this, if none of the suffixes in the suffix dictionary exist in the inputted word, then it leads to step six where conditions are checked for necessary action on the resulting stem (*see section 4.7 below for detail*). After application of necessary condition, if any, on the finally identified stem the word is considered as a stem and

recorded in the stem dictionary. The procedure continues reading next word, EOF not reached, for the same stemming task. Unfortunately none of the conditions set for this stemmer are applicable in the resulting word (stem) *zigir*. Finally, this word is recorded in the stem dictionary. IF EOF not reached, it accepts new unstemmed word, stops processing otherwise.

4.7 CONDITIONS/RULES CONSIDERED BY THE STEMMER

The stemmer developed for stemming Wolaytta text is context sensitive. The stemmer is made context sensitive mainly to get better performance result. Two context-sensitive actions are employed in the stemmer developed. These are:-

Action 1. Remove characters ee

Action 2. Replace by characters y

The following are the conditions checked in order to take anyone of the actions, action 1 or action2. These are:

Condition 1. After suffix removal, if number of radicals is one in the remaining stem and the radical (Consonant) is followed by characters ee. (In this case action 1 will be taken by the stemmer).

Condition 2. If characters 'sh' are part of identified stem word. [In this case action 2 will be taken by the stemmer].

The following two examples illustrate how the algorithm works with respect to the above conditions and necessary actions following them.

4.8. EVALUATION OF THE FIRST STEMMER

The stemmer was trained on the 3537 words (80%) selected for training purpose. Out of these words 14.6% (516 words) were understemmed and 3.2% (113 words) were overstemmed. Totally this version of the stemmer generates 17.8% (629 words) stemming error. As a result, the accuracy of the stemmer becomes 82.3%.

In order to improve the performance of the stemmer, problems of the stemmer in this first version were first investigated. Accordingly, the following problems were identified:

i. The suffix list used by this version of the stemmer was kept initially in ascending order of the number of characters. Some of the Wolaytta suffixes appear alone as well as being part of another suffix. For example, in the word **tara** 'with me', there seems two suffixes {*-a*} and {*-ra*}. Especially the later one {*-ra*} is not considered as a suffix if *-a* is removed in the first iteration. In the suffix list, if {*-a*} comes before {*-ra*} then we get then the stem *tar* which has no meaning at all in the language and considered as understemmed. This was the main problem for the majority of the problems faced by the first version of the stemmer.

ii. There were suffixes that were not included in the suffix list. As a result, words that should have been conflated to same stem are not done so. This is because those words that possess a suffix which is in the suffix list are reduced to their stem whereas those to which no suffix exists are left as they are.

iii. There were words that should not be conflated to the same stem but stemmed to the same stem.

iv. In some cases, character(s) that should not be removed are done so because the character(s) exist(s) in the suffix list.

In terms of compression, i.e., reduction of dictionary size, percentage of compression is calculated using the formula (Girma, 2001):

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

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 = 3537
- Number of stems = 2311

Hence, the percentage of compression for Wolaytta text based on the training text for this initial stemmer becomes 100 * (3537- 2311) / 3537 = 34.67%.

4.9 IMPROVED STEMMER

To solve the problems identified on the first version of the stemmer, the following minor enhancements were introduced to the algorithm.

i. To solve the problem described under section 4.9(i), the suffix list is arranged in descending order of the number of characters.

ii. To solve the problem described under section 4.9(ii), more suffixes which were not previously available in the suffix list are included manually.

iii. To solve the problem described under section 4.9(iv), the following new condition and respective action are introduced.

Condition 3: If the number of radicals in the remaining stem is 1 followed by character a or aa.

Action 3: Don't remove character a or aa.

The improved stemmer looks the following:

1. READ the next word to be stemmed
2. OPEN stopword file
   Read a word from the file until match occurs or End of File reached
   IF word exists in the stopword list
   Go to 5
3. OPEN suffix list file
   IF number of radicals of word is greater than 1
   READ a suffix from the file until End of File reached

The test prevails that 9% (79 words) are understemmed and 4.1% (36 words) as overstemmed which makes the accuracy 86.9%. The distinct stems after conflation are 540. Accordingly, the dictionary size of the test set is compressed by 38.92%.The main cause for most understemmed words is the fact that few additional suffixes (suffixes not listed in the suffix dictionary) are identified.

In general, reasons for the understemming and overstemming problems are:

1) It was difficult to come up with the complete list of suffixes because of the complexity of the language.
2) More conditions/rules are required based on a detailed study of the morphology of the language.

4.10 CHAPTER SUMMARY

A context-sensitive iterative stemmer is developed in this study for Wolaytta text and it is capable of conflating word variants of same words. Sample of 884 words were used as a test set to evaluate the performance of the stemmer and the result shows 86.9% accuracy. For the same set of words, a compression of 38.92% was found. The next chapter presents conclusions of findings and recommendations for future research.

CHAPTER FIVE

CONCLUSLONS AND RECOMMENDATIONS

5.1 CONCLUSIONS

Wolaytta is a morphologically complex language. A single word in the language has a number of variants. The language is rich in both inflectional and derivational morphologies. The language uses suffixation technique to create variant words and suffixes are concatenated one after the other to create another variant of the word. In this way the length of a word can be very long and some times a single word in Wolaytta is equivalent to a sentence in English. The morphological complexity of Wolaytta language is witnessed by word distribution ratio, as stated in Adams (1983); Lamberti and Sottile, 1997), and the possibility of forming large number of word variants from a single base word (see appendix III). The analysis of word ratio of total words to distinct words calculated from sample text also shows that Wolaytta is morphologically complex language than English but less complex than three local languages such as Amharic (Nega, 1999), Affan Oromo (Wakshum, 2001) for which stemmer has been tried. Therefore, it is time taking and cumbersome, if not impossible, to conflate words manually for Wolaytta. This concatenation of suffixes is highly common in the language. Most function words are assimilated as part of content bearing words in the form of suffixes. This property of the language has its own role in the morphological complexity of the language.

In the study, it is found that iterative approach is more appropriate for developing the stemmer for Wolaytta 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) and the varying nature of order of concatenation.

The stemmer developed uses a context sensitive conditional rules. The evaluation for the final stemmer reveals that there is significant difference between stemming and non-stemming for Wolaytta. The modified stemmer (improved or final stemmer) was run on the training data. Accordingly, the number of understemmed and over stemmed words were 8.6% (304 words) and 0.8% (28 words), respectively. The total errors account for 9.4% (332 words) and the performance of the stemmer is improved to 90.6%. In terms of dictionary size, the compression became 41.2%. The result shows that the accuracy of the stemmer on the test set (unseen part of the sample text) is 86.9% with only 9% and 4.1% understemming and overstemming errors respectively and with 38.92% data compression.

As indicated by Tesfaye (1987), conflation algorithms have inherent limitations and certain linguistic problems that are common to all conflation algorithms, irrespective of their ultimate use. The general assumption is that if two words have the same underlying stem then they refer to the same concept. However, this is not always the case since some times words of the same stem need to be distinguished while words which are essentially equivalent may mean different things in different contexts. Thus, it is inevitable that such systems will produce errors. What has happened for Wolaytta text in this context is also not far different from this fact. Even if the stemmer is improved to eliminate those errors occurred in the first trial of the algorithm, the final

and improved algorithm has also shown few errors by combining words together that should not be brought together and by separating words that should not be. However, experiments have shown that the proportion of errors caused by these circumstances does not lessen retrieval effectiveness too much (porter, 1980; Lovins 1968).

Few context sensitive rules are included in the algorithm. But it seems that the language requires more context sensitive rules for more effective conflation (this needs further research).

5.2 RECOMMENDATIONS

This stemmer is the first trial of its kind for Wolaytta language. It is my belief that the stemmer should be improved by further research to attain better performance and hence bring it to an operational level. Accordingly the following recommendations are suggested:

- Come up with a proposal for developing an operational stemmer for Wolaytta language.
- Compilation of corpus useful for Natural Language Processing of Wolaytta in general and development of a stemmer in particular.
- The procedures followed to develop the Wolaytta stemmer can be used as a source for developing stemmers for other Omotic family languages.
- Longest match approach can be implemented to see whether it performs better than the approach used in this research or witness the approach used in this research work.
- One can add more context sensitive rules (formulate it) in order to increase the accuracy of the stemmer tried in this research work.
- After improving the algorithm to its appropriate level, the stemmer can be an important tool for those researchers who are interested to study the Wolaytta language morphology.
- By incorporating necessary elements, the stemmer can also be used as a component for developing other computational tools like morphological analyzer, parser, spell checker, thesaurus, word frequency counting and the like of the language under consideration.

BIBLIOGRAPHY

Adams, B.A. (1983). *A Tagmemic Analysis of the Wolaitta Language*. Ph.D Thesis. University of London (unpublished).

Ahmed, el al. (1996). Experiments with Stemming Algorithm for Malay Words. In *Journal of American Society for Information Science* 47(12): 909-918.

Al-Kharashi, Ibrahim et al (1994). "Comparing Words, Stems, and Roots as Index Terms in an Arabic Information Retrieval System." In *Journal of American Society for Information Science* 45(8): 548 - 560.

Alemayehu, Nega and Willet, P. (2002). *Stemming of Amharic words for Information Retrieval.* In Literary and Linguistic Computing. 17 (1): 1-17.

Baeza-Yates, Richardo. (1999). *Modern Information Retrieval*. New York: McGraw Hill.

Bender, M. Lionen , et al. (1976). *Languages in Ethiopia*. London: Oxford University Press.

Central Statistics Office. (1999). *The 1984 EC*[1] *Population and Housing Census of Ethiopia: Analytical Report at National Level*, Addis Ababa, Ethiopia.

[1] EC Ethiopian Calander

Croft, W. Bruce and Xu, Jinxi (1995) *Corpus-Specific Stemming using Word Form Co-occurrence* at http://citeseer.nj.nec.com/croft95corpusspecific.html

Dawson, J.L. (1974): *Suffix removal for word conflation* at http://www.comp.lancs.uk/computing/reseach/stemming/general/dawson.htm

Ekmekcioglu, et al . (1996). *Stemming and N-gram Matching for Term Conflation in Turkish Texts* at http://informationr.net/ir/2-2/paper13.html

Frakes, William B. (1992). Stemming Algorithms. In Frakes, William B. and Baeza-Yates, Richardo, eds. *Information Retrieval: Data Structures & Algorithms.* New Jersey: Prentice Hall PTR.

Girma Berhe. (2001). *A Stemming Algorithm Development for Tigrigna Language Text Documents.* MSc Thesis. (Unpublished.)

Harman, D. (1991). "How Effective is suffixing" In *Journal of American Society for Information Science* 42(1): 7 - 15.

Hmeidi, Ismail et al. (1997). "Design and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents." *In Journal of American Society for Information Science.* 48(10): 867-881.

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

Hull, David A. and Grefenstette, Gregory (1996) *A Detailed Analysis of English Stemming Algorithms* at http://citeseer.nj.nec.com/hull96detailed.htm

Kosnov, Serhiy. (2003). *Evaluation of N-grams Conflation Approach in Text-Based Information Retrieval* at http://www.yslab.ceu.hu/~serge/ir2001/ir2001_kosinov_s.pdf

Kraaij, Wessel and Pohlmann, Renee (1996) *Viewing stemming as Recall enhancement* at http://citeseer.nj.nec.com/kraaij96viewing.html

Krovetz, R., (1993). *Viewing morphology as an inference process* at http://www.comp.lancs.ac.uk/computing/research/stemming/general/krovetz.htm

Lamberti, Marcello and Sottile, Roberto. (1997). *The Wolaytta Language*. Koln: Rudiger Koppe Verlag.

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

Lovins, J. B. (1968). *Development of Stemming Algorithm*. Cambridge: Electronic System Laboratory, MIT.

APPENDIX I: Stopwords compiled from the sample text

taagaa

neegaa

aagaa

iigaa

nuugaa

hiinteegaa

eetaagaa

hiinteero

issiippe

taani

uray

nuuiyu

garsa

aakko

ideta

garsana

garsan

aana

taayoo

taassi

neeyoo

neew

taana

taani

naee

banta

mati

eeti

hagani

ii

hegatuu

heggeti

nuu

eetaa

henna

taa

nee

hegetane

hegeti

hanna

hage

he

eeta

taa

hinte

issi

ha

daroo

guyaani

ubbagallas

ubbawode

ubbayoho

cora

nuuyoo

issoy

aybaakka

issibaanne

happuni

happunee

awni

awayba

ani

aymalaa

APPENDIX III: Examples of a stem and its different variants

The following is just a list of few of the variants of the word *kall* 'follow'. All the variants have meaning related to the base word *kall*.

kallasi	'ተከተለ'
kallokkona	'አይከተሉም'
kalla	'ተከተል'
kalladasa	'ተከትለሃል'
kallane	'ትከተላለህ? '
kallideta	'ተከትላችኃል'
kallisi	'ተከትሏል'
kallida	'ተከትለናል'
kalluku	'አትከተልም'
kallasu	'ተከትላለች'
kallidosona	'ተከትለዋል'
kallabeikke	'አልተከተለችም? '
kallabeikka	'አልተከተልክም'
kallabeikku	'አልተከተለችም'
kallibenna	'አልተከተለም'
kallibokko	'አልተከተልንም'
kallibekketa	'አልተከተላችሁም'
kallibokkona	'አልተከተሉም'
kallidana	'ተከትያለሁ?'
kalladi	'ተከትለሃል? '
kallade	'ተከትላለች?'
kallide	'ተከትሏል? '
kallido	' ተከትለናል? '
kallideti	'ተከትላችኃል? '
kallidona	'ተከትለዋል? '
kallabeikkina	'አልተከተልኩም? '
kallabeikki	'አልተከተልክም? '
kallabekke	'አልተከተለችም? '
kallibenne	'አልተከተለም? '
kallibokkoni	'አልተከተሉም? '
kallibekketi	'አልተከተላችሁም? '
kallebekkona	'አልተከተላችሁም? '
kallana	'እከተላለሁ'
kallikke	'አልከተልም'
kallakka	'አትከተልም ለወንድ'
kallukku	'አትከተልም ለሴት'

kallenna 'አይከተልም'

kallokko 'አይከተሉም'

kallekketa 'አትከተሉም'

kallokkona 'አይከተሉም'

kallute 'ትከተላለህ? '

kallikkina 'አልከተልም? '

kallikki 'አትከተልም? '

kallekke 'አትከተልም? '

kallenne 'አይከተልም? '

kallokkoni 'አንከተልም? '

kallekketi 'አትከተሉም? '

kallokkona 'አይከተሉም? '

kallisi 'ተከትሏል'

kallasa 'ተከትለሃል'

kallausu 'ተከትላለች'

kallesi 'ይከተላል'

kallosi 'እንተላለን '

kalleta 'ትከተላላችሁ'

kallosona 'ይከተላሉ'

kalliya 'እንከተላለን '

kallyana 'እከተላለሁ?'

kallai 'ትከተያለሽ? '

kalli 'ይከተላል? '

kalliyo 'እንከተላለን? '

kalleti 'ትከተላላችሁ?'

kalliyona 'ይከተላሉ?'

kallu 'ትከተል'

kallo 'ይከተል'

kallite 'ተከተሉ'

kallona 'ይከተሉ'

kalloppa 'አትከተል ለወንድ'

kalluppu 'አትከተል ለሴት'

kalloppo 'አይከተል'

kalloppona 'አይከተሉ'

kallennani 'ሳይከተሉ'

kallini 'ሲከተሉ'

kallaidda 'ተከተልን '

kalliddi 'እየተከተልን '

kallisini 'ስንከተል'

kallennasini 'ሳንተከተል'

kallidasine 'እንደተከተልን '

kallikko 'ከተከተልን'

kallanatayikko 'ካልተከተልን '

kallaiddakka 'እየተከተልኩም'

kallennanka 'ሳልከተልም'

kallikkonne 'ቢከተልም'

woottadasa yeekkiyoogaappe yootadi aaggaa acca aliyaa asee aseeppe asiyu awgeta aybaani aybee azeloosona bade baanne balaagoosoona barayyiya baysi beanaaw biidaakko biidoogaappe bitaanney bitanniyaa-ssi dandayaysi deesshiya dendidi dooyyidaakko eetaassi ekettiisi ekkade fatuumaa fuuttuwa gaadiyaani garsaappe gelisoona giddooni gideennaga gidoppa guyeera haasayey hanidooga haykkisoona heeya hegatuuppe hegeti heyrayso hiinttero hiintena hinteero hupiyaappe hupiyaaw iissi imattay immokkoshini issuwaassi ittikko kaanteese kamma keettaani kiyisoona kooyyaasa kuundiisi maaddeennaani maanaagadani maasa mara mata mataafati meeccaasi miidaakko miisshay miiziya naaw naeera neegaa neeppe nepee ola oollaappe oonne oottadi ootteennaani oyddata shaaya

b)Stemmed text (excluding stopwords)

imm mar er heer isha maskoot mat oosuw taar yoot meecc oott talaal micc oott tamaar miid oott tarappeez aara oott m toho miissh oydd tooss aay aas miitt shaaf ufay aas miiz shaay ura aay naa sham ura naa shamm usuppun adukk naa shimm usuppun adukk na shoc waan agn na shor waan shucc ali shucc wont asa sind wor asa neer siy wott asa siy yassh asat nuu soh y asa ol soo yeess ashkar ooll suwik y assi bana bitannetu bar bitann bar bitann b od bayn b og ayb bayz b ol ayb bayzz boogaanc omars ayba bayzz daabb omars ayb bead daabb oog ayba b daabb ookk ayb bei daabdaabb ookk aybaykka bey daabdaab ookk ayll bitanne daann ooll bid daar y mataaf asa taa n ba eet kuma taa laagg yoot ta y haatt keett laagg hageet naati keett laagg macc ciinc ubba abeba tee naa taan garsa gela guy kooyy wor bi biy biis eh gela laagg na aduss dylleet baasu bitann deo keett oyta ayll eh hegg koyro naa yoot ayb booll gar hann im imm naa oona yiida awde ayll daabdaabb kitt mar wod bad be de horaatta imm keett maadd mayy na shor uyy y yoot awg ayb b ciinc eh ekk ekk gelaa gel heer hiint kaam laagg maadd mat na og saabb shamm shucc sukkaar taay waatade woott y aliy as ashkar baa b bitann b dar dees dooyy ekkis ere fakkad garad gelaaw ha happun gaap imm ishol kar kiy kooyy laagg maadi macc mataaf y miiz naa oydd shamm shoc woott yeekk yoot acc alak aliy asee asee as azel bad baann balaag barayy b b b bitaann bit danday deessh dend dooyy ekett ekk fatu fuutt gaad gars gel gidd gid gid gu haas hanido haykk he he hup hup imatt imm iss ti kaant kamm keett kiy kooyy kuund maadd maa maa ma ma mataaf meecc miid miissh miiz naa nee ooll oott oott oydd shaay

APPENDIX V: The test set

issi guutta ne ha mataafa asati taaw neeni ba eeti kuma taani laaggeti yoota tani yissi haatta keetta laaggey hageeti naati hagani keettata laaggeta macciya ciinca ubba abeba neni teega hara naa taana aaw garsa gelaayu guyaani kooyyeese woriya biisi nuuna aw biisoona ehida gelaaya laaggiya naay aappe adussaa-nne dylleeta baasu bitanniya deosoona keettati oytta aylliya ehaasu heggeti iippe koyro naatura nuni oo yootaasi aybaaw boollaani gadiya garsana hanninti imam immiisi naata oona yiida awde aylliyu daabdaabbiya kitta marata neessi wode badasa beidaakko deiyuuna ehiisi horaatta immaysi keettay maadde ennaantaikko mayyuwa naiya shorotu uyyiisi wookku yiidaakko yootiisi awgetee aybaassii beabeykke wursetta gelaawta beabeykke gosshaancay wursetta naakko biisi issi issi naaassi heezza heezza imma hacci issi issi asaassi tammu tammu bira immaysi ii guutta kuma ba laaggetura issiippe moogaappe guyaani taani soo ta macceekko baasi nuuni taaran taaran yiite issoy issoy yiite eeti naay naay gelisoona eeti naati naati gelisoona hiinte naati hiintee nanaati nuuni naati nuuna naata nuuni oyddata nuuna oydda ti nuuna oyddata hiinte iccashati eeta usuppunata he usuppunati heggeta usuppunata eeta heezzatakka beaasi ha ogiyara ya baade maammada bea baade maammada kooyya neew guutta ashuwa shamma hupiyaassi guutta ashuwa shamma hegara ba hagan nanga asati ubba gallas maakkinaani addis abeba boosoona asati ubba wode maakkinaani addis abeba boosoona asati ubba gallas kaamiyaani addis abeba boosoona asati maakkinaani addis abeba hai boosoona asati kaamiyaani addis abeba hai biisoona baa biida ciincoosoona ehay ekka ekkiisi gelaawota geleese heeraappe heggani hiinteeyoo immanashini kaamiyaani laaggetura maaddiyakko mati naatuussi nayu ogiya saabbida shammideta shuccata sukkaariya taayoo waatade woottiisi yaasu yeessaasu yohuwa aliyaara aseessi ashkareti awni baawa bay bare beabeeyekke bitanniyaappe booppa dareese dees dooyyaasu eetaara ekkisoona ereeti fakkadiisi garadiy gelaawiyu gita hai happuneekko hegaappe hiinteena iikko immanaagashini isholatti kare kiyidooga kooyyaakko laaggee maadi macceessi mataafata miizata naassi naeekko nuuyoo oow miidde deedooga koppiisi ta ootteennaani baasi ne oottabeennaaga ii koppiisi ne oottanaaw dandayabeyakka ne taana maaddikko mataafa neew immana ne taana maaddikko mataafa neew immaysi ta neena maaddikko mataafa taaw immana ta neena maaddikko mataafa taaw immaasa ii o maaddikko mataafa aaw immana ii o maaddikko aa mataafa aaw immawsu nu hiintena maaddikko hiinte nuu

mataafa immeeta ta neena maaddiyakko ne taaw mataafa immanashini ne taana maaddiyakko ta neew mataafa immanashini hiinte nuuna maaddiyakko oyddu shammidooga shociisi taagaa taara tuma woottadasa yeekkiyoogaappe yootadi aaggaa acca alakaassi aliyaa asee aseeppe asiyu awgeta aybaani aybee azeloosona bade baanne balaagoosoona barayyiya baysi beanaaw biidaakko biidoogaappe bitaanney bitanniyaa-ssi dandayaysi deesshiya dendidi dooyyidaakko eetaassi ekettiisi ekkade fatuumaa fuuttuwa gaadiyaani garsaappe gelisoona giddooni gideennaga gidoppa guyeera haasayey hanidooga haykkisoona heeya hegatuuppe hegeti heyrayso hiinttero hiintena hinteero hupiyaappe hupiyaaw simmidde ba shoruwaa naeera ogiya boollaani gaayettiisi ii soo simmidde bare shoruwaa naeera ogiya boollaani gaayettiisi ii miiziyu bayzzidi soo simmidde ba shorotu macca naatura gaayettiisi ii miiziyu bayzzid soo simmidde bare shorotu macca naatura gaayettiisi ii miiziyu bayzzidoogaappe guyaani soo simmidde ba shorotu maccia naatura gaayettiisi ii miiziyu bayzzidoogaappe guyaani soo simmidde bare shorotu macca naatura gaayettiisi ii miiziyu bayzzidoogaappe guyaani soo simmidde ba shorotu macca naatura ogiya boollaani gaayettiisi taani suwikaydde baysi aa suwikaydde baasu ne oottade badasa ii uyyide zineese taani oottidooga ii koppiisi ne oottishine ii koppiisi ii oottide deishine aa koppaasu aa oottanaaga koppiisi hiinte oottanaaw kooyyideta ta baada wursaasi nu miiddi uttida ii nee ni miidooga koppiisi ii ii miidde deedooga koppiisi ta ootteennaani baasi ne oottabeennaaga ii koppiisi ne oottanaaw dandayabeyakka ne taana maaddikko mataafa neew immana issi imattay immokkoshini issuwaassi ittikko kaanteese kamma keettaani kiyisoona kooyyaasa kuundiisi maaddeennaani maanaagadani maasa mara mata mataafati meeccaasi miidaakko miisshay miiziya naaw naeera neegaa neeppe nepee ola oollaappe oonne oottadi ootteennaani oyddata shaaya shammiisi shocida shuccaara sindiya siyida sooppe taafaysi taageta talaali tarappeezati toossa uraappe usuppun waanade weysa woriisi yasshay yeessisoona yiidoona yiite aakko aara aasa aayoo adukkawsu agnaappe aliyow asakka asatoow ashkaret attadasa atteese awga awunna aybaakka aybaaranne aybaw aybaykka aymalaa baadi baanta baggaani baggati baha barakka bay bayzettaasu bayzzidi beay beykke bidoogaappe biideta bite biossappe birshaasi bitanneta bitanniyaaw biya biye daabbiya daabbuwokka daabdaabeeta daarati daddosoona dandayabeyakka dandayeese deana deay deeenna deenna deennadani deeshati deesshataaw deesshatuussi deessheeyoo deibeenna deiis deisoona demmida demmosoona deoppo desshataayoo dussaanne duummare eet eettappe ekettide ekkeda ereesi erettosoona erokkoona essaasi fakkadeese fengiya fulaattawsu gaattanokka gakkanaashi gatimaani gediyaani gelaawata gelaaweeppe gelide

gidabeikke gidanaaw giddo gidibeekkoona gididosoona gidiyakko gidokkoona gidoppite giishini gisshi gooday gosshaancaassi gosshaancayoo gosshaancey haasayiisi haasayoppa hagaadani hagana hakime hanana hanibeekketa hanidaakko hanite hanno hanoona happuni haykkaasu haysiya heeraaakko heezzatukko hegeta heggeta hiinteera hiinttegaa hinteegeta hupee iccashati iigaa iiro iiyoo imattaaw imattaw immaasa immakkashini immeeta immokkoonashini ishatey issuwaayoo kaamma kaantaasu kaantosoona kaawoy kamiyaani kariya kattaappe keetaatuuppe keettaaway keettatuni keettidey kessosoona kiyanaw kiyeese kiyoppo koffaysi koossheese kooyya kooyyayi kooyyideta kooyyosoona koppaasu koyiide kumaani kuttuwa kuundisoona laaggeeti laankiya laggeti looabeekku looanaaw looppa maaddeennakko maahe maala maattaanne naccciya macceera masmuriya mentide miidde miidooga miitettaasi miizataaw miizati miizeetuussi miiziiyu minuweese mishireeyoo missha mittaakko mizade naaassi naantona naeeyoo nagaasa naiyu nayo neenara nuppe nuukko nuunara nuuro ogiyara omarsaakko oogey ookkokka oolla oonanne oonaranne ooppe oossanne oottabeennaaga oottade oottanaaga ootteenna oottidaakko oottidooga oottishine oraatta oycciisi saatiniyaani shaafaakko shammaasu shammaney shammida sharmuuta shoceeta shoosshaa shoosshatta sinaani siyiisi soon sooni taafida taaafishini taakko taro talalasi tamaaree-soo teegiisi tiillee tohuwaani tuuggaasu ubbaykka uraassa usshaasu uttida uyyana uyyaysi waanidi waasani wara waraasu wootta woottadi woottideta wottaasu wuriya ya yaadasatte yaade yaasitte yayyeennaani yeddade yeenna yeessaasi yeesseda yeessideta yekkey yohaannisaappe yuushuwa zineese

azinay guutta kuma hinte deesshataaw efiisi aylliyey eeta sooppe kiyeese asati baanta keetaatuuppe kiyisoona asati eeta keettatuuppe kiyisoona miitta acca mitta mata keetta acca keetta mata mitta laankiya keetta boollaani mitta boollaaani ogiya boolaani keetta garasa keetta garsan keetta geliisi soon woottiisi soo garsani woottiisi ii keettaappe oollaakko biisi ii keetta garsaappe oollaakko biisi keetta yuushuwa keetta gupiya mitta gupiya he asatu sintaani ekka mitta garsa mittaakko gakkanaashi heeraakko biisoona heera bagga biisoona keetta garsaappe keettaappe karee bagga karee bagga satiniyaani taappe sinta aylliya gaadiya kaantade baasu maskootteera teella yohaannisaappe guyeera kattaappe guyeera kammaakko omarsaakko omarsay gakkanaasha heezzu agnaappe ta gishata ne gishata a gishata ii gishata ta laaggiya gishata asatu giddo asatu garsa miissha gaasoota nu gaasoota ta gaasoota ne gaasoota a gaasoota ii gaasoota tammu dakiika garsa olaa wode asa oliisi taara ekettiisi taanara ekettisi ii taanara ekettide kaiisi ii asnay gediyaani biisi ii asnay tohuwaani biisi imattay kamiyaani biisi a macciya kulfiyara fengiya dooyyaasu a macciya kulfiyara ifitta dooyyaassu a macciya kulfiyara kariya dooyyaasu ta samunaani mayyuwa meeccaasi yasshay bayeennaani yasshay yayyeennaani neenara gideenna ne bayeennaani neeni bayeennaani miisshay bayeennaani ta ta hupee ta ta talaali ne ne talaali ne ne talaasi ne duummare ii duummara eeti ubbay nu ubbay nuuni ubbay nuuppe baggati nuuppe guuttati eetaappe guuttati hiinteppe guuttati ooni yiide ooni bii oona beadi neeni oow yootadi neeni ooyoo yootadi neeni oossi yootadi haatta oow ehay haatta ooyoo ehay neeni haatta oossi ehay haatta neeni oossi ehay ookko baadi oow gadiya shammadi ooyoo gadiya shammadi oossii gadiya shammadi oonara yaadi ooppe guutta miissha ekkade hagee oossi hagee oogey hegee aybee aybee hanide aybaa maadi ne aybaa koyyey aybaa garsani woottadi aybaa baa baa haasayey aybaaw haasayey aybaaw yekkey aybaaw yaadi aybaayoo yaadi aybaassi yaadi waanade baay waatade baay waatade oottadi waanade oottadi waatade maadi waatidi oottanadan kooyyey waanidi oottanadan kooyyey aymalee aymalaa deay awde yiide awde yeey awde hanide awni hanide ii awppe mataafa shammaney aw baay aw maadi awppe yeey wookku oytta kooyyey wookku sukkaariya nuuni kooyyu appuni laaggeti neew deiyuuna appuni laaggeti neeyoo deiyuuna appuni laaggeti neessi deiyuuna appuni isholatti iiyu deyuuna appuni isholati iiyoo deiyuuna appuni isholati iissi deiyuuna eeti appuni mataafati ooni yiidaakko taaw yoota ooni biyaakko aaw yoota eeti aybaa beidaakko taaw yoota aybaa beidaakko taaw yoota eeti oona beidaakko taaw yoota oona ii azinay guutta kuma hinte desshataayoo eflisi ii azinay guutta kuma hinte deesshataaw efiisi aylliyey eeta sooppe kiyeese

asati baanta keetaatuuppe kiyisoona asati eeta keettatuuppe kiyisoona miitta acca mitta mata keetta acca keetta mata mitta laankiya keetta boollaani mitta boollaaani ogiya boolaani keetta garasa keetta garsan keetta geliisi soon woottiisi soo garsani woottiisi ii keettaappe oollaakko biisi ii keetta garsaappe oollaakko biisi keetta yuushuwa keetta gupiya mitta gupiya he asatu sintaani ekka mitta garsa mittaakko gakkanaashi heeraakko biisoona heera bagga biisoona keetta garsaappe keettaappe karee bagga karee bagga satiniyaani taappe sinta aylliya gaadiya kaantade baasu maskootteera teella yohaannisaappe guyeera kattaappe guyeera kammaakko omarsaakko omarsay gakkanaasha heezzu agnaappe ta gishata ne gishata a gishata ii gishata ta laaggiya gishata asatu giddo asatu garsa miissha gaasoota nu gaasoota ta gaasoota ne gaasoota daabbuwokka miisi ubba haatakka uyyaasu issiippe biida asay issiippe yiisoona eeti ubbaykka hai nagosoona ne maanaaw kooyyaasa maanaaw kooyyayi aley soo biye neew geleese neew geliye neessi geleese neessi geliye nau attuma naati nau attuma naati yiisoona taani nau naata beaasi heezzu asati heezzu asata he heezzu asata hegeti heezzu asata oyddu macca yootaasi appuni deosoona aybaaw gelaawta banta keettati boollaani beadi malaanca gaayettiisi biisoona oytta gadiya eetaappe taappe garsan ehida aylliya garsana gelaawiya dees gelaawti gelaaya ehaasu hanninti hupiya hanna heezzu laaggiya heggeti imma maaddeennanitayakko hiinteeppe immeennaagashini naay iippe immiisi shammaasi iita maaddikko simmidde gishata awgeti yiide horaatta aybaassi adukkoosoona iccashu aylley awde immaysi baa aybaa issoy bayeennaani aylliyu keettay biida mayyuwa fuuttuwa oottidde shimmaynneti mayyuwa daddosoona yayyiddeenne kokkoriddeenee oossha haasayiisi ta ta macceera ooyetto gisshaassi sohuwara oo birshaasi ifitta mentide ta keetta garsa gelide boogaancati eeti demmida ubba miisshaka efisoona taani ne oraatta baa siyido gisshi maalalettaasi macca asati sindiya gaaccidde tiillee kessosoona macca asati sindiya gaaccidde aappe tiilliya demmosoona nuuni shaafa biidoogaappe guyaani bitanney taassi tuma yootiisi nuuni shaafaakko biidoogappe guyaani bitanney taassi tuma yootiisi nuuni shaafa bidoogaappe guyaani bitanney taaw tuma yootiisi ii hagaadani giishini miiziya sohuwara bayzettaasuazinay loonne biitta daabdaabbiya maaddeennaantaikko ciincoosoona gaasoota maakkinaani eetaaw kitta mayyuwa ehay koppiisi naanto ehiisoona marata naiya ekka naya oossi ekkaasu neessi shorotu ekkiisi tarappeezay tarappeeza gallas wode uyyiisi gelaawota aley wigiyaani gelaawoti badasa wookku geleese naatura ziniisi ta laaggey daroo macca naatura ziniisi ta laaggey heyrayso macca naatura zi niisi ubba gallas taani cora daabaabbiya taafaysi ubba gallas taani daroo daabdaabbeta taafaysi hara bitanney yiisi hara asay yiisi hara macca

efaasu ii hiinte garsana ii hiinte garsan deees hinttera yiisoona hiinteera issiippe yiisoona hiinteenara yiisoona hiinteenara issiippe yiisoona gosshaancey hiinteeppe guutta miissha ekkiisi hiinte eeta ereta ashkareti eetaaw guutta mitta ayfiya ehisoona ashkareti eetaayoo guutta mitta ayfiya ehisoona ashkareti eetaa ssi guutta mitta ayfiya ehisoona aa macciyta wottiba eetaaw efaasu aa macciya eetaayoo efaasu aa macciya oottiba eetaassi efaasu ashkaret mitta ayfiya ekkogaappe guyanni yiisoona ii eeta garsana eeta garsan deees ii azinay eetaara yiisi ii azzinay eetaara isippe yiisi ii keettaaway eetaara issiippe yiisi ta tani guutta maattaanne oytta eetaappe ekkaasi asay miisi issi asay miisi asay aliyaara uyyiisi asay alliyaara issiippe uyyiisi issi uray aliyaara issiippe uyyiisi banta giddooni erettosoona eeti banta giddooni imota imettosoona neeni gosshaancaaw guutta miissha saatiniya garsani woottadasa neeni gosshaancayoo guutta miissha saatiniya garsani woottadasa neeni gosshaancaassi guutta miissha saatiniya garsani woottadasa ha mataafay taagaa ha keettay neegaa ha keettay aagaa hakeettay iigaa ha mataafay nuugaa ha tararppeezay hiinttegaa ha gatimay eetaagaa ha daangatiriya hiinteero ha daangatiriya hinterro hanna daangatiriya hiinteero ha keettay taagaa ha mataafay neegaa he tarappeezay nuuga gideese he billaway eetaaga gideese ha keettati nuuna maaddeennanitayakko nu hiinteeyoo mataafa immeennaagashini hiinte nuuna maaddeennanitayakko nu hiinteeyoo mataafa immokkoshini eeti eeta maaddeennanitayakko eetaaw mataafa immeennaagashini eet eeta maaddeennanitayakko eetaaw mataafa immokkoonashini ta oottanaaw koossheese ne yaawodiyaani ta oottana taani yaawodiaani ne oottanna ne yeenna wodiyaani ta oottikke ta ootteenna wodiyaani ne kare booppa taani oosuwaani miitettaasi taani oosuwaani beykke heezza immakka immanashini taafibeynna bagga taayoo bare immikke tammu bayzzidoogaappe kaamiyaani waatade beabeeyekke kuundibeynna wodiyaani beida laaggetura woottiisi bitanniyaappe maaddidaakko yaadi bitanniyaw maaddiyakko yaasu booppa mataafay yeekkiyaakko daangatiriya mati yeessaasu dareese naakko yelisoona daroosoona naatuussi yohuwa dees nai aassi diggaasu nayu aliyaara dooyyaasu nuussi asatu eetaaga ogiya aseessi eetaara saaba aseeyoo efiisi saabbida ashkareti ekkisoona saatiniya attuma ereenna shammideta awni ereeti shocabeyakka baanadani eroos gedeene kulfiyara taafaasu asey gelaawiyu laaggee taagaa taanara asiyu gelaawtaappe laaggetukko asnay gita maadi taara awgeta guuttati macceekko teesa awppe hai macceessi tuma aybaani hanide masshara ubbay aybaanne happuneekko mataafata woottadasa aybee heezzanto miisi yaana azela hegaappe miizata yeekkiyoogaappe azeloosona hiinini muuliya yibeenna baada hiinteena naassi yootadi

baade hiinteew naee ziniisi baakka iikko naeekko aagaa baanne immana nau aagata balaage immanaagashini nuuyoo acca balaagoosoona immibookkoona oottanaaw aduckkeese barayyeti isholatti oow alakaassi barayyiya kaawuwa ooyoo alakay baykka kare oyddu aliyaa baysi keettaappe shammadi asaakko bea biidooga gaaccidde hege iyoo biidoogaappe gaadiyaani hegeti kaanteese billawaani gameela heggatuuppe kaantiisi bitaanney garsaappe heyrayso kamma bitanneti geelaawta hiinteenara kammi bitanniyaassi gelisoona hiinteero keettaani dandayaasa geliye hiinteessi keha dandayaysi giddooni hiintena kiyisoona deaydde gideenna hiinteppe koommo deesshiya gideennaga hinteero kooyyaasa deidooga gideese hospuni kusshiya dendidi gidoppa hupiyaappe kuundiisi hupiyaassi laappuni diggaasi gupiya hupiyaaw maaddeennaani dooyyidaakko guyeera guutta kuma nu miizeetuussi ehida nuuni guutta kuma nu miizataayoo ehida nuuni guutta kuma nu miizataaw ehida a macciya guutta kuma hinte deessheessi efaasu a maccciya guutta kuma hinte deessheeyoo efaasu a macciya guutta kuma hinte deesshaaw efaasu ii azinay guutta kuma hinte deesshatuussi eflisi ii azinay guutta kuma hinte desshataayoo eflisi ii azinay guutta kuma hinte deesshataaw efiisi aylliyey eeta sooppe kiyeese asati baanta keetaatuuppe kiyisoona asati eeta keettatuuppe kiyisoona miitta acca mitta mata keetta acca keetta mata mitta laankiya keetta boollaani mitta boollaaani ogiya boolaani keetta garasa keetta garsan keetta geliisi soon woottiisi soo garsani woottiisi ii keettaappe oollaakko biisi ii keetta garsaappe oollaakko biisi keetta yuushuwa keetta gupiya mitta gupiya he asatu sintaani ekka mitta garsa mittaakko gakkanaashi heeraakko biisoona heera bagga biisoona keetta garsaappe keettaappe karee bagga karee bagga satiniyaani taappe sinta aylliya gaadiya kaantade baasu maskootteera teella yohaannisaappe guyeera kattaappe guyeera kammaakko omarsaakko omarsay gakkanaasha heezzu agnaappe ta gishata ne gishata a gishata ii gishata ta laaggiya gishata asatu giddo asatu garsa miissha gaasoota nu gaasoota ta gaasoota ne gaasoota a gaasoota ii gaasoota tammu dakiika garsa olaa wode asa oliisi taara ekettiisi taanara ekettisi ii taanara ekettide kaiisi ii asnay gediyaani biisi ii asnay tohuwaani biisi imattay kamiyaani biisi a macciya kulfiyara fengiya dooyyaasu a macciya kulfiyara ifitta dooyyaassu a macciya kulfiyara kariya ifitta maammada eetaagata haasayaakko eetaassi haasayey issi maanaagadani eetaayoo hagaappe imattassi maanaaw ekettiisi hanidooga imattay maasa ekkaasi hanukku immiyoogashini maatta ekkade haykkisoona immokkoshini mara erikke heera ishati maskootiyara mate oosuwaani taaran yootogaappe meeccaasi oottadi talaali aakko miccati oottanadan tamaare aanara miidaakko ootteennaani tarappeezati aara oottidde miidoogaappe tohora aaro miisshay oyddata toossa aayu

aasa miitta shaafa ufayettaasi aasati miiziya shaaya uraappe aayoo naatuuyoo shamma uray naaw shammiisi usuppun adukkawsu naayoo shimmaynneti usuppunata adukkeese naeera shocida waanade agnaappe naeessi shoruwaa waani ali neegaa shuccaara weysa aliyow neegeta shuccati wonta asaassi neeppe sindiya woriisi asakka neera siyaasi wottaappe asaoow nepee siyida yasshay asatoow nuuga sohuwara yeddiisi asatukko ola sooppe yeessisoona ashkaret oollaakko suwikaydde yeey assi awa bana bitannetukko awga barakka bitanniyaaw nuura awge baranne bitanniyakko nuuro awunna bay biya odogaappe awunno baynne biyaakko ogiyara aybaakka bayzettaasu biye oliisi aybaarakka bayzzid boogaancati omarsaakko aybaaranne bayzzidi daabbiya omarsay aybaayoo beadasa daabbuwa oogey aybaw beay daabbuwokka ookko aybayinee beisoona daabdaabbeta ookkokka aybaykka beykke daabdaabeeta ookkonne aylliyey bianney daannay oolla aymalaa bidoogaappe daarati oonakko azina bii daaray oonanne baadi biideta daddosoona oonara baana biidi dakiika oonaranne baanta biite dandayabeyakka oonikka baara billaway dandayay ooppe baggaani biossappe dandayeese ooppenne baggara bira dea oossanne deedooga deokkoona essaasu oossha deeenna deoppo fakkadeese oottana deeesi dersa fatuumey oottanaaga deenna desshataayoo fengiya oottanna deennada deyuuna fula ootteenna deennadani dussaanne fulaattawsu oottida deennaga duummara fulaattoosoona oottidaakko aybaaranne neeni shocabeyakka aybaarakka neeni shocabeyakka issi baranne neeni shocabeeyaakka issi barakka neeni shocabeyakka imattati guutta asa beisoona gooday guutta laaggetukko biisi wottadaarati isi heeraappe koommo baggara yiisoona woottadaarati koommo baggaani issi heeraappe yiisoona guutta sukkaariya taassi imma guutta sukkaariya taaw imma guutta kitta nuussi imma guutta oytta nuuyoo imma guutta haatta aaw imma guutta haatta aassi imma issoy taaw guutta oytta immiisi issoy taassi guutta oytta immiisi eettappe issoy guutta haatta kooyyeese hagani guutta sukkaariya deees imattaw guutta laaggeti deosoona imattassi guutta laaggeti deosoona imattay guutta laaggetura deeesi ii azinay guutta daabdaabeeta taafiisi neessi cora kitta immaysi neew darro kitta immaysi neeyoo daroo oytta immaysi he asati ubba wode darro haatta kooyyosoona nu alakay cora shaaya uyyiisi nu alakay daroo shaaya uyyiisi nuppe issuwaassi cora laaggeti deosoona nuuppe issuwaassi daroo laaggeti deosoona nuupee issuwaayoo cora laaggeti deosoona nuuni heyrayso asata eroos nuuni cora asata eroos nuuni daroo asata deeshati duummare gaattanokka oottidi deesshaaw eesuwaani gakkanaasha oottidooga deesshataaw eet gakkanaashi oottikke deesshati eetaaro gakkideta oottishine deesshatuussi eettappe gatimaani ooyetto deessheessi efisoona gatimay oraatta deessheeyoo

ekettide gediyaani oycca dei ekkani geelawiyu oycciisi deibeenna ekkeda gelaawata oyddati deibeykkoona ekkogaappe gelaaweekko saatiniyaani deiis ereesi gelaaweeppe samunaani deishine ereta gelaawotukko shaafaakko deisoona erettosoona gelide shamiisi demmibeenna erokko geliisi shammaasu gidanaw haasayeenna hasayaakko shammadasa giddo haasayiisi haykkaasu shammidoonaakko gidekketa haasayooga haykkiisi sharmuuta gidibeekkoona haasayoppa neena kaamma muuliya yeekkiyoogaappe diggaasi taani neeni kamma muuliya yeekkiyoogaappe essaasi aa nuuna wottaappe essaasu aa nuuna wottaappe diggaasu ii nuuni he yohuwa baa haasayeyooga siyiisi ii neeni guutta kuma kammi maanaagadani fakkadiisi ii neeni guutta kuma kammi maanaagadani fakkadeese taani neeni woonto tamaare sooni guutta kuma miyooga koffaysi ii neeni guutta haatta kmmi uyyana maala fakkadiisi ii guutta kuma ba laaggetura issiippe miidoogaappe taani soo ta maccekko baasi ii guutta katta ba laaggetura issiippe miidoogaappe taani soo ta macceekko baasi ii guutta kuma ba laaggetura issiippe moogaappe guyaani haysiya shocaasu gididooga haatakka heeni shoceeta gididosoona hagaadani heeraakko shociis gidisoona hagan heezzatakka shoosshaa gidiyakko hagana heezzatukko shoosshati gidokko hakeettay hegara shoosshatta gidokkoona hakime hegeta sinta gidoona hakimiya hegge sintaani gidoppite hanana heggeta siyido gidu hananaw henni siyiisi giishini hanibeekketa hiinteera sookko gisshaassi hanibeykko hiinteo soon gisshi hanidaakko hiinttegaa soona gisshoo hanideta hinteegaa sooni gooday hanite hinteegeta taafaasi goondallaanne hanniippe gosshaancaassi hospun taafida gosshaancaaw hanno hupee taafiisi iira hanokkonaa hupiyaayoo taafishini kaantade kiyoona taafiyooga iiro kaantosoona kiyoppo talala iitadani kaawootu koffaasu talalasi iiyoo kaawoy koffaysi tamaani taara ekettiisi taanara ekettisi ii taanara ekettide kaiisi ii asnay gediyaani biisi ii asnay tohuwaani biisi imattay kamiyaani biisi a macciya kulfiyara fengiya dooyyaasu a macciya kulfiyara ifitta dooyyaassu a macciya kulfiyara kariya dooyyaasu ta samunaani mayyuwa meeccaasi yasshay bayeennaani yasshay yayyeennaani neenara gideenna ne bayeennaani neeni bayeennaani miisshay hupee ta ta talaali ne ne talaali ne ne talaasi ne duummare ii duummara eeti ubbay nu ubbay nuuni ubbay nuuppe baggati nuuppe guuttati eetaappe guuttati hiinteppe guuttati ooni yiide ooni bii oona beadi neeni oow yootadi neeni ooyoo yootadi iiyu kaiisi kooncissa tamaareesoo imattaaw kamiyaani koossheese tayakko imattati kammaakko kooyaakko teegiisi imattaw kariya kooyya teella imettosoona katta kooyyakko tiillee immaasa kattaappe kooyyayi tiilliya immaasi kattani kooyyeenna tohuwaani immakkashini keetaatuuppe kooyyideta tooranne immawsu keettaana kooyyiisi tuuggaasu

immeeta keettaaway kooyyosoona ubbageti immekketashini keettatuna kooyyu ubbaykka immokkoonashini keettatuni koppaasu uddupuni imota keettatuuppe koriddeenee uraassa ishatey keettidey koyiide uraassi isippe kessiisi koyyey usshaasu issuwaayoo kessosoona kumaani usuppunati iyu kiyana kumi uttida laaaggey macciyta misshiriya uyyaasu laaggeeti masmuriya mittaakko uyyide laaggetupee mayota miyooga waanidi laankiya mentide mizade waannaassi laappun miccatey moogaappe waasani laggeti miidde naaassi waatidi loittade miiddi naaggade wara looabeekku miidooga naantona waraasa looana yida miisshaka naayka waraasu looanaaw miitettaasi naeeyoo woonto lookko miizaa naey wootta looppa miizataaw nagaasa woottaasi maaddeenanitayakko miizataayoo nagosoona woottadi maaddeennakko miizati naiyu woottida maaddeennani miizeessi nanga woottideta maahe miizeetuussi nayo woottisoona maahiya miizeeyoo neekko wottaasu maala miiziiyu neenara wurestta maalalettaasi minu neero wuriya maattaanne minuweese nuppe wursaasi maccaasatoow mishireessi nuugeta ya yaabare yaanatte yaasitte yaasutte yayyeennaani yayyiddeenne yibeenna yeddade yeddiyosaappe yeenna yeesiisi yeessideta yeessiisi yekkey yekkiyaakko yohaannisaappe yohoy yuushuwa zammarida