ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

DESIGNING A STEMMER FOR GE'EZ TEXT USING *RULE BASED APPROACH*

A Thesis Submitted to School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Science

BY

ABEBE BELAY ADEGE

July, 2010

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

DESIGNING A STEMMER FOR GE'EZ TEXT USING RULE BASED APPROACH

A Thesis Submitted to School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Science

BY

ABEBE BELAY ADEGE

Name and Signature of Members of the Examining board:

	Name	**signature**	**Date**
Chair person,	_____	_____	_____
Advisor,	_____	_____	_____
Examiner,	_____	_____	_____

Dedication

This work is dedicated to my father, Belay Adege Mekonen, and my brother Mulugeta Belay, who were breathe their last breath in my B.SC studies.

Acknowledgements

First, I need to thank my mother St.Mary (the Blessed ever vergin Mary) with her son 'Jesus christ' who help me by making things easier and give me persistency to all faces up in my life.

Subsequently, my deepest and my heartfelt thank goes to my research advisor Ato Ermias Abebe for his unlimited advices, constructive comments and suggestions.

My deepest gratitude also goes to Dr.Million Meshesha who sows the seeds of IR into my mind and initiated the area of this thesis. The expert person: M/r Dessie Qeleb also deserves special thanks for his contribution and encouragement.

I also greatly thank all my friends for their dear advices and supports, especially Ato Getachew Tesfa. I also deeply acknowledge Ato Melaku Assefa and W/o Hirut Hailu (hire out of my residence) for their supports and encaragments. I am grateful to Ato Abere Ashagre for his unreserved supports at any time I need him during this study.

Finally, my thanks go to my love Weyneshet Melese, my families, colleagues and work partners who helped in my study at School of Graduate Studies, Addis Ababa University.

June, 2010
Abebe Belay
Ababa, Ethiopia

Table of contents

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS AND SYMBOLS

Abbreviations/symbols	Meaning
➔	becomes
' '	gloss
1p.sg	1st person singular
1p.pl	1st person plural
2m.sg	2nd Person singular masculine
2f.sg	2nd person singular feminine
2m.pl	2nd person plural masculine
2f.pl	2nd Person plural feminine
3m.sg	3rd Person singular masculine
3f.sg	3nd Person singular feminine
3m.pl	3rd Person plural masculine
3f.pl	3rd Person plural feminine
nom.	Nominative
acc.	Accusative

ABSTRACT

In this study, a stemmer of Ge'ez text was developed. In designing processes, different concepts such as background for the thesis, literatures on conflation of the stemming algorithms, morphological nature of Ge'ez language, stemming techniques and other realted things were discussed in order to model and develop an automatic procedure for conflation.

When inflectional and derivational morphologies of the language were discussed, affixations such as prefixing, infixing and suffixing are the main word formation processes in Ge'ez language. The language is morphologically complex. This is because different words can be formed due to the wide concatenations of affixes.

For the experiment, two techniques were used: affix removal and morphological analysis techniques. To evaluate the stemmer, manually error counting technique was used.

From the experiment, three types of errors are observed: over stemmed (6%), under stemmed (4.27%) and structural problems (7.31%). When the stemmer runs on the sample texts, it performed with an accuracy of 82.42%.

The dictionary reductions of the stemmer were 29.9% to the stemmed words and 62.8% to root words.

Lastly, the possible recommendations to future works and improvements of this work were reported.

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND TO THE STUDY

Morphology describes how various forms of words are created, and studies structures of words in the language. Suffixing and Prefixing are the main and common ways of creating word variations in natural language text [18]. Morphology (the internal structure of words) can be broken down into two subclasses: inflectional and derivational [4, 26]. Inflectional morphology describes predictable changes of words. These are as a result of syntax, such as changes in person, number, tense and gender. These changes have no effect on a word's part-of-speech [1], (for instance, a noun still remains a noun after pluralized). For example, kick, kicks, kicking, kicked.

On the other hand, changes of derivational types may affect a word's meaning in part of speech. For example, affix changes from adjective to nouns, from verb to nouns, from noun to verb, and so on; like friend, friendly, friendliness, and friendship.

From either type of morphologies, depending upon the complexity of the morphological natures of language types, very massive variants of words may be resulted from single word. These large numbers of occurrences have strong impact on information retrieval systems. Thus, there is a need of automated procedure that can reduce the size of the various words to manageable level, and also capture the strong correlations existing between different word forms [15]. Even if the

complexity of morphological variations may differ from one natural language to others, stemming is widely used in information retrieval (IR), with the basic reason that morphological variants represent similar meaning.

Morphological processing becomes widely used for effective and efficient information retrieval [13], machine translation (MT) and text classification [5]. Hence, morphological processing becomes particularly important for information retrieval (IR) because IR needs to determine the appropriate forms of words as index [17].

Information retrieval particularly automatic information retrieval system is an information processing activity which is carried out with the help of automatic equipment. As [19] defined automatic information retrieval system is a software/hardware package that lets different users to access query and retrieve information from the database.

Stemming is a natural language processing technique that identifies a stem/root word from morphologically conflated words using various techniques on inflectional and derivational affixes [18]. It helps to reduce morphological variants of words into a common form. As an example, ‘throws’, ‘thrower’, ‘throwers’ and ‘throwing’ are the common variances of a stemmed word ‘throw’. Therefore, stemming can be applied in different languages to increase searching efficiency and reducing vocabulary size of indexed files in information retrievals [1, 4].

In stemming, all of the word's true prefixes and suffixes are removed to produce the stem or the root word [11]. This is import to classify texts and index builders. It is also used to make operations fewer dependants on particular form of words rather than enclosing all possible forms. The computational process gathers all words that share the same stem and have some semantic

relations. It is also used for document matching and classification to convert all likely forms of a word in the input document to the form in reference document [11].

There are four major types of automatic stemming strategies [18]: affix removal, table look up, successor variety, and n-gram. Affix removal stemming is imitative, simple, and can be implemented efficiently. This type of strategy removes suffixes and/or prefixes from terms to give a stem. This type of strategy was done by [1, 3, 11]. Table look up strategy looks for the stem of a word in a table of the dictionary. This strategy is done by [4] to stem Amharic morphological words. It is a simple procedure and depends on size of stemmed data for the language. But, this strategy requires an available stemmed dictionary before the activities will be performed. It also causes to require considerable storage spaces. Successor variety stemming is done based on the determination of morpheme on boundaries, uses knowledge from structural linguistics, and is more complex than Affix removal. N-gram stemming is used based on the identification of di-grams and tri-grams and is more of a term used for clustering than stemming. It is done based on numbers of n-grams. This strategy is done by [12] to stem language independent terms using single gram.

Unlike languages like English that has less morphological variants, there are languages like Amharic, Arabic and Ge'ez that are much rich in morphology. So, like Amharic [1] and Arabic [11], Ge'ez involves dealing with prefixes, infixes and derivatives in addition to suffixes.

1.2 STATEMENT OF THE PROBLEM AND JUSTIFICATION OF THE STUDY

This research is the first work to Ge'ez stemmer. It will be a starting for other works such as designing indexers, summarizers and cross linguals to the language.

Today, Ge'ez courses are given at various colleges, religious centers and higher instituites. Different Ge'ez books and other reference materials are publishing now. For example, Ethiopian Orthodox Church Theology Colleges are giving now various courses and publishing now various books. Ge'ez is also given in higher institutes like Addis Ababa University. Therefore, designing automatic Ge'ez stemmer helps to develop indexers for search engine improvments of efficentcy and effectiveness of the information retrieval systems.

There are a number of available Ge'ez literatures, some of which are translated to Amharic and Tigrigna languages. The translated documents may not be exactly matched with the original one. So, it is better to access information or documents that are written in Ge'ez automatically with the application of natural language processing(NLP), information retrievals(IR) and machine translation(MT).

Ethiopia is a country with several cultures and languages. Languages such as Ge'ez, Amharic, Tigrigna, Guragegna, Afaan Oromo and so on are giving now services in the country, Ethiopia. These languages play crucial roles for the people in social, political, cultural and economic activities. Other languages, especially Amharic and Tigrigna benefited a lot from Ge'ez. Their many words are borrowed from Ge'ez. Hence, designing a stemmer to Ge'ez language helps to improve information retrieval efficiency of other languages that are borrowed from it. This improvement will be done by incorporating Ge'ez stemmer with other language's stemmer.

Information can exist in various forms, but accessing the right information at the right time is a pre-requisite for functional efficiency [6]. Now, accessing information is becoming critical to the success of all users within a given society. This is because the economic development of any country depends upon the effective utilization of the stored information, especially for developing countries [6]. Today, a large amount of information is available in the form of written text. Developing countries like Ethiopia can facilitate its vision of development by making this store of knowledge accessible to their citizens. This can be done with the help of different techniques such as natural language processing, speech recognition, machine translation and information retrieval. These techniques mainly depend upon morphological analysis (or stemming) for their effectiveness [1].

Ge'ez, the classical language of Ethiopia, is still used as a liturgical language by the Ethiopian Orthodox Tewahido Church (EOTC), Ethiopian Catholic Church (ECC), the Bete Israel Jewish community of Ethiopia, and Eritrea [8]. Hence, developing a stemmer is used to increase large electronics Ge'ez documents and is good opportunities to increase information retrieval system's efficiency.

As Ge'ez is morphologically complex language, there is a need for automated procedures that can reduce the size of lexicon to manageable level and improve the application of information retrieval and natural language processing. Retrieval systems allow for catching information easily and timely from a large body of sources [18]. This is because larger body of information and research available mostly in the language of the primary researcher [9]. Hence, much information cannot be shared between research communities without extensive translation and

considerable time delay. Therefore, by developing retrieval tools for Ge'ez language, one can design a retrieve system for the language.

Since stemmer is language specific [14, 20], it is not possible to make use of a stemmer developed for other languages. This is because of structural differences of the languages. So, it seems quite difficult to follow the same stemming pattern and apply the same technique for all languages. Therefore, the stemmer of one language cannot be effectively applied to other languages [20].

Different researchers attempted to develop stemming algorithms for Amharic [16], Tigrigna [9], Afaan Oromo [2] and wollayta [14] to simplify the problem of morphological complexity of the respective languages. But, as far as the researcher's knowledge goes, there is no work on the development of a stemmer for Ge'ez language. Therefore, this study is initiated to explore the rule based algorithm for stemming words in Ge'ez language.

1.3 OBJECTIVE OF THE STUDY

The research has the following general and specific objectives.

1.4.1. GENERAL OBJECTIVE

The general objective of this research is to develop a stemmer for Ge'ez language using Rule based approach. To design the rule based stemmer, affix removal and morphological analysis techniques are used.

1.4.2. SPECIFIC OBJECTIVES

The following are specific objectives of this research.

- Review literature to have a clear understanding of the area and identify different stemming algorithms that have been developed for other languages;
- Review properties of the Ge'ez language in order to get familiar with the different aspects of the language and know how the separation of words is achieved;
- Build a stop word list in Ge'ez language and compile a list of affixes used in Ge'ez;
- Select appropriate stemming algorithm for Ge'ez language;
- Develop a stemmer that identify inflectional and derivational affixes of Ge'ez language;
- Test the performance of the stemmer on the selected Ge'ez text and report the finding of the study
- Forward recommendations for further improvements/studies.

1.4 METHODOLOGY

1.4.1. LITERATURE REVIEW

Reviewing literatures such as books, articles, magazines and Internet were done to understand the subject matter in detail as well as to gather key ideas, information, concepts and experiences. There are a number of stemming algorithms for different languages such as for Amharic [1], Arabic [20], Tigrigna [9], and Wollayta [14] languages that are used to get basic ideas about this work.

A review of literatures were also conducted to familiar with morphological patterns, rules and sequences of Ge'ez text features in relation to information retrieval. Discussions with professionals at Addis Ababa University, Ethiopian Orthodox Tewahido Theology School and other experts are also used to understand and define the morphologies of Ge'ez language.

1.4.2. PROGRAMMING TECHNIQUE

A prototype stemmer for Ge'ez language was written using the Python programming language (Python2.6 and Python3.0). This is because it is relatively easier to manipulate text (string manipulation) and also the researcher has some experience in writing programs using Python programming language.

1.4.3. DATA SOURCES FOR EXPERIMENT

A text corpus is one of the resources required in IR research. 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 data set was organized from collections of data sources. These sources are historical books such as Abune Habte Marrian History [55], cultural as well as religious books like Bible in Ge'ez [7] and Wdase Marriam [54]. The reason of selecting these sources is the materials are available readily. The selections of topics from all materials were done randomly.

1.4.4. EXPERIMENTATION METHODS

Evaluation was done to measure the performance of the stemmer. The accuracy of the stemmer was measured by comparing the actual data set with the stemmed one.

In the experimentation, the stemmer was run on the sample texts that were selected randomly. The data set contains 1866 words. The evaluation of the stemmer was done via a manual error counting mechanism.

1.5 SCOPE AND LIMITATIONS OF THE STUDY

In this research, a stemmer developed for Ge'ez mainly focused on derivational and inflectional word variants of the language. The research does not focus on compound and irregular words because of the high morphological complexity of the language. Irregular words follow various structures. Hence, it is difficult to design a stemmer in a short period of time.

The stemmer contains prefix striping, suffix striping and structural transformation techniques. The structure was designed for various forms of words in one. This causes for a stemmer to produce unmatched structure of words. For example, a structure can hold and stemmed in a similar fashion for both verbs and nouns if they have identical consonant-vowel structures.

There are lacks of computational resources and lack or limitations of professional persons. Hence, it was difficult to get timely response from language experts.

1.6 APPLICATION OF THE STUDY

This study has various significances, as Ge'ez is the base for many of the Semitic languages in Ethiopia and this research is the first work to Ge'ez stemmer. It will use as a start for doing further researches on computational work of the language.

This research can also help to develop tools like spell checkers, grammar checkers, thesauri, word frequency counters, document summarizers and indexers. It can also be used to reduce word variants and to minimize total numbers of files.

Last but not least; it can be used for developing cross lingual retrieval systems. For example, one can feed a Ge'ez text to search engine, this entered word may be translated to other languages like English and accessed documents in English or other languages.

1.7 ORGANIZATION OF THE THESIS

This thesis consists of five chapters. Each helps to introduce the significance of text stemmer and its needs, particularly Ge'ez texts. The stemmer works by conflating word variants in Ge'ez language (inflectional and derivational morphologies).

In chapter one, background of the study, statement of the problem and justification of the study, methodology, scope and limitation of the study are presented.

In chapter two, reviews of various related works on conflational techniques and stemming algorithms for different languages are reviewed. The approaches to stem words and types of stemmers were discussed in this chapter.

Morphologies of Ge'ez text are reviewed in chapter three. The inflectional and derivational morphologies of the language are the main concerns of chapter 3. Word formation processes for Ge'ez nouns and verbs are presented in detail.

Development and experiment of the stemmer for Ge'ez text were dealt in the fourth chapter. The compilation of stop word lists, prefix striping, infix striping suffix striping and morphological analysis techniques are presented in this chapter.

Finally the result obtained from experiment, conclusions of the work and recommendations for further researches are discussed in chapter 5.

CHAPTER TWO

REVIEW OF RELATED LITERATURES

2.1. INTRODUCTION

The main concepts included in this chapter are review of literatures about the need of term conflations, approaches in term conflations and various stemming algorithm types. The advantages and disadvantages of various approaches for term conflations and stemming algorithms as well as concepts about stop words are discussed.

2.2. CONFLATION TECHNIQUES

In information retrieval, the relationship between a query and a document is determined primarily by frequency of terms, which they have in common [21]. However, some words may have different forms that need some forms of natural language processing. The existence of these variants for a word is mainly caused by morphological variants. Hence, morphological variants are the main problem by making a word to have various forms in the documents and queries, and these have problems in indexing and retrieving systems [3].

As discussed by [22], gender, number, tense, person, mood, or voice can characterize variants of words. In addition to these, the requirements of grammars (e.g. "connecting" and "connectional"), valid alternative spellings (e.g. "recognize" and "recognise"), antonyms (e.g. “ability” and

“disability”), and misspellings as well as abbreviations are also cause to word variants. In general, morphological variantce is the main source of text word variants [1]. Hence, morphology uses to study and describe the forms of words in a language, and commonly includes inflectional, derivational, and compounding.

As [42] discussed, derivational morphemes make new words from old ones, like *creation* which is formed from *create*, but they are two separate words. It causes to change part of speech or basic meaning of a word. On the other hand, inflection morphemes vary (or "inflect") the form of words in order to express grammatical features, such as singular/plural or past/present tense. For instance *boy* and *boys* are two different forms of the "same" word. Generally, inflectional morphemes do not change basic meanings or parts of speeches.

Compounding is the process of constructing morphologically complex words from two or more unbound morphemes. That is, compounding is the joining of two linguistic forms, which are functionally independents [43].

According to [21], semantic interpretation of morphological variants of the same words considered as similar and equivalent terms for the purpose of IR applications. Hence, we have to take into consideration the morphological variants of words as one reference of all word forms, because simplifying these complex forms of a word is comparatively important in determining the relevance of documents to users’ queries in information retrieval. These help to perform stemming in different languages fruitfully using a number of stemming or conflation methods to get root/stem of morphological variants.

The variations of a word are not recognized equivalently without some form of natural language processing (NLP), since word form variation has huge impacts on information retrieval system effectiveness [21]. Therefore, it is better to use a mechanism that uses to reduce variation of word forms. Accordingly, equivalent word forms should be detected and reduced to a single form [23, 28]. To alleviate the word morphological problems, we use a method that helps to bring together semantically related words to a single form, which is conflation technique [30].

As defined by [22], conflation is the process of merging or lumping together non-identical words that refers to the same principal concept [22]. It can also be defined as the process of matching morphological term variants by mapping term variants to a single form, usually a unique well-formed root for each word[23]. It performs by bringing similar words that have similar meaning with in a common form. Thus, conflation technique helps to solve morphological variants problems.

According to [40], conflation has the following major functions: reducing the requirement of storage media, increasing retrieval effectiveness and decreasing process costs as well as increasing the speed of retrieving processes. It also uses to capture the strength in the relationship between different word forms, to overcome vocabulary mismatch problems as well as to better generalizations [28].

There are two major types of conflation techniques [22, 30]: manual and automatic conflations. Manual conflation is performed at search time with right-hand truncation[1]. It is applied only to words at the query, but not applied in the document.

Automatic conflation, on the other hand, is designed using programs called stemmers. It improves the problem of word variants by matching different forms of the same word automatically. This type of conflation is applied to words in queries as well as in documents. Automatic conflation has two classes [20, 30]: stemming algorithms and string-similarity algorithms. Stemming algorithms or stemmers are techniques for reducing words to their grammatical root/stem form. These types of algorithms are usually designed to handle morphological variants with in a language (see the detail in 2.3). In contrast, string-similarity algorithm is usually language independent and designed to handle all types of variants. This approach involves conflating by calculating and measuring the similarity between an input query terms and each of the distinct terms in the database. Terms that have a high similarity to a query terms are then displayed to the user for possible inclusion in the query. The common example to string-similarity algorithm is N-*gram* matching techniques (see the detail about N-gram technique in section 2.3).

In both manual and automatic conflations, over stemming and under stemming problems may occur. In over stemming, the stem is too short (too much of affix is removed from a term) and will result in unrelated words being stemmed to the same stem; while under-stemming arises if too short affixes are removed and may result in related terms being described by different stems.

[1] Removing letters manually from right hand-side of the word by the searcher is known as truncation.

Example; the stem of a word 'Instructors' may be considered as 'Instructor', rather than 'Instruct'. These errors are occurred when a word is not stemmed properly. These causes to irrelevant documents are retrieved or only few relevant documents are retrieved. The errors that may occur in stemming could be corrected manually using different exceptional lists, or by means of rules in the stemmer [28].

As [33] discussed, the effectiveness of the retrieval system is usually measured/evaluated using a combination of both recall and precision. Precision is a proportion of the retrieved documents that is actually relevant to user interest; on the other hand, recall is a ratio of the relevant documents that is actually retrieved from the file. Achieving high recall and retaining a high precision use to improve the salient of information retrieval systems [33]. Hence, to improve the recall level, there are various types of methods such as use of term phrases, use of truncated terms or stems, and use of the synonyms [23].

Another issue that helps to illustrate retrieval effectiveness is stop word lists. Stop words are removed because of the following impacts on information retrieval [28]: they can affect the retrieval effectiveness and weighting process. This is because they mostly have a very high frequency and tend to diminish the impact of frequency differences among less common words. They can also affect efficiency due to their nature and fact that they carry little meaning, which may result in a large amount of unproductive processing. Therefore, removing stop words results to increase effectiveness of retrieval by reducing storage requirements and increasing the matching of a query with index terms of a document.

2.3. STEMMING ALGORITHMS

“Stemming algorithm is a computational procedure that uses to reduce all morphological variant of words to their common form i.e. root/stem, usually by stripping each word of its derivational and inflectional affixes” [24, 32]. Stemming algorithm is also defined by [27] as “an algorithm which maps different morphological variants to their base form (stem). The stem of words is obtained by removing all or some of the affixes attached to the word, and this helps to make information retrieval a faster process [15]. Thus, stemming algorithm uses to know morphological properties of different words in a language.

In information retrieval, morphological variants of words that have similar semantics can be stemmed and grouped together to increase the success of matching documents to query. This is because the high inflection of words in a language results to reduce the accuracy of information retrieval systems [30]. Therefore, it is necessary to use certain stemming techniques that improve success of Information retrieval (IR).

According to [18, 21, 32], basic stemming methods are classified as follow: affix removal, table lookup (Dictionary method), successor variety, and n-gram stemming approaches.

The affix removal is a technique that uses to convert different forms of words to their base form (root or stem) by removing prefixes, infixes and/or suffixes from word variants [3]. For example; Porter algorithm consists of a set of condition/action of rules for affix removal stemmer. For instance, stripping affixes from each word variants like stripping of “-ES” from a word “SEARCHES”, “-S” from “READS”. Affix removal stemming algorithm can also be characterized as: the longest-match and shortest-match/iterative approaches.

The longest-match stemmer technique removes the longest possible string of characters from a word based on the allocated rules. It is made by removing affixes one after the other with single iteration until no more characters can be removed. In the case of longest-match stemming algorithm, the suffix lists are sorted in decreasing length to reduce programming complexity and the longer endings are first scanned; if a match is not found, then the shorter one is proceed. That is, if more than one suffix matches the end of the word, the longest one is removed by possible combinations of suffixes. As [36] and [2] discussed about longest-match algorithm, it is often easier to design a program, but it has a problem of requiring a much longer dictionary since frequent combinations of short suffixes must be included. Hence, in this type of approach the list of affixes is very large compared to the iterative method.

On the other hand, shortest-match (iterative) approach removes the shortest possible suffix from the word and re-conceders the result words. This method is done by starting from the end of a word and working toward its beginning recursively to remove strings in each order-class one at a time [35]. For example; if a word “PEACEFULNESS”, is given “-NESS” will be removed in the first iteration and get again a word “PEACEFUL”, and then the suffix “-FUL” will be removed and leave “PEACE” as final stem.

As discussed by [20], stemming using affix removal has the following challenges: improper removal of some affixes may appear as prefix, infix or suffix, failure to extract the singular form of a broken plural and wrongly stemmed some exceptional words in the language. Hence, it is better to use this type of algorithm with certain exceptional rules.

Another type of stemming algorithm is table lookup (dictionary-base) stemming algorithm, which is constructed using very large dictionaries of morphological forms such as stems, roots, and affixes [41]. Even if it is a simple to proceed with this technique, there should be a manually synthesized table or a large dictionary for checking words with corresponding different word forms. This type of technique has inconveniences because of storage overhead, difficult to construct the table and multiple dictionary entries for similar concept. But, its result is generally correct [31]. Hence, due to the above problems and unavailability of Ge'ez dictionary, the technique is left to this paper. From Ethiopian languages, Amharic stemmer is designed using this technique by [4] for converting morphological variants of Amharic to their citation. The researcher could achieve an approximated result value of 60 percents and 75 percents from old fiction and new article texts respectively.

According to [32], successor variety strategy is a method that does a task based on structural linguistics. It is used to determine root words rather than the affixes based on the distribution of phonemes from a set of unique words in the corpus. This type of technique requires knowledge from linguistics, and more complex than affix removal technique. It should be applied on a very large collection of data to get sufficient statistical information, and works with letters instead of phonemes. In this type of stemming method, when a word is stemmed from two different domains, the result of the stemmers may differ.

As discussed by [29], N-gram stemming technique can work as well for indexing terms using overlapping sequences of n characters; many of the benefits of stemming can be achieved without any knowledge of the target language. This is because some of the n-grams of the word that do not exhibit morphological variation. For example, words like juggle, juggling and

jugglers share the common 5-gram juggle. It is entirely language-neutral; no knowledge of a language is required to apply n-gram tokenization to the language beyond selection of a suitable value for n. Its problem is in retrieval performance and disk usage, but not in retrieval accuracy. Because each character of a text begins a new n-gram, an n-gram representation of a text contains many more indexing terms than does a word or stem representation, increases the number of disk seeks required to locate all of the postings for a query. It is done by a number of diagrams instead of terms and mostly uses to clustering procedure than a stemming.

As [16,35] discussed, stemming algorithm can also be characterized as context-free and context-sensitive stemmers, where context refers to mean any attribute of the remaining stem after stemming. They are used to make a decision whether remained word actually represents the morphological variances or not in affix removals.

Context-free stemmer carries out by removing affixes without regarding the remaining stem. It has no any restrictions which are placed on the removal of suffixes. Thus any ending which matches is accepted for stripping. It can remove strings that are similar to, but actually that are not similar. For example; removing of “re-” from “regular” or “-al” from “metal” in English are context-free stemming types. But this method gives poor results to morphologically complex languages like Amharic and Ge’ez languages. It has no qualitative or quantitative restrictions on the removal of endings except a one quantitative restriction to the length of a stem should at least two [35].

In contrast, context-sensitive stemmer is done based on various restrictions that are placed on affix removal program such as an affix dictionary to give the exact affix forms, and a set of rules

that helps to define the morphological situation of the affixes. This algorithm is done by considering the remaining stem using many exceptional rules [16]. For example, rules like do not removing an ending that begins with “-en”, following “-e" because violation of this rule would change seen to se-, which is ambiguous stemming. Hence, better result can be obtained from context-sensitive than context free approach [24].

As [24, 25] proposed, a better stems can be found using a context-sensitive approach by adding constraints to the stripping operation such as:

1. Quantitative constraints: the remaining stem length must exceed a given number. Example, from a word “sing”, by removing “-ing” the system cannot drive “s” because with a minimum of length two characters are required.
2. Qualitative constraints: the stem ending must satisfy a given condition, like removing the suffix “-en” if the remaining stem does not end with “se-”).
3. Recoding rules: these are rules that are used for spelling checking and improve the accuracy of conflation in affix stripping algorithm.

For various languages, different stemmer researches are designed using different approaches. Some of the works that are done in Ethiopian languages include Amharic [1,4,16], Afaan Oromo [2],Tigrigna [9] and Wollayta [14]. From non-Ethiopian languages, there are different stemmer works such as English [25, 26], Dutch [27], Arabic [17,20,23], Indonesia [13], Malay [15 and so on.

The Amharic stemmer, which was done by [16], used a context-sensitive iterative approach that removes both prefixes and suffixes. In the work, the achievement was an accuracy of 95.9% from 1221 sample tasted words of data.

Afaan-Oromo stemmer, which was done by [2], used the longest-match context-sensitive approach that removes prefix and suffix. The stemmer was evaluated by counting stemming errors, and reducing of dictionary size. The stemmer performance was an accuracy of 92.52% from sample data of 1061 words.

Tigrigna language stemmer was done by [9], and used a context-sensitive iterative approach like Amharic stemmer designer. The researcher used rules that remove prefix, suffix, prefix-suffix pair and reduplication of single and double letters. Like that of Afaan Oromo stemmer, the accuracy of the stemmer is tested based on error counting techniques. From the experiment, the result of the stemmer accuracy is 84% from 1568 sample data words.

According to [14], stemmer in Wellayta language was designed, and an accuracy of 86.9% achieved. The stemmer used context sensitive iterative approach to this work. Error counting technique was employed to evaluate the performance the stemmer using error-counting techniques.

Stemmers can be judge with in different contexts [13]: correctness, retrieval effectiveness and compression performances. The correctness of stemmer describes over-stemming and under-stemming which are evaluated based on error counting [28,37]. Hence, computing of under stemming and over stemming is necessary to distinguish the proposed stemmer is performed effectively in IR systems or not.

Different researchers can use different algorithms to stem texts in different languages. The combinations of various rules of various algorithms help to improve the stemmer performance. For example, [41] used the combination of affix removal, dictionary based and morphological analysis techniques to stem Arabic language. According to the researcher, designing rules for morphologically complex words like Arabic help to improve information retrieval systems.

The rule based approach uses to handle exceptional things by applying different techniques based on linguistic rules. For example, [41] used rules on affix removal and morphological analysis to Arabic Language. The researcher also used dictionary based for checking and minimizing stemmer errors. But, researcher couldn't get any available dictionary to Ge'ez document.

As discussed by [44,45], a number of difficulties for the stemmers such as over-stemming and under-stemming could be alleviated by the application rules with different algorithms. Applying important rules with stemming algorithms helps also to filter as well as to increase overall IR performances. For instance, in Ge'ez verb: ቀተለ can be written in various morphological forms such as ይቀትል→ይቅትል, but ከወወ does not follow similar pattern, rather it can be written as ይከወው→ይከውው; not change ከ to ክ.

This can be handled using exceptional rules. Hence, a rule based approach was used to design Ge'ez stemmer.

CHAPTER THREE

MORPHOLOGY OF GE'EZ

3.1. INTRODUCTION

Ge'ez language has a characteristic of conveying different messages with a single word alone [51]. It is rich in vocabulary because a single word can appear in different forms by adding affixes or changing alphabetical patterns.

Developing a stemmer for a language requires studying and modeling the language phenomenon in terms of word formation. This is because a stemmer plays a high role in identifying a word-stem from several possible forms. Therefore, it is necessary to study Ge'ez morphology in order to model it and develop automatic procedures for conflation of words in the language.

This chapter focuses on concepts about inflectional and derivational morphologies of Ge’ez language. This is because morphology plays a high role to guide the development of the stemmer, and describe the nature and characteristics of affixations in Ge’ez words.

3.2.GE'EZ ALPHABETS AND THEIR SOUNDS

Ge’ez language was the most widely used written language in the historical Ethiopia and Ethiopian Orthodox Tewahido Church. Art works, governmental documents and religious scripts which were widely available in the church and governmental possessions’ are inherited to different users. The writing system of different Ge’ez alphabets of similar sound are written differently in the early times.The different writing of alphabets with similar sounds would raise questions.

According to [53], the most likely deriving factor for the creation of those letters in Ge’ez language were primarily the nasal, flap/trill, dental, and velar phonemes. In the table below some common characters with similar sound are listed.

Number	First order (ግእዝ)	Second order (ካዕብ)	Third order (ሳልስ)	Fourth order (ራብዕ)	Fifth order (ሃምስ)	Sixth order (ሳድስ)
1	ሀ፣ ሐ፣ ኀ			ሃ፣ ሓ፣ ኃ		
2	አ፣ ዐ			ኣ፣ ዓ		
3						እ፣ ዕ
4	ሠ፣ ሰ					
5	የ				ዬ	
6			ዪ			ይ

Table 3.1 Sound similarities in Ge’ez characters/Alphabets

The interchangeabilities of Alphabets/Fidels were widely encouraged after Tigrigna, Amharic, and other languages were becoming a spoken language in Ethiopia. Though there are sound similarities among characters, generally should not be used interchangeably for Ge'ez language [53], because the meaning of the word will be changed.

In the early days of the language's history, however, the letters that now represent the same sound, used to have phonological variants .There are, for example, the three types of ሀ 'ha'. A word such as ሀ is only written in this way but not either as ሐ or ኀ . Similarly ሐ is not written as ሀ or ኀ. The same is true to ኡ and ዐ. Using these interchangeabilities are full mistakes compared to the early writing system. At this time, although in the linguistic history of Ge'ez, certain phonemes become confused in the pronunciation, one will notice that quite several Ge'ez phonemes correspond to one phoneme, as shown in Table3.1.

Ethiopic or Ge'ez Fidel has the phonetic structure of seven columns with 26 syllographs: starting from ሀ 'ha' up to ፐ'pä'. Each 26 characters of Ge'ez Fidel combine with each of the seven vowels to create sounds of the language. In addition to 26 basic alphabets, there are five alphabets which have not the second and the third orders. See lists of consonant translations, numbers, punctuations and vowels of Ge'ez language in appendix I**.**

3.3. MORPHOLOGY

Morphology is a branch of linguistic that studies and describes how words are formed in languages [21]. It deals with the internal structure of words in natural languages. The word form variations in Ge'ez could be formed by inflectional and derivational morphologies. Both types of morphologies can be created using affixations.

Affixation is the process of adding or attaching affixes in some manner to root/stem of a word [48]. Affixes cause for word variants.

3.3.1. GE'EZ MORPHEMES

A morpheme is the minimal linguistic unit of a language that carries meaning and that cannot be further decomposed into meaningful units [50]. A morpheme in Ge'ez can be free or bound, where a free morpheme can stand as a word on its own where as a bound morpheme cannot occur on its own as a word [48]. In free and bound morpheme of Ge'ez language, it is assumed that the later is typically derived from the former. Therefore, bound morphemes exist with other morphemes, and they are usually either affixes or roots [9].

3.3.2. WORD FORMATION IN GE'EZ

As discussed by [48], affixing (adding suffixes, prefixes, and infixes), compounding, duplicating (reduplicating) and different vowel patterns are used to create various word forms in Ge'ez language.

The addition of suffixes, infixes and prefixes are the common ways of word formations in Ge'ez like the other Semitic languages such as Amharic [16] and Tigrigna [9]. Continuous affixation results in long word formation. Hence, the complex morphological structure of a single Ge'ez word can give rise to a very large numbers of variants.

The pattern of vowels in a word can also create various word forms in Ge'ez text. For example, according to [52], verbs can be classified into three classes: Type A, Type B and Type C verb classes. They differ by vowel patterns for perfect and jussive tense descriptions. Class A is

unmarked class. In the base past and jussive it has two sub classes: A1 has stem vowel /ä/ in the past and stem vowel /a/ in jussive, whereas A2 has stem vowel /a/ in the past and stem vowel /ä/ in the jussive. B is the class of verbs geminating middle radical (second radical from the left). C is class of verbs with stem vowel /a/ after the first radical/consonant (see the details in Table 3.20). The internal plural formations of Ge'ez nouns are also changed by various consonant vowel patterns (see the details in Table 3.6 up to 3.11).

A process of compounding is another way to create various Ge'ez word forms. Compounding is the joining of two linguistic forms that are functional independents (see Table 3.22).

The Ge'ez language can also inflect with duplications and reduplications by attaching one affix with another affixes fully or partially. These types of word formations are created by the combination of prefixes, infixes and/or suffixes. For example, ተቀታትልክሙ'killed each other' (see the detail in Table 3.19).

3.4.INFLECTIONAL AFFIXES OF GE'EZ

As pointed out by [1], both inflectional and derivational morphologies can result for a very large numbers of variants of a single word. But, these variations depend on the morphological complexity of the language.

Inflectional affixes describe word stems are combined with grammatical markers for things like person, gender, number, tense and case. Nouns and verbs can be marked for these different grammatical markers. This is because verbs and/or nouns are very rich in morphological

characters to agree with person, number, and gender of subjects in Ge'ez language, the discussions are mainly focused on noun and verb morphologies.

To study the morphological natures of the language, the researcher used different sources such as [50], [51], [52], [53], [54], [55] and [56].

3.4.1. NOUNS

Ge'ez noun can be inflected for genders, numbers and cases. They have their own phonetic structures. The basic phonetic structures of the Ge'ez nouns consist of various character sequences.

3.4.1.1GENDER

Gender in Ge'ez is more important category because for one thing (e.g. for third person singular), gender distinguishes in second and third person pronouns for both singular and plural nouns differently. For example, for 2m.sg and 2f.sg gender is shown differently (see Table 3.2).

In Ge'ez, the gender markers are not limited. Gender is distinguished for both singular and plural, masculine and feminine. The gender marker to third person pronouns in Ge'ez nouns is the feminine marker: the suffix ት ¼t¼ to masculine nouns. For example; the feminine form of ብእሲ b'asi is ብእሲት b'asit 'man, woman'. There are also suffixes to indicate the masculine and feminine for possessive pronouns. Table 3.2 illustrates how the noun ዜና 'zena' is expressed with both genders.

Gender	Number	Noun	Suffixes
Masculine	2m.sg	ዜና -ከ 'zena-kä'	-ከ –kä
	3m.sg	ዜና -ሁ zena-hu	-ሁ -hu

	2m.pl	ዜና -ክሙzena-kmu	-ክሙ-kmu
	3m.pl	ዜና -ሆሙzena-homu	-ሆሙ-homu
Feminine	2f.sg	ዜና -ኪ zena-ki	-ኪ -ki
	3f.sg	ዜና -ሃ zena-ha	ሃ –ha
	2f.pl	ዜና -ክን zena-kn	-ክን –kn
	3f.pl	ዜና -ሆን zena-hon	-ሆን –hon

Table 3.2: Gender markers for possessive pronouns

The suffix –ክ, -ሁ, -ክሙ; -ሆሙ; -ኪ, -ሃ, -ክን and -ሆን are possibly employed Gender as well as possessive markers.

3.4.1.2 NUMBER

The Ge'ez noun system recognizes different number types, namely singular, plural, dual and plural of plural forms [53]. Ge'ez words having the different forms in their plurals are called dual forms. This form of plural is related to the dual form of Arabic (see Table 3.3) [53].

Singular	Dual/plural	plural of plural
ዕዘን	ዕዘን	አዕዛን
ግር	ዕገር	አጋር
ዕብን	ዕበን	አዕባን
ብርክ	ሠርክ	አብራክ
ደቅ	ደቂቅ	ደቃቀቅ

Table 3.3: Dual nouns formations

In the other case, the nouns can be singular or plural. Plural of plural refered to the number which is more than two (see in Table 3.4).

Singular	Plural	plural of plural
ንጉስ	ነገስታት	ነገስታት
ሊቅ	ሊቃናት	ሊቃን፡ ሊቃውንት
ከረምት	ክራምት	አክራም
ሣን	አሣን	አሥማ ፡ አሣአን
ነገር	አንጋር	ነገራት
ልፍ	አላፍ	አላፍት

Table 3.4: Plural nouns formation

The number markers such as suffix, prefix, infix or their combinations form the plural nouns. These number markers in Ge'ez are usually present in nouns, adjectives, and verb conjugations.

Generally, the Ge'ez nouns can be pluralized using two ways: using external plural markers and internal plural markers [51]. Nouns that are created by external plural markers are carried out by adding suffixes and/or prefixes at a stem. The following affixes are some of plural number indicates. These are the prefix አ- 'ä-, suffixes like -ያን –yan, -ኣን –'an, -ያት-yat, -ኣት–'at, -ው-w, -ዋት -wat , ሙ-mu and ት -t (see Table 3.5).

To use the external plural markers, there are some rules that help to create grammatical agreements in Ge'ez text. This is because word forms in Ge'ez language are very complex. Therefore, it is important pursuing the following rules [51]:

- Nouns having the fifth radical at the end add the suffix ያት-yat
- Nouns having the seventh radical at the end add suffix ዋት- wat
- Nouns having the third radical at the end add the suffix ያን - yan to masculine or ያት –yat to feminine.

- Nouns having the fourth radical at the end add the suffix ት-t. If the nouns having ended by the sixth radical, change the end radical to fourth and then add ት-t.
- When አ ‘ä is placed at the beginning of words, change the second radical of the last alphabet of a word in to the sixth order and change the third radical into the fourth order and may add -at. Sometimes because of the presence of the gutturals, the second radical may be changed in to the fourth order. See the following example (Table 3.5) to illustrate plural nouns formations.

Singular plural	**Plural**	**Prefix**	**infix**	**suffix**
ሰባኪ: säbaki	ሰባኪያን: säbakiyan			ያን -yan
ብሩህ : baruh	ብሩሃን: baruhana			ኣን-an
ፈረሳዊ: färäsawi	ፈሪሳዊያን:Färisawiyan		-እ- ‘a-	ያን -yan
ጽጌ: śage	ጽገያት: śagäyat		-አ - -‘ä-	ያት -yat
ብን : ban	በን :bäna		-አ-, -‘ä-	
መጻብሕ:Mäśäbaħa	መጻቢሕያን: äśäbiħayan		-አ- ’ä- , እ a	ያን-yan
ምሳሕ: masaħa	ምሳሐት: masaħät			ኣት-at
ቤት: bet	አብያት: ‘bayata	አ ‘ä-	ያ-ya-	

Table 3.5: Plural noun formations using affixes

Internal plural markers also create plural noun. These markers are used to make plural nouns for most tri-consonantal (three consonants) stem nouns and for some quadric-radical (four

consonants) stem nouns. These types of nouns are created by changing the alphabetical order (see Table 3.6). Infixes such as አ ʻä, ዐ ʻa, ዖ ya are enabled to create internal plurals. In general, plural nouns that are created by internal plural markers follow one of the following patterns [52]:

CCäC (consonant + consonant + vowel አ + consonant): this pattern is not that much familiarized to Ge'ez nouns. But there are some nouns which follow this pattern.

Singular	**Gloss**	**plural**
እዝን ʻazan	ʻear'	እዘን ʻazän
እግር ʻagar	ʻfoot'	እገር ʻagär
እኅ ʻa'ha	ʻbrother'	እኁወ፡ahew

Table 3.6: List of nouns that follow CCäC pattern

äCCaC (Vowel አ + Consonant + consonant + vowel አ + consonant): this pattern is the most common for three consonantal stem nouns.

Singular	**gloss (singular)**	**Plural**
ልብስ lbs	ʻgarment'	አ-ልብ-አ-ስ ʻälbas
ቤት bet	ʻhouse'	አ-ብ-የ-አ-ት ʻäbyat
ፈረስ färäs	ʻhorse'	አ-ፍር-አ-ስ ʻäfras
ጾም śom	ʻfast'	አ-ጽ-ው-አ-ም ʻäśwam
ስም sm	ʻname'	አ-ስም-አ-ት ʻäsmat

Table 3.7: List of nouns that follow ʻäCCaC pattern

äCCuC: nouns with this pattern precede a pattern as vowel ä + consonant + consonant + u followed by consonant (see Table 3.8):

Singular	**Gloss**	**Plural**
አድግ ‘ädg	‘ass’	አ-እድ-ኡ-ግ ‘ädug
ሀገር hagär	‘city’	አ-ህግ-ኡ-ር ‘ähgur

Table 3.8: List of nouns that follow äCCuC pattern

äCC(a)C-t: this pattern follow a sequence of Vowel ኣ + consonant + consonant + vowel ኣ + consonant followed by ት.

Singular	**Gloss (singular)**	**plural**
ብር br	‘birr’	አግብርት ‘ägbrt’
ራስ ras	‘head’	አርእስት ‘ärst
ግብር gbr	‘slave’	አግባርት ‘ägbart

Table 3.9: List of nouns that follow äCC(a)C-t pattern

Nouns with quadric-consonantal and some tri-consonantal nouns follow the following pattern. Tri-consonantal nouns that take this pattern must have at least one long stem vowel ኢ ‘i , ኤ ‘e, ኦ ‘o, ኡ ‘u. This pattern is **CäCaCC(t)** i.e. consonant + vowel ኣ + consonant + vowel ኣ +consonant + consonant /ት/. (See the detail from Table 3.10).

Singular	**Gloss**	**Plural**
ድንግል dngl	‘virgin’	ደናገል dänagl
መኳፍን mesfn	‘prince’	መኳፈንት mäsafnt
ኮከብ kokäb	‘star’	ከዋክብት käwakbt
ዶርሆ dorho	‘checken’	ደዋርህ däwarh
ሌሊት lelit	‘night’	ልያልይ läyaly
ባሕር bahr	‘earth’	በሃወርት bähawrt

ወሕዝ wh‘az	‘river’	ወሃይዝት wähayzt
ቀሲስ ḳäsis	‘priest’	ቀሳወስት qäsawst

Table 3.10: List of nouns that follow CäCaCC(t) pattern

Note that: the ት‘–t ‘ is added to the plural if it is absent in the singular and the noun is not feminine.

äCaCC/t/: it has a pattern of vowel ኣ + consonant + vowel ኣ + consonant + vowel ኣ +consonant /ት/. This pattern works to four stem words (see Table 3.11).

Singular	**Gloss**	**Plural**
በግዕ ‘bäg’a’	‘sheep’	አባግዕ ‘äbag‘a
ጋኔን’ganen’	‘devil’	አጋንንት‘ägannt

Table 3.11: List of nouns that follow äCaCC(t) pattern

3.4.1.3 CASES

There is one morphologically marked case form in Ge’ez, the accusative construct [52]. Accusative simply designed by “of” in glosses when it makes the possessive construct configuration, otherwise by “acc”. The accusative is formed by suffixation of –ä to the unmarked form of the noun. Thus, nominative bet ‘house’, accusative bet-ä. This form is used both for the direct object of the verb as in ሰርሃ ንጉስ ቤት-አ särha ngus bet-ä ‘the/a king built the/a house’ [built king [nom.], house [acc.]], and for the head (first) noun in the so-called possessive construct configuration as in bet-ä ngus ‘’the/a house of the/a king where house of king [nom.].

In both constructions, morphological indication of case can be replaced by syntactic paraphrases. In the case of the direct object, the construction verb noun [acc.] can be replaced by verb+ lä-

nouns where lä is the preposition ‘to’. Thus instead of särhä bet-ä ‘he made the/a house’ one can have särho (särhä+hu) läbet ‘he made the house’.

For the possessive construct of noun1 and noun2, there are two possibilities፡ either noun1 zä-noun2 or noun1 lä-noun 2. Thus instead of betä ngus, one can have either beta zängus or betu längus.

Therefore, nouns can have two cases in Ge’ez: nominative which is not marked and accusative nouns that are marked with suffixation of ኣ ‘–ä’ or the paraphrase of –ä such as ለ lä- or ዘ zä.

3.4.2. VERBS

Ge’ez verbs are formed from bi-radical, tri-radical and quadri-radical roots with two, three and four consonants respectively. However, the common Ge’ez verbs are tri-radicals. Generally, the numbers of Ge’ez-verb’s alphabets could not be less than two and could not be more than seven [46].

Ge’ez verb is used to create or form verbal nouns, adjectives and/or another verb forms (eg.infinitives, jussives etc). It helps to make an agreement among number, tense, and gender with nouns. To make this agreement, suffix, infix and/or prefix are added to a verb.

The main verbs in Ge’ez are perfection and imperfection. Perfection is usually past or completed action. It includes past perfect, past continuous, past participle with the relative pronoun ዘ ‘of’. The imperfect one is usually present, continuous and future action. The end of all perfect verbs is the first order while all imperfect verbs have in the end the sixth order except the verb ይበ ybe. These all are under the pronoun ውእቱ ‘He’. The other verbs categories are subjunctive, infinitive

and gerundive. The conjugation of a verb or a stem is highly dependent on gender, number and pronouns.

3.4.2.1 PERFECTIVE

In Ge'ez verb, the perfect natures of verbs use to create the past and completed actions. These types of verbs are the base to other forms that may inflect to their base. Table3.12 illustrates how suffixes are employed to perfect forms.

Persons	Perfect	Gloss
1p.sg	ቀተልኩ 'ቀ̣ätäl ku'	'I killed'
2m.sg	ቀተልከ 'ቀ̣ätäl ke'	'you killed
2f.sg	ቀተልኪ 'ቀ̣ätäl ki'	'you killed'
3m.sg	ቀተለ 'ቀ̣ätäl ä'	'he killed'
3f.sg	ቀተለት 'ቀ̣ätäl ät'	'she killed'
1p.pl	ቀተልነ 'ቀ̣ätäl nä'	'we killed '
2m.pl	ቀተልክሙ' ቀ̣ätäl kmu'	'you killed'
2f.pl	ቀተልክን 'ቀ̣ätäl kn'	'you killed'
3m.pl	ቀተሉ 'ቀ̣ätäl u'	'they killed'
3f.pl	ቀተላ 'ቀ̣ätäl a'	'they killed'

Table 3.12 Perfect verb formation

The morphemes that are employed in inflecting verbs in the perfect verb formations included the following suffixes -ኩ/–ku/, -ከ /-kä/, -ኪ/-ki/, -አ/-‘a/, -ኣት/at/, -ነ /-nä/, -ክሙ/-kmu/, -ክን-kn/, -ኡ/- ‘u/, -እ/-/ ‘ä. These are the general forms of perfect Ge’ez verbs.

3.4.2.2 IMPERFECT (NON-PAST)

The imperfect verb in Ge’ez language uses to describe non-past action. It uses suffixes and/ or prefixes. Table 3.13 illustrates how prefixes and suffixes are employed to imperfect verb forms.

Pronouns	**imperfect (future)**	**Gloss**
1sg	እቀትል ‘a - ḳätl’	‘I will kill’
2m.sg	ትቀትል ’t- ḳätl’	‘you will kill’
2f.sg	ትቀትሊ ‘t- ḳätl -i’	’you will kill’
3m.sg	ይቀትል ‘y- ḳätl’	’he will kill’
3m.sg	ይቀትላት ‘y- ḳätlät’	‘she will kill’
1p.pl.	ንቀትል ‘n- ḳätl’	’we will kill’
2m.pl	ትቀትሉ ‘t- ḳätl -u’	’you will kill’
2f.pl.	ትቀትላ ‘t- ḳätl -a’	‘you will kill’
3m.pl	ይቀትሉ ‘y- ḳätl -u’	‘they will kill’

3f.pl	ይቀትላ ‘y- ḳätl -a’	‘they will kill’

Table 3.13: Imperfect verb forms

From the above morphemes, the prefixes such as ኦ ‘a-’, ት ‘t-’, ይ ‘y’ and ን ‘n’ are employed. The present tense and future tense have similar pattern/forms in Ge’ez verb inflection. Generally, there are four possible prefixes to show imperfect verb inflection. These are ተ ‘tä-’, ኦ ‘ä-’, የ ‘yä-’, and ነ ‘nä-’.

3.4.2.3 SUBJECTIVE/JUSSIVE

Subjective describes an action depending up on a preceding verb of volition or conjunction. These types of verbs use to describe behaviors. Table 3.14 illustrates how subjective affixes are employed to subjective verb formation.

Persons	subjective/jussive	Gloss
3m.sg	ይቅትል/y- ḳtl	’let him kill’
1p.sg	ኦቅትል/ ’a- ḳtl	’let me kill’
2m.sg	ይቅትል/y- ḳtl	’let you kill’
2f.sg	ትቅትሊ/t- ḳtl –i	‘let you kill’
3f.sg	ትቅትል/t- ḳtl	’let her kill’
1p.pl	ንቅትል/n- ḳtl	‘let us kill’
2f.pl	ትቅትል/t- ḳtl	‘let you kill’
2m.pl	ትቅትሉ/t- ḳtl –u	‘let you kill’

3m.pl	ይቅትሉ/y- ḳtl –u	‘let them kill’
3f.pl	ይቅትላ/y- ḳtl –a	‘let them kill’

Table 3.14 Subjective verb formation

The employed affixes are y-, t-..i, t-, n-, t-..-u, y-…u, and y-…a. Jussive follows similar patterns with subjective. But in the second person pronouns, there is no any added prefix for jussives case. Commands follow similar pattern with subjective.

3.4.2.4 INFINITIVE

This is created by making the last alphabet 6th order. See Table 3.15 to illustrate how infinitive verbs are formed.

Infinitive	**Gloss**	**perfect**
ቀቲል/ ቀቲሎት‘ḳät -i-l/ot’	‘to kill’	ቀተለ ḳätälä
ቀዲስ/ቀዲሶት ‘ḳäd-i-s/ot’	‘to praise’	ቀደሰ ḳädäsä
ሐጺጽ/ሐጺጾት ‘ẖaṣ-i-ṣ/ot’	‘to decrease’	ሐጸ ẖäṣä
ወሂብ/ወሂቦት ‘wh-i-b/ot’	‘to give’	ወሀበ wähäbä

Table 3.15 Infinitive verb formation

The vowel ‘ä’ is transferred to –i- and add the suffix –ot. The above table is indicated to third person singular noun (3p.sg).

3.4.2.5 GERUND

This verb category is used to indicate an action whether it is done or not and express occurrences of things or not. This type of verbs cannot close the sentence. This is shown using ten pronouns

to the verb በልዐ ‘bäl‘ä’ in table 3.16:

Person	**Gerundive**	**Gloss**
1p.sg	በሊዕየ	‘I having eaten’
2m.sg	በሊዐከ	‘you having eaten’
2f.sg	በሊዐኪ	‘you having eaten’
2m.pl	በሊዐክሙ	‘you having eaten’
2f.pl	በሊዐክን	‘you having eaten’
1p.pl	በሊዐነ	‘We having eaten’
3m.sg	በሊዖ	‘He having eaten’
3f.sg	በሊዓ	‘She having eaten’
3m.pl	በሊዖሙ	‘They having eaten’
3f.pl	በሊዖን	‘They having eaten’

Table 3.16 Gerund verb formation

Gerundive is formed be changing a preceding of the last radical into third order and then add the suffixes ከ, ኪ, ክሙ; ክን and ነ to 2m.sg,2f.sg, 2m.pl, 2f.pl and 1p.pl respectively. The last radical is changed to 7th order for 3m.sg and to 4th order for 3f.sg. By changing the last radical to 7th order, the suffix ሙis added to 3m.pl and ን is added to 3f.pl nouns.

3.5 DERIVATIONAL AFFIXES OF GE’EZ

Derivation is the process of creating new words by adding affixes from existing words. Mostly, Ge’ez verbs can derive others such as verbal nouns, adjectives and adverbs, rather than derived from these.

3.5.1 NOUN DERIVATION

Verbal nouns can be derived from verbs which have not more than three radicals [50]. When nouns are derived from verbs, the first alphabet from the left and its follower are changed to 6th orders and then the third alphabet will change to first order, and then add ት‘t’ as suffix. See at table 3.17 to understand how verbal nouns are derived from Ge’ez verbs.

Verbs	**Derived nouns**
ሰበከ	ስብከት ‘ säbäkä, sbkät’
ስብሐ	ስብሐት ‘sbḫä, sbḫät’
ሞጽአ	ምጽአት ‘mäṣ‘ä, mṣ’ät’
ቀተለ	ቅትለት ‘ḳätälä, qtlät’

Table 3.17: Verbal noun derivation by changing alphabetical orders and a suffix ት -t

Verbal nouns can also be derived by changing only alphabetical orders (See Table 3.18).

Verbs	**Gloss (verbs)**	**derived nouns**	**Gloss(nouns)**
ፈጸመ ‘feṣṣämä’	Complete	ፍጹም ‘fṣṣum’	‘ completed’
ቀተለ ‘ḳätälä’	Kill	ቀታሊ ‘ḳätali’	‘killer’
ቀተለ ‘ḳätälä’	Kill	መቅተትል ‘mästeqtl’	’ killer’
ቀየሰ ‘ḳäyäsä’	Prise	ቄስ ‘ḳes’	'priest’

Table 3.18: Noun derivation from verb by changing alphabetical orders

3.5.2 VERB DERIVATION

Ge'ez verbs have an ability to create new words that are not similar with the original verb form. This new word form can be created by verb derivations [16].

As [46] discussed, Ge'ez verbs can appear in some or all of the following possible verb derived forms using prefixes: base (simple past, shown in Table 3.1), causative (prefix አ- 'ä-), passive reflexive (prefix ተ- 'tä-' , if not preceded by a subject prefix , otherwise ት t-), and causative passive (prefix አስተ- 'ästä-'), (see Table 3.19).

Simple past → causative → reflexive-passive → causative-passive			
3m.sg ቀተለ	---አቅተለ/ä-qt äl- ä /	----------ተቀትለ/tä-qätl-ä /	----------አስተቃተለ/ ästä -qatl-ä /
1p.sg ቀተልኩ	--አቅተልኩ/ ä-ḳtäl-ku/	---ተቀትልኩ/t ä-ḳätl-ku/	---አስተቃተልኩ
2m.sg ቀተልክ	--አቅተልክ/ ä -ḳtäl-k	ä/---ቀተልክ/t ä-ḳät	äl-kä/--አስተቃተልክ
2f.sgቀተልኪ	--አቅተልኪ/ ä-ḳtäl-ki/	----ተቀተልኪ/t ä-ḳätäl-ki/	---አስተቃተልኪ
3f.sgቀተለት	--አቅተለት/ä-ḳtäl- ät/	----ተቀትለት/tä-ḳätlät/	----አስተቃተለት
1p.plቀተልነ	-አቅተልነ /ä-ḳtäl-n	ä/---ተቀተልነ /tä-ḳätäl-nä/	----አስተቃተልነ
2m.pl ቀተልክሙ	-አቅተልክሙ/ä-ḳtäl-kmu/	--ተቀተልክሙ/tä-ḳätälkmu/	--አስተቃተልክሙ
2f.plቀተልክን	---አቅተልክን/ä-ḳtäl-kn/	---ተቀተልክን/tä-ḳätäl-kn/	---አስተቃተልክን
3m.pl ቀተለ	---አቅተሉ/ä-ḳtäl-u/	-----ተቀትሉ/tä-ḳätl-u/	--------አስተቃተሉ
3f.plቀተላ	---አቅተላ/ä-ḳtäl-a/	----ተቀትላ/tä-ḳtäl-a/	---------አስተቃተላ

Table 3.19: Verb derivations from verbs

From the above verb ‘ḳätälä ’ /he killed/, the causative and the reflexive passive show the meaning ‘he caused to kill’ and ‘he was killed’, and also tä-ḳätätäl-ä ‘ killed each other’, ‘ästä-ḳätäl-ä ‘he caused to be killed’. The others are translated in similarly fashion.

As discussed by [52], the whole lexicalized Ge’ez verbs can be designated with letter A, B and C. Type A is the unmarked class. In the base past and jussive, there are two subclasses of type A: A1 has stem vowel ኣ /ä/ in the past and stem vowel ኣ/a/ in the jussive, whereas A2 has stem vowel ኣ/a/ in the past and ኣ/ä / in the jussive. The corresponding to stem vowel alternations form to type A class are: past CäCäC→ present jussive CCuC, past CäCiC→present jussive CCäC.

Type B class of verbs is categories with geminating middle radical. Type C is the class of verbs with stem vowel /a/ after the first radical consonant. A verbal entry must be marked in the lexicon as either class A, class B or class C in Ge’ez, and if it occurs in one class, it will not occur in another. An exception to this general rule is the class of passive reflexive C (e.g. tänagärä). Table 3.20 illustrates Type A, Type B and Type C to perfect and Jussive tense categories as follow:

Types/verb classes	**Past**	**Jussive**
A	nägärä /ነገረ	ንጋር/ngar
B	fäṩṩämä /ፈጸመ	ፈጽም/fäṩṩämä
C	masän /ማሰን	ማስነ/msanä

Table 3.20: Verb classes

3.5.3 ADJECTIVE DERIVATION

There are also adjectives that are derived from Ge'ez verbs. The followings are possible adjectives derived from a verb ቀተለ: ቀታሊ ‘ kätälä, ḳätal-i’ to 3m.sg, ቀታልያን ‘ḳätal-yan’ to 3m.pl, ቀታሊተ ‘ḳätal-ite’ to 3f.sg, and ቀታልያት ‘ḳätal-yat’ to 3f.pl.

In derivational morphemes, suffixes such as ዊ and ይ are used to create adjective like ፈያታይ ‘The thief’,‘thief’ and ፈያታዊ ‘The thief’, ‘thief ’.

3.6 PLURAL OF PLURAL NOUNS

Plural of plural is the other phenomenon of Ge'ez nouns. It is an additional pluralizing of plural nouns (plural of plural). This can be created by adding suffixes such as -ኣት ‘-ät’. For example; to a plural noun ነገሥቱ nägäst ‘kings’ , the plural of plural can be formed as ነገሥታት ‘kings.

Adding prefix-suffix to plural nouns is the other possibility to create double plural nouns (see Table 3.21).

Singular	*Plural*	*plural of plural*	*Prefix-----Suffix*
ንጉሥ	ነገሥቱ	ነገሥታት	-ኣት
ሊቅ	ሊቃናት	ሊቃን፡ሊቃውንት	-ኣት
ደብር	አደባር	አደባራት	-ኣት
ዐምድ	አዕማድ	አዕማዳት	አ-...ኣት
ረምኽ	አርማኽ	አርማኻት	አ .....ኣት

Table 3.21: Plural of plural words formations

3.7 COMPOUNDING

Most compound words are created by combining two different words. For example; for a word ዲቤ dine, ዲቤ -በሐር /dibe -bäḫar/ ‘above (the) sea’ and ዲቤ-ተቅዋማ /dibe täqwama/ ‘Above the home of the candle ’ are possibly formed. Compound can also be created by inserting ኤ (e) as infix in Ge’ez noun for all pronouns (see the detail in Table 3.22):

Compound	Gloss
አሜየ [‘ämeyä]	‘In my time’ (1p.sg)
አሜከ [‘ämekä]	‘In your time’ (2m.sg)
አሜኪ [ämeki]	‘In your time’ (2f.sg)
አሜክሙ[ämekmu]	‘In your time’ (2m.pl)
አሜክን [ämekn]	‘In your time’ (2f.pl)
አሜነ [ämenä]	‘In our time’ (1p.pl)
አሜሁ [ämehu]	‘In his time’ (3m.sg)
አሜሃ [ämeha]	‘In her time’ (3f.sg)
አሜሆሙ [ämehomu]	‘In their time’ (3m.pl)
አሜሆን[ämehon]	‘ In their time ’ (3f.pl)

Table 3.22: Compound word formation

3.8 NEGATION OF GE’EZ VERBS

The common negation prefix in Ge’ez verb is ኢ/‘i-/. This is when it comes with perfective form of verbs. Table 3.23 illustrates negation of verbs;

Verbs	**negative of verbs**
ቀተለ/ḳätäl-ä	ኢ-ቀተለ ‘ ‘i-ḳätäl-ä ’
ይቀትል/y-ḳätl	ኢ-ይቀትል ‘i-y-ḳätl’
ይቅትል/y-ḳtl	ኢ-ይቅትል ‘ i-y-ḳtl’

Table 3.23: Negations of words

3.9 SUMMARY

This chapter discussed about Ge'ez language morphology: inflectional and derivational morphologies. Both the inflectional and derivational morphologies involve suffixing, infixing, prefixing. Hence, Ge'ez words are very complex morphologically.

The complexity of the language is one of the main reasons for a language to desire a stemmer. A stemmer helps to conflate variants of words as well as to access documents in natural language text.

Since verbs and nouns are very rich in morphological characters to agree with person, number, and gender of subjects in Ge'ez language, the discussions are mainly focused on noun and verb morphologies.

The internal and external plural nouns are also discussed in this chapter. External plural marker includes suffixes, infixes and prefixes. Internal plurals are created by changing the order of radicals.

The forms of Ge'ez verb such as perfect (past and completed actions), imperfect, jussive, subjunctive and infinitive are discussed. Its derivational characteristics: causative reflexive, passive reflexive, and causative passive are also seen.

The next chapter presents the development of a stemmer to conflate variants of Ge'ez words with respect to concepts in this chapter.

CHAPTER FOUR

DEVELOPMENT OF A STEMMER FOR GE'EZ TEXT

4.1. INTRODUCTION

This chapter presents the tasks that were done in developing a Ge'ez stemmer. It includes an affix removal and morphological analysis techniques. Furthermore, the preprocessing steps such as stop word and non-functional word removals are also part of the discussions.

Finally, a report is made on how the developed stemmer was done on the sample set of texts to evaluate its performance.

4.2. SAMPLE TEXT

For this work, sample texts were prepared from different sources: ወዳሴ ማርያም 'wdase maryam' (prayer book)[54], history of Abune Habtä Marriam [55] and Bible in Ge'ez (የሉቃስ ወንጌል)[7]. These documents are selected only because they are readily available to the research. The selection of topics and chapters from each resource was done randomly.

All sample texts were collected from hard copies and were typed using visual Ge'ez 2006. A python script was written to translate the sample texts into Latin equivalences. This was done to increase efficiency. For the purpose of the experiment, python2.6 and python3.0 codes were written. To test the performance of the stemmer, the whole sample data sets were used.

4.3. WORD DISTRIBUTION OF GE'EZ

As discussed by [9] and [14], word distribution in sample text documents of a language helps to study language's behavior, and this distribution can be shown using word-ratio (numbers of distinct words to total numbers of words), and percent frequency ratios (e.g. total words which have frequency equals to 1 to total numbers of words). These help to show how much words are morphologically distributed within a document.

Name of text	Total words	Distinct words	Word-ratios in percent	%of words with frequency 1	%wordswith frequency more than 1
Lukas	1,866	1,064	57.02	38.75	61.25

Table[2] 4.1: Number of words and their distributions to Ge'ez sample data sets

From table 4.1, words with frequency of one constitute 38.75 percent (%) of the total. This means more than one-third (1/3) of the total words in the sample texts composed of frequency equals to one. More than half of the sample texts were also distributed uniquely as shown word-ratios in table 4.1. This implies that there are existences of more variants of words in Ge'ez sample data set text.

To compare the sample sets of Ge'ez word distributions with other languages such as Tigrigna [9] and Amharic [16], the whole data sets were taken and shown in table 4.2. The data sets of Arabic and English languages were also adopted from [14].

[2]- Word-ratio is the ratio of numbers of distinct words to total words
- %of words with frequency equals to 1 is ratio of words with frequency of equals to one to total number of words
- % words with frequency more than 1 is ratio of words with frequency greater than 1 to numbers of total word

Language	Text	Total numbers of words	Distinct words	Word-ratio (distinct to total words)
Ge'ez	Lukas	1,866	1,064	57.02%
Tigrigna	Text1(SRUGGLE)	1,632	918	56.25%
Amharic	Text1(AMTHES)	4781	2663	55.70%
Arabic	Text1	1,600	902	56.38%
English	Text1	1,600	621	38.81%

Table 4.2: Comparison of word distribution ratios of the sample data set

The word ratios obtained from table 4.2 are almost similar to Ge'ez, Amharic, Tigrigna and Arabic texts. However, it is absolutely different from English text. This similarity among Ge'ez, Amharic, Tigrigna and Arabic might be due to the fact that each is in Semitic language groups. As shown in table 4.2, larger numbers of unique words are found in Ge'ez document when it was compared with other languages, especially with English. Hence, Ge'ez language has more distinct words and is morphologically very complex language when its word distribution is compared with others based on their sample data sets.

The zipfians law could also be used as an indication of complexity in the morphology of a language [9]. But, according to [16], this law is not obeyed by many languages. This law is done based on frequency of words as follow:

$$f*r=k$$

where f is a frequency of a word in document, r is rank of a word(in highest to lowest frequency arrangement) , and k is a constant value. The law says the product of the rank and frequency of a word gives us a constant value for all words of the text.

But, for example; for the sample data set, more than 700 words have a frequency of one, hence for a word which is settled at the end (i.e. at r=1,064) and a word at r=342, frequency to each word is 1. Therefore, k to the first and the second words are 1064 and 342 respectively, which are not matched. This means the zipfians law is not sufficiently support to conclude Ge'ez language's word distributions. Hence, it was left out to this work.

4.4. COMPILATION OF THE STOPWORD LIST

Stop words can be compiled from the sample text using rank-frequency distribution and by checking from list of known stop word lists (dictionary). The rank-frequency distributions of words, especially to a language which have high morphological complex words are not convenient to determine stop words. But, the frequency may guide in selecting stop words [16].

Some words such as prepositions, conjunction and articles in Ge'ez can exist affixed to words. For example, the adverb ነየ 'näyä' may exist as ነየኪ 'näyäki', ነየከ 'näyäkä', ነያ 'näya', ነዮሙ 'näyomu', ነዮን 'näyon', ነየክሙ 'näyäkmu', ነየነ 'näyänä', which makes identification problematic.

The common stop word list includes lists such as connecters, articles, infinitives and so on. It can also includes verb to be words such as አነ/' 'änä/ 'am, was', አንተ /'äntä/ 'are, were' አንቲ /'änti/ 'are, were', ወአቱ /w'atu/ 'is, was', ይአቲ/y'ati/ 'is, was', ንህነ /nhnä/ 'are, were', አንትሙ/äntmu/ 'are, were', አንትን/äntn/ 'are, were', ወአቶሙ/w'atomu/ 'are,were', ይአቲን /y'atin/ 'are, were'.

Demonstrative adjectives such as ዝንቱ zntu/z/ ‘this’ (2m.sg), ዛቲ z/zati/ ‘this’(2f.sg), እሉ ‘alu/‘alontu/ ‘these’(2m.pl), እላ ‘ala/’alantu/ ‘these’(2f.pl), ዝክቱ zktu/zku/ ‘that’(3m.sg), እንታቸቲ ‘antacäti ‘that’(3f.sg), እልክቱ ‘alktu ‘those’(3m.pl), alktu ‘those’ (3f.sg) are part of stop word lists.

Prepositions such as ዲቦ /dibo/ ‘on’, ላእለ /la’alä/ ‘with---on’, ታህተ /tahtä/ ‘down’, መትህተ /mäthtä/ ‘to---down’, ወስተ /wstä/ ‘in’, ዋስጤ /waste/ ‘within’, ማእከለ /ma’akälä/ ‘between’, እምኅ /’amn ä/ ‘from’, እም/’am/ ‘from’, ጊዜ /gize/ e.t.c are also included in stop list.

Most words in the sample text are occurred with frequency of one as shown in Appendix-II and discussed in section 4.3. Therefore, different words including stop words can exist in various forms due to affixes. Hence, it is better to use stop word list to remove stop words. For this work, lists of stop words dictionary was prepared from different books such as [51],[53], [54] and [55], and then automatic checking was done before stemming as well as during each stemming processes.

There are more than 241 stop word lists in the dictionary. The complete list of the stop word lists compiled from the sample text is given in Appendix III. See some stop word lists that are acquired from sample data set in appendix IV.

4.5. COMPONENTS OF THE STEMMER

The stemmer removes the affixes by applying various rules to each affix and this was done using an application of context sensitive rules. These rules are designed based on morphological natures of a language to each sequence of activities. There are also affixes that are stripped using

iterative approach with rules of the language. For example; lists of characters such as ወ ‘wä’, ለ ‘lä’, ዘ ‘zä’ and ከ ‘kä’ can be occured by concatenating each others as prefix of words.

The root of Ge’ez text can be obtained by stripping out all the vowels from the stemmed words [49] and [50]. But, when the first character of a word is vowel, it may be considered as consonants. For example, when አ ‘ä comes at the beginning of a word, if it is not removed as prefix, it is not considered as vowel and is not removed from a word. For instance from a word አንስት ‘änst’, ‘ä’ is considered as consonant.

In this paper, the stemming method incorporates two different stemming techniques. These techniques are: affix removal and morphological analysis techniques. The following sections describe the detail techniques of each method.

4.5.1. THE AFFIX-REMOVAL TECHNIQUES

Before this technique is applied, some common stop words from sample text are removed by matching with stop word lists. In the process of affix removals, the length of each word is also checked.

The stemming process is done whenever a word have more than two radicals or length of a word is greater than three depending up on the length of the affixes. This is because the minimum numbers of radical to Ge’ez words are two, and the most common words contain three radicals and above. The stemmed word, after affixes are removed is also checked with list of stop words and if it is part of stop words, it will be removed from the sample text.

This technique helps to determine all possible affixes (prefixes, suffixes and their combinations) that can be attached to Ge'ez words. List of common affixes with their rules are collected from different books such as [50], [51] and others. The common suffixes and prefixes commonly occur by attaching with Ge'ez words, but these have their own rules when they come with nouns and verbs. For example, the word ኮነ 'konä' and ኀጠኣትነ ' `haṭ'ätnä' both have the suffix '–nä' but nä is not removed from 'konä', but '-nä' is removed from ` haṭ'ätnä'. This is done using the assigned rules such as length of words, number of radicals and sequences of characters.

In the process of suffix or prefix striping, the length of words is checked iteratively, i.e. if the length of the stemmed word is greater than or equal to a suffix or a prefix, greater than two, and is not in the stop word dictionary, the suffix and/or prefix are stripped out from the word, otherwise, no stripping would be carried out. For example; a suffix ክሙ 'kmu' can be stripped out from a verb ቀተልክሙ'ḳätälkmu'. In this case, the length of ክሙ 'kmu' is not greater than the stemmed word length and is not in stop word list. Hence, the last three letters are seen as a suffix. In second example, the word ዜናየ 'zenayä', stemming can be done by removing the suffix የ 'yä' from the word to produce the stemmed word *zena('news').* But, this does not mean that the length of affixes is always less than the length of the stemmed word. For example, from a word 'ḳätlikmunä' the length of suffix (-ikmunä) is greater than the length of the stemmed. But, this can be caught using rules. Here, the length of examined word without considered affix is checked and if it satisfies the condition and number of radicals is not less than three, the assigned actions are taken based the appropriate rules.

Basically three actions are taken in the stemming processes. Only the two actions are applied to affix removal. These are:

Action1: do not remove any affix

Action2: remove the concerned affix

The third action is applied in morphological analysis technique (at section 4.5.2). To take any one of the above actions, there are conditions that are used to check the rules and apply an action 1 or action 2. These are:

Condition 1. After getting the assumed prefix or suffix, if number of radicals is not more than two or length of words less than three for a word without affix, take action 1.

Condition 2. If part of the assumed affix is obtained and number of radical without assumed affix is greater than two or length of a word greater than three and is not in stop wordlists, take action 2.

In the stemming process, based on the above conditions the appropriate action is taken. To take each action in the conditions, there are rules that are designed to each sequence of characters of words. For example to remove the prefix አ ‘ä, it is necessary seeing each follower characters and the third character from the left and so on in addition to word and radical lengths.

In removing of affixes, the sequences that are done in this work are: [checking word length], [checking prefix], [prefix removing], [suffix checking] and [suffix removing]. In each activity, there are stemmed word length checkers, stop word checkers and character sequence checkers. The followings are parts of affix removal techniques that are designated to this work.

4.5.1.1. PREFIX STRIPING TECHNIQUE

This procedure takes a word and checks the existence of a true prefix, and removes it whenever the condition is fulfilled. For example a verb which begins with መ‘me’ and its last two characters are sixth order and its length more than three, then መ‘mä’ is a prefix of a word. For instance ዘማሪ zämärä→ማዘምር mäzämr, መሃረ mäharä→መምህር mämhr, ፈከረ fäkärä→መፈክር mäfäkr and so on.

The common prefixes that may attach to the left side (beginning) of Ge’ez words are ኣ‘ä’, ኢ‘i’, እ‘a’, መ‘mä’,ሙ‘mu, ተ‘tä’,ቲ‘ti’,እን‘än’,አስተ‘ästä,ታ‘ta’, ት‘t’, ኣ‘a’, የ‘yä’, ያ‘ya’, ይ‘y’, ኑ‘nu’, ና‘na’, ን‘n’, ወ‘wä’, ዘ‘zä’ and ለ‘lä’.

The basic rules that check whether the assumed prefix is true prefix or not are checking word length (represents the length of a word without a prefix), prefix structure (represents a prefix and its follower, end of alphabet that could be attached to a word and so on) and whether the word is part of stop word list (represent whether a stemmed word is part of stop lists or not). These rules are used to minimize the over stemming and under stemming problems. To strip prefixes, the following algorithm is used.

1. Get WORD

2. Open stop word files

Read a WORD from the file until match occurs with stop word lists or End of a File reached

IF word exists in the stop word list

Remove a word

3. Count number of radicals (consonants) of a WORD and length of a word

4. If number of radical (length of a word) <=2, stop and Return WORD

ELSE:

IF length of prefix>length of stemmed, then stop and return WORD

ELSE IF length of WORD <3, stop and return WORD

ELSE GOTO step 5

5. If length(word)<length(prefix)+3 or the first length(WORD)-length(prefix) in stop lists, then remove a word

Else go to step 6

6. Check the rule:

If it satisfies the rule GOTO 7.

Else return WORD.

7. Remove prefix

8. IF number of radicals of stemmed WORD >2 and a WORD is with prefix, THEN GOTO 2

ELSE stop and Return WORD

9. IF end of file not reached

Go to 1

ELSE

Stop processing

Figure 4.1: Prefix striping algorithm

4.5.1.2. SUFFIX STRIPING TECHNIQUE

This process is done after prefixes are released out from the lists. Like that of prefix striping process, at each action the whole documents are checked with sequences of rules before stemmed words.

The common suffixes that appear in Ge'ez words are ን'n', ት 't', ው 'w', ይ 'y', ኡ 'u', ኣ 'a', ም 'm', ክ 'k', ን 'n', ሁ 'hu', ሃ 'ha', ሙ 'mu', አን ' 'an', አት 'at', 'አት' ot, አም 'am', ከ 'kä', ኩ 'ku', ያት 'yat', ያ 'ya', ዊ 'wi', ና 'na', የ 'yä', ኪ 'ki', ነ 'nä', ክን 'kn', ያን 'yan', ክሙ 'kmu', ዋት 'wat', ወያ 'wya', ሆሙ 'homu', እ ' 'a', ክሙ 'kämu' and their combinations.

In the process of suffix striping, word length and number of radicals which represent the length of a word without suffix. Suffix checker (represents a condition that checks whether the assumed suffix is true suffix or part of a word). If the above conditions are fulfilled and a word without part suffix is not found in stop word lists, suffix striping is done based on the assigned rules. The following algorithm is used to strip suffixes.

1. *Get WORD*
2. *OPEN stop word files*

   Read WORD from the file until match occurs with stop word lists or reached at End of File

   IF WORD exists in the stop word list

   Remove WORD

   ELSE

   Count number of radical and length of WORD
3. *If number of radical<=2*

   Return WORD

   Else

   If length of words <3 then stop and return WORD

   Else GOTO step 4
4. *IF length(WORD)<=length(SUFFIX)+2 OR length(WORD without SUFFIX)<length(SUFFIX), then stop and return WORD*

   ELSE

   IF the first length (WORD)-length (SUFFIX) in STOPLISTS, THEN remove WORD

   ELSE apply RULES and check them

   IF satisfy RULES, GOTO step 5

   ELSE stop and then return WORD
5. *Remove suffix*
6. *IF number of radical of a stemmed WORD >2 and a WORD is with suffix, THEN GOTO 3*

   ELSE stop and Return WORD
7. *IF end of file not reached, Go to 1*

   ELSE

   Stop processing

Figure 4.2: Suffix striping algorithm

4.5.2. THE MORPHOLOGICAL ANALYSIS TECHNIQUE

This technique helps for minimizing improper removal of some affixes from words which have main letters that appear as infixes and suffixes. This can be achieved by patterns of a language. These patterns are collected from various Ge'ez books such as [50] and [53]. From words which do not satisfy the pattern, the affix is considered as part of a word. Hence, such types of affixes are not true affixes.

This method is used to stem words with 3 or 4 radicals. In this technique, there is one action and one condition to take the action. These are:

Action3: replace an affix with other affix or remove vowel or consonant with other consonants.

Condition 3: If the assumed affix is part of an identified consonant-vowel structure in consonant - vowel patterns [In this case action 3 will be taken by the stemmer]. This condition is done after the possible prefixes and suffixes are removed.

The rule includes checking word length and checking consonant vowel patterns. When the conditions of rules and word patterns are matched, action 3 is taken.

This technique helps to convert the internal plural words (infix) to their singular forms. Affixes that are not removed by affix removal can be simplified by this technique. For example; a word ደናግል 'dänagl' and አብያት 'äbyat' could not be stemmed by affix removal technique. But, this method could handle them and change to singular word pattern (using C1äC2aC3C4→C1C2C3C4 and äC1C2aC3→C1eC3 forms respectively).

This technique is very difficult unless the words' patterns are known. Therefore, it helps to minimize number of unstemmed words. The rules are taken to both nouns and verbs in one. This causes to unknown word creations. Therefore, it may be possible to improve the performance of the rule by separating rules based on part of speeches. For example, a rule to noun should be implemented to nouns only because when we implement one rule that is not designed to certain documents, it causes to decrease efficiency of the stemmer.

Words which have prefix but not striped from the word are also held and stemmed with it. For example; ደብር →አድባር 'däbr→ 'ädbar', ፈረስ→አፍራስ 'färäs→ 'äfras', and so on. In these types of nouns, the prefix አ 'ä' not removed using prefix striping technique because the condition is not fulfilled, but this technique can solve this problem. See the following examples how one structure of words is changed to other form:

Unstemmed words' forms Stemmed words' forms

C1C2eC3----------→C1C2C3 or C1C2

e.g. 'äbäw----eb

'ager------'agr

C1äC2aC3C4/t/----------→C1C2C3C4 if not have t,

e.g. dänagl-----dngl

C1äC2aC3C4/t/----------→C1äC2C3C4 if a word has ት 't' at the end and not have ው 'w', C1oC3C4o if c2='w' and has not 't',

e.g: mäsafnt--------mäsfn

Säwarh-------Sorho

C1äC2aC3C4/t/----------→ C1äC2C4 if C3=’w’ and has ‘t’ at the end.

e.g. ḳäsawst---------ḳäss

äC1C2uC3----------→äC1C3

e.g. äddug----------ädg

C1äC2äC3ä----------→C1C2aC3

e.g. nägärä----------ngar

where C refers to radical (consonant) with in a word.

Table 4.3 Sample structural analysis of Ge’ez words (verbs and nouns)

4.6. EVALUATION OF THE STEMMER

To evaluate the performance of the stemmer, manual counting technique was used. This helps to compare numbers of errors that are not conflated correctly with the correct one. In the stemming process, three types of errors are observed. These are under stemming, over stemming and structural errors. See some errors of the stemmer in Table 4.4:

Words	Resulting stem	Expected stem	Error type
`slstu	`slst	‘sls	Under stemmed
Heqefkiyu	Hef	Hqf	Structural
IgziebHEr	GzebHr	IgzebHr	Over stemmed
Imnske	Sk	Nsk	Structural
Eyesus	Yss	Iyss	Over stemmed
Kemeznu	Kz	Kmz	Structural
Genete	Gt	Gnt	Structural
Weizeiyasekr	Ziyskr	Askr	Structural
Wereseyene	Ry	Rs	Structural
YaIqob	Iqb	YIqb	Over stemmed
YtwaHyu	TwHy	Why	Under stemmed

Table 4.4: Examples of the stemming errors

Using manual assessments of the stemmer, under stemmed holds 4.27 %, over stemmed covers 6 % and due to structural problems 7.31% from sample data sets are observed. Totally this version of the stemmer generates 17.58% stemmer errors. Consequently, the accuracy of the stemmer becomes 82.42 %.

The errors that are observed from the stemmer could be due to the following reasons:

1. Because of the complexity of the language, it was difficult to come up with the complete list of affixes at a time.
2. More conditions/rules are required based on a detailed study of the morphology of the language.
3. The whole rules are designed to common words. There are exceptional words and these may cause to increase stemmed errors.

The stemmer is also evaluated with percentage of compression. For calculating the compression rate(C), the expression used by [9] was used. The expression is shown as follow:

$$C = \frac{100*(W-S)}{W}$$

,Where W is the total word of the text and S is the total stem or root numbers.

Figure 4.3: Expression for measuring compression rates

The dictionary size and the compression obtained for stem and root from sample text are given as follow:

Number of stems 29.90 % reduction

Number of roots 62.8% reduction

4.7. SUMMARY

When a stemming process is done, over stemming and under stemming problems are observed from affix removal technique, and structural problems are mainly observed from morphological analysis technique. The last problem may be because both a verb and a noun may have similar consonant vowel pattern, and stemmed equivalently. Because the structural patterns for nouns and verbs are designed in one.

For experiment, there are 1866 words, which are collected randomly from available sources. From these data, performance of the stemmer is evaluated and the result shows 82.42 % accuracy.

For these data sets of words, 29.90 % compression of stemmed words and 62.8% compression of root words were found. The total errors that are observed from the experiment were 17.58 %.

The stemmer conflated derivational and inflectional affixes. It is not conflate irregular and compound words. These words are very complex and follow different (non-constant structures) patterns.

The next chapter presents conclusions of findings and recommendations for this work and future research.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

Ge'ez is a language with root pattern structure typical of Semitic languages. A single word in the language has a number of variants. Ge'ez is rich in both inflectional and derivational morphologies. The main word formation process is done through affixation. It uses prefix, infix, suffix and reduplication of a word. Longer words can be created in Ge'ez text via the concatenations of affixes or reduplications of words.

As shown in chapter three, Ge'ez's word distribution ratio shows the complexity of it as compared to other languages such as English. The analysis of word ratio of distinct words to total words calculated from sample text showed that Ge'ez is highly morphological complex language than English. But, it is related with Amharic by [16] and Tigrigna by [9] though there are some differences. Its most words are distributed throughout the texts and are singletons.

Stemmers developed for other languages could not be applied for this language. This is because of the morphological complexity and differences in feature of the language as discussed by [16]. However, commonly used methods of stemmers such as using stop word lists and context-sensitive rules are employed. Some techniques are also adopted from [9] and [16] in developing the stemmer.

The stemmer contains procedures of affix removals such as prefix and suffix removals. It also holds structures to verbs, and internal plural checker and converts to singular forms of nouns.

In the experiment, the developed stemmer was evaluated 1866 words which are selected randomly from available documents. From the experiment of the stemmer, the number of under stemmed covers 4.27%, over stemmed words holds 6% and structural errors covers 7.31% from the sample texts. Hence, the total errors account 17.58%. Therefore, the performance of the stemmer is 82.42%.

In terms of dictionary size, the compression of the stemmer became 29.90% for stemmed data and 62.8% to root sample data set.

In this work, rules are not enough when compared with the language's complexity. This is due to the limitation of time and complexity of the language. In general, conflation algorithms have inherited limitations and certain linguistic problems that are common to all conflation algorithms [14]. The two words that have the same underlying stem refer to the same concept. However, this is not always the case since sometimes words of the same stem need to be distinguished while words which are essentially equivalent may mean different things in different contexts. As a result, it is expected that such systems increase the errors numbers in addition to lack of rules.

Although additional rules and conditions may necessary to increase the success and achievement of the stemmer, the accomplished result of this work is comparatively balanced when compared with other stemmers that are developed for other languages. Hence, the experiment shows that the proportion of errors does not diminish retrieval effectiveness too much.

However, the performance of the work can be improved if different preconditions are fulfilled: corpus and sufficent soft copy documents as well as additional rules are prepared (this is recommended for future works).

5.2 RECOMMENDATIONS

The study in this work is done on the limited size of sample texts and not tested in IR environment due to time constraints and limitation of freely available Ge'ez texts in soft copy. So, it is possible to design a better stemmer by applying on a corpus of soft text documents.

Since this stemmer is the first trial of Ge'ez language, improving the stemmer is better to attain better performance, and to come up with operational levels. As a result, the following recommendations are suggested:

- One can add more rules in order to increase the accuracy of this stemmer.
- One can do the stemmer by designing taggers which able to differentiate part of speeches and stemmed each based on their assigned rules.
- Different approaches such as N-gram approach, Iterative approach, and/or their combinations can be applied to see whether better performance can be achieved or not.
- Experimenting the stemmer in IR environment to measure its performance in actual retrieval session
- Studying the effect of ordering the stripping procedures on the performance of the stemmer and selecting the best possible order.

- After improving the algorithm to its appropriate level, the stemmer can be an important tool for those researchers who are interested to study the Ge'ez language morphology.
- It is possible to use the stemmer by incorporating other components for developing other computational tools like morphological analyzer, parser, spell checker, thesaurus, word frequency counting, and summarizers.

REFERENCES

[1] Alemayehu,N and Willett,P. (2002), Stemming of Amharic words for information retrieval , Literary and Linguistic computing ,vol.17,No.1, pp.1-17

[2] Wakshum Mekonnen. (2000). Development of Stemming Algorithm for Afaan Oromo Text. M.Sc. Thesis . Addis Ababa Universty

[3] Ananthakrishnan Ramanathan & DurgeshD.Rao, ALightweightStemmer for Hindi, National Centre for Software Technology, NaviMumbai400614, India,pp.1-6

[4] Atelach Alemu Aregaw and Lars Asker, (2007), Amharic Stemmer: reducing words to their citation forms, Association for computational linguistic proceeding of the 5th work shop on important unresolved matters, Swiden, Czech Republic, pp.104-110

[5] Basem O. Alijla , (2009), Stemming in Natural Language, Faculty seminar, Department of Information Technology System, Faculty of Information Technology, The Islamic university of Gaza

[6] Birungi, P. (1995). Improved Strategies for Employment and Human Resources Utilization in the Information and Documentation Sector. Strategies for Human Resources Development for information Management in Africa, Ed. 49-57. Addis Ababa: UNECA, PADIS.

[7] Ethiopian Orthodox church teaching, The Ge'ez Language and Kine (Poetry, http://www.eotc-patriarch.org/index.htm

[8] Gabriella F. Scelta, (2001), The Comparative Origin and Usage of the Ge'ez writing system of Ethiopia, pp.1-9

[9] Girma Berhe (2001), Stemming Algorithm Development for Tigrigna Language Text Document, M,Sc Thesis, Addis Ababa University.

[10] Grishman,R. (1984), Natural language processing JASIS 35(5), pp.291-296

[11] Hayder K. Al Ameed, Shaikha O. Al Ketbi, Amna A. Al Kaabi, Khadija S. Al Shebli,Naila F. Al Shamsi, Noura H. Al Nuaimi, Shaikha S. Al Muhairi (2005),ARABIC LIGHT STEMMER: ANEW ENHANCED APPROACH, The Second International Conference on Innovations in Information Technology (IIT'05), College of Information Technology, UAE University

[12] James Mayfield and Paul McNamee (2003), Information Storage and Retrieval content analysis and indexing, Single N-gram Stemming.

[13] Jelita A.Hugh , E. Williams & S.M.M. Tahaghoghi (2005), Stemming Indonesian language, 28th Australasian Computer Science Conference (ACSC2005),Conferences in Research and Practice in Information Technology, Vol. 38, pp.1-8

[14] Lemma Lessa (2003), development of stemming algorithm to wolytta text, M.Sc Thesis, Addis Ababa University.

[15] M.Taufik Abdullah, F.Ahmad, R.Mahmod, T.Mohd & T.Sembok. (2009), Rules Frequency Order Stemmer for Malay Language , IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, pp.433-438

[16] Nega Alemayehu and Peter Willett (2003), the effectiveness of stemming for information retrieval in Amharic words for information retrieval. Journal: electronic library and information systems, Vol.37, num.4, pp.254-259

[17] Haidar Harmaneni, Walid Keirouz, Saeed Raheel (2006), A rule based extensible stemmer for information retrieval with application to Arabic, The International Arabic Journal of Information Technology, Vol.3, No.3,pp. 265-272

[18] Ricardo Bauza-Yates & Bertnier Ribeiro-Neto (1999), Modern information Retrieval, India

[19] Salton, Gerald (1983), "An Introduction to modern Information Retrieval", New York.

[20] AbduelBaset M., Gowede.R, Husien A, and Abdulselam M. (2008), "A Hybrid method of stemming for Arabic Text" Libya .pp.1-7

[21] David A. Hull (1995). Stemming Algorithms - A Case Study for Detailed Evaluation pp.1-27

[22] S.Srinivasan and ZP.Thambidurai (2006), Stans Algorithm for Root Word Stemming, Information Technology Journal 5 (4): 685-688, India.

[23] H.Harmanani, W.Keirouz and S.Raheel (2006). The international Arab Journal of information technology, vol.3, No.3, A Rule-based Extensible Stemmer for Information Retrieval Application to Arabic, pp.265-271.

[24] J. Savoy (1993). Stemming of thus helping the analysis of a sentence syntactically. Words on Grammatical, Categories, Journal of American Society for Information Science, 44(1), pp. 1-9.

[25] Porter, Martin (2003) .The Porter stemming algorithm at http://www.tartarus.org/~martin/PorterStemmer/, pp.1-4

[26] Porter, M.F (1980). An algorithm for Suffix Stripping. Program, 14, 130-137.

[27] Wessel Kraaij and Renee Pohlmann , Porters stemming algorithm for Dutch. Pp.167-180

[28] B.L.Narayan and SankarK.Pal, Distribution based stemmer refinement Machine intelligence unit, Indian Statistical Institute, Calcutta- 700108,India. Pp.1-6

[29] James Mayfield and Paul McNamee(2003), Single N-gram Stemming , The Johns Hopkins University Applied Physics Laboratory, Toronto, Canada.pp.415-416

[30] F. Çuna Ekmekçioglu, Michael F. Lynch, and Peter Willett (1996), Stemming and N-gram matching for term conflation in Turkish texts, Information Research, Vol. 2 No. 2, pp.

[31] J.Leveling, A comparison of sub-word indexing methods for information retrieval, Centre for Next Generation Localization (CNGL), DublinCityUniversity, Dublin9, Ireland

[32] Don Erick J.Bonus, The Tagalog Stemming Algorithm (TagSA), Jose Resal University, Mandaluyong Phillippines 1552.

[33] Djoerd Hiemstra (2001), A book for Using language models for information retrieval

[34] Ilia Smirnov (2008), Overview of Stemming Algorithms ,DePaul University (algorithm is given here based on Dr. Porter description).

[35] Julie B. Lovins (1968), Development of a Stemming Algorithm, Mechanical Translation and Computational Linguistics, vol.11, nos.1 and 2, pp.22-30

[36] Tesfaye Biru (1987). Incorporation of relevance Data in the Term Discrimination Value. The University of Sheffield (unpublished).

[37] Eduard Heindl, Stemming Technology: E-Business Technologies at http://webuser.hs-furtwangen.de/~heindl/ebte-09ss/stemming-technology.html

[38] Ibrahim Abu EI-Khair(2006), Effects of stop words Elimination for Arabic Information retrieval: A comparative study, International Journal of computer and information sciencesvol. 4, No.3.minia university- Egipt.

[39] Guanglai Gao, Wei Jin, Fei Long, Hongxu Hou (2008), A First Investigation on Mongolian Information Retrieval, The Second International Workshop on Evaluating Information Access (EVIA), Tokyo, Japan

[40]. Harman,D (1999). “How effective is suffixing”, Journal of the American society for the information science. 42, pp.7-15

[41] Frakes, William B. (1992). Stemming Algorithms. Information Retrieval: Data Structures & Algorithms. New Jersey: Prentice Hall PTR

[42] Potsdam (2009), Linguistic Theory and African Language Documentation , 38th Annual Conference on African Linguistics: Inflectional vs. Derivational Morphology in Tagdal Cascadilla Proceedings Project Somerville.

[43] Abreham Dilnesaw (2003), Addis Ababa university School of Graduate Studies word formations in Oyda, Thesis iin addis ababa university.

[44] Jose M.Gni-Menoya,Jilio Villena-Roman, Ana M.Garcia-Serano (2007), miracles Hybrid Approach to bilingual and monolingual Information Retrieval. university of Madrid.

[45] P.Nakov (2003). Building an Inflectional Stemmer for Bulgarian International Conference on Computer Systems and Technologies - CompSysTech'

[46] Muluken Andualem, Ge'ez Verb Classification in the Three Traditional Schools of Qəne, sited at http://etd.aau.edu.et/dspace/bitstream/123456789/1363/1/Muluken%20Andualem.pdf

[47] Ghelawdewos Araia (2004), Institute of development &Education for Africa,Inc. At http://www.africanidea.org/index.html

[48] Anna Siewierska and Dik Bakker, Inclusive and exclusive in free and bound person forms, in Lancaster University and University of Amsterdam, pp.162 -190

[49] Bernice Varjick Hecker (2007), The Biradical Origin of Semitic Roots , Thesis, Presented to the Faculty of the Graduate School of the University of Texas at Austin

[50] Dillmann, the Ethiopic language book (1958), vo.2, n.3, p.107-112

[51] ሊቀ ጉባኤ አባ ተ/ሃይማኖት ወልዱ (2001), ቀላል የግዕዝ ቋንቋ መማሪያ መጽሐፍ በቅዱስ ጳውሎስ መንፈሳዊ ኮሌጅ

[52] Robert Hetzron, The Semitic language Book: Ge'ez (Ethiopic) , p.241-265.

[53] መምህር ደሴ ቀለብ (2002), የግእዝ ቋንቋ መማሪያ መጽሐፍ-1, Addis Ababa University (unpublished)

[54] ተስፋ ገብረ ስላሴ ዘቢሐረ ቡልጋ (1992), የዘወትር ጸለት፣ ወዳሴ ማርያም ፣ገጽ 31-36

[55]አሳታሚ የደብረ ብፁዓን ገዳም (1995)፣ ሚኪከአ አቡነ ሀብተ ማርያም ታሪክ፣ አዲስ አበባ፣ ገጽ 3-19

[56] በላይ መኮንን ሥዩም (2000)፣ ሕያው ልሳን ግእዝ -አማርኛ መዝገበ ቃላት

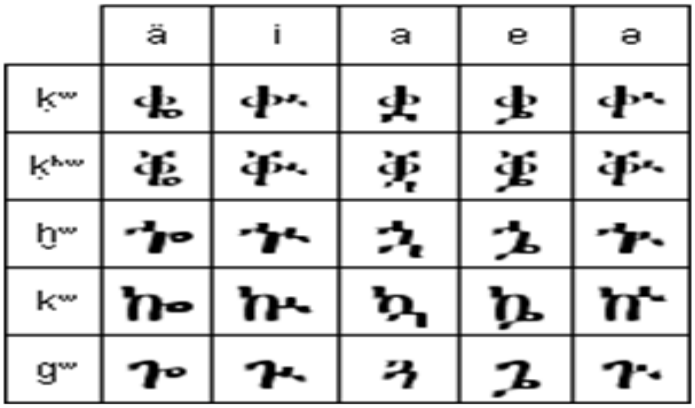
APPENDIX I: TRANSLATION OF GE'EZ SCRIPTS.

In general, the translations applied in this thesis are as follows:

A. The Ge'ez script (Zelalem, 2001)

	ä	u	i	a	e	ə	o	wa	yä
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ		
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ	ሏ	
ḥ	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ	ሗ	
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ	ሟ	ሟ
ś	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ	ሧ	
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ	ሯ	ሯ
ś	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ	ሷ	
ḳ	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቋ	
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	ቧ	
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ	ቷ	
ḫ	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኋ	
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ	ኗ	
ʾ	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ	ኧ	
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኰ	
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ	ዏ	
ʿ	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ		
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ	ዟ	
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ	ዯ	
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ	ዷ	
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ	ጓ	
ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ	ጧ	
p	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ	ጷ	
ṣ	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ	ጿ	
ḍ	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ	ፇ	
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ	ፏ	ፚ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ	ፗ	

Letter Variants (labiovelars)



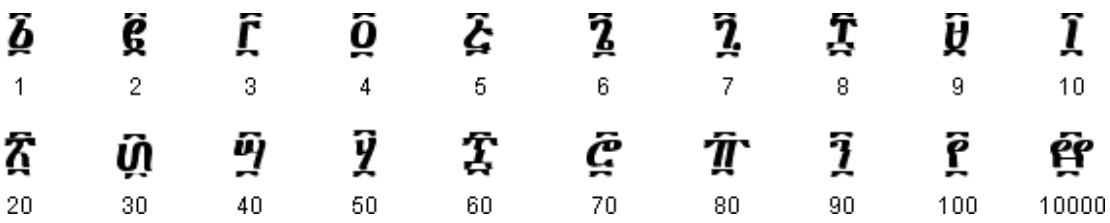
	ä	i	a	e	ə
kʼʷ	ቈ	ቍ	ቋ	ቌ	ቍ
kʰʷ	ኈ	ኍ	ኋ	ኌ	ኍ
gʷ	ጐ	ጕ	ጓ	ጔ	ጕ
kʷ	ኰ	ኵ	ኳ	ኴ	ኵ
gʷ	ጐ	ጕ	ጓ	ጔ	ጕ

B. Translations of characters for implementation:

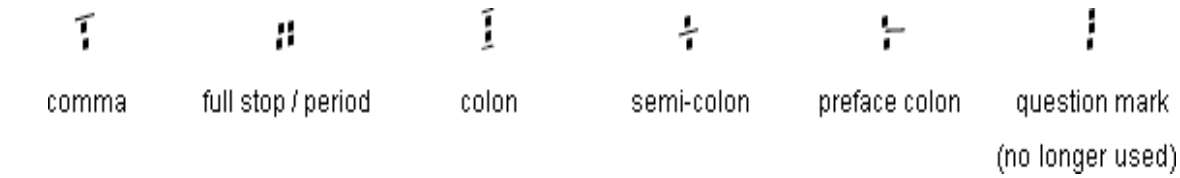
ሀ	he	አ	e	ፈ	fe	ኑ	nu	ጱ	Su	ቺ	ci
ለ	le	ከ	ke	ፐ	pe	ኙ	Nu	ፁ	`Su	ኺ	`hi
ሐ	He	ኸ	`ke	ሁ	hu	ኡ	u	ፉ	fu	ኒ	ni
መ	me	ወ	we	ሉ	lu	ኩ	ku	ፑ	pu	ኚ	Ni
ሠ	`se	ዐ	`e	ሑ	Hu	ኹ	`ku	ሂ	hi	ኢ	i
ረ	re	ዘ	ze	ሙ	mu	ዉ	wu	ሊ	li	ኪ	ki
ሰ	se	ዠ	Ze	ሡ	`su	ዑ	`u	ሒ	Hi	ኺ	`ki
ሸ	xe	የ	ye	ሩ	ru	ዙ	zu	ሚ	mi	ዊ	wi
ቀ	qe	ደ	de	ሱ	su	ዠ	Zu	ሢ	`si	ዒ	`i
በ	be	ጀ	je	ሹ	xu	ዩ	yu	ሪ	ri	ዚ	zi
ቨ	ve	ገ	ge	ቁ	qu	ዱ	du	ሲ	si	ዠ	Zi
ተ	te	ጠ	Te	ቡ	bu	ጁ	ju	ሺ	xi	ዪ	yi
ቸ	ce	ጨ	Ce	ቩ	vu	ጉ	gu	ቂ	qi	ዲ	di
ኀ	`he	ጰ	Pe	ቱ	tu	ጡ	Tu	ቢ	bi	ጂ	ji
ነ	ne	ጸ	Se	ቹ	cu	ጩ	Cu	ቪ	vi	ጊ	gi
ኘ	Ne	ፀ	`Se	ኍ	`hu	ጱ	Pu	ቲ	ti	ጢ	Ti

ጪ	Ci	ና	na	ሄ	hE	ዔ	`E	ስ	s	ግ	g
ጲ	Pi	ኛ	Na	ሌ	lE	ዜ	zE	ሽ	x	ጥ	T
ጺ	Si	ኣ	a	ሐ	HE	ዠ	ZE	ቅ	q	ጭ	C
ጿ	`Si	ካ	ka	ሜ	mE	ዬ	yE	ብ	b	ጵ	P
ፊ	fi	ኻ	`ka	ሤ	`sE	ደ	dE	ቭ	v	ጽ	S
ፒ	pi	ዋ	wa	ሬ	rE	ጀ	jE	ት	t	ፅ	`S
ሃ	ha	ዓ	`a	ሴ	sE	ገ	gE	ች	c	ፍ	f
ላ	la	ዛ	za	ሼ	xE	ጠ	TE	ኀ	`h	ፕ	p
ሓ	Ha	ዣ	Za	ቄ	qE	ጨ	CE	ን	n	ሆ	ho
ማ	ma	ያ	ya	ቤ	bE	ጴ	PE	ኝ	N	ሎ	lo
ሣ	`sa	ዳ	da	ቬ	vE	ጼ	SE	እ	I	ሖ	Ho
ራ	ra	ጃ	ja	ቴ	tE	ፄ	`SE	ክ	k	ሞ	mo
ሳ	sa	ጋ	ga	ቼ	cE	ፌ	fE	ኽ	`k	ሦ	`so
ሻ	xa	ጣ	Ta	ኄ	`hE	ፔ	pE	ው	w	ሮ	ro
ቃ	qa	ጫ	Ca	ኔ	nE	ሀ	h	ዕ	`I	ሶ	so
ባ	ba	ጳ	Pa	ኘ	NE	ለ	l	ዝ	z	ሾ	xo
ቫ	va	ጻ	Sa	ኤ	E	ሕ	H	ዥ	Z		
ታ	ta	ፃ	`Sa	ኬ	kE	ም	m	ይ	y		
ቻ	ca	ፋ	fa	ኼ	`kE	ሥ	`s	ድ	d		
ኃ	`ha	ፓ	pa	ዌ	wE	ር	r	ጅ	j		
ቆ	qo	ዖ	`o	ፎ	fo	ቷ	tWa	ጯ	CWa	ቋ	qWa
ቦ	bo	ዞ	zo	ፖ	po	ቿ	cWa	ፏ	Pwa	ኋ	SWa hWa
ቮ	vo	ዣ	Zo	ኗ	lWa	ኈ	hWe	ፏ	fWa	ኳ	kWa
ቶ	to	ዮ	yo	ኋ	HWa	ኗ	nWa	ፗ	pWa	ጓ	gWa
ቾ	co	ዶ	do	ሟ	mWa	ኟ	NWa	ቊ	qWu	ቈ	qWE
ኆ	`ho	ጆ	jo	ሧ	`sWa	ኰ	kWe	ኍ	hWu	ኌ	hWE
ኖ	no	ጎ	go	ሯ	rWa	ዟ	zWa	ኲ	kWu	ኴ	kWE
ኾ	No	ጦ	To	ሷ	sWa	ዧ	ZWa	ጕ	gWu	ጔ	gWE
ኦ	o	ጮ	Co	ኧ	xWa	ዷ	dWa	ቊ	qWi	ኸ	ea
ኮ	ko	ጰ	Po	ቈ	qWe	ጇ	jWa	ኍ	hWi		
ኾ	`ko	ጸ	So	ቧ	bWa	ጐ	gWe	ኩ	kWi		
ዎ	wo	ፆ	`So	ቭ	vWa	ጧ	TWa	ጕ	gWi		

C. Ge’ez-Numbers.



D. **Punctuation**



APPENDIX II: Lists of Sample Text with Their Frequency

('zewe`Ie', 2)
('eron', 1)
('ebase', 1)
('yt`eqebwo', 2)
('Hzb', 1)
('lekulomu', 1)
('kone', 2)
("bzu'han", 1)
('itab', 1)
('estesalma', 1)
('Igzine', 1)
('ma`hyewi', 2)
('merTuleke', 1)
('be`haTie', 1)
('`heb`ani', 1)
('Ski', 1)
('wequr', 1)
('`hebEhu', 3)
('iytgemer', 2)
('emlakomu', 1)
('`sereSa', 1)
('bzu`han', 3)
('nesHu', 1)
('hzb', 1)
('beemkurab', 1)
('kelIta', 1)
('msb`it', 1)
('meleke', 1)
('yngru', 1)
('Tbl', 2)
('bemnt', 1)
('kebetet', 2)
('`ebiye', 1)
('Ibrotu', 1)
('brhan', 2)
('yaIqob', 3)
('konke', 1)
('Kflo', 1)
('webelElit', 1)
('iyadengSa', 1)
('bluye', 2)
('H', 1)
('`helefe', 2)
('ySelyu', 2)
('yeHewr', 2)
('`alem', 4)
('wetelIkwo', 2)
('`helqu', 1)
('bekalIt', 2)
('zeleke', 1)
('nguse', 1)
('weImz', 4)
('fTret', 2)
('kulu', 8)
('fqerE', 3)
('taHte', 2)
('mel`Ite', 2)

('mebreqe', 1)
('egannt', 1)
('HSan', 1)
('ysIlo', 1)
(`Ifret', 2)
('ymeSI', 1)
('yuHens', 3)
('ra`Iyu', 2)
('nebere', 2)
('dengeSe', 1)
('tled', 2)
('zeegbo', 2)
('welde', 4)
('tf`sHte', 1)
('burket', 2)
('WsTe', 1)
(`alem', 1)
('webe', 2)
('lemelake', 1)
('leweld', 1)
('bequre', 2)
('leIrqt', 1)
('hayle', 1)
('e`hegur', 1)
('Sdqe', 1)
('nbl', 2)
('wereseyene', 3)
('kEfa', 1)
(`haTietene', 5)
('menfesawit', 3)
('kme', 1)
('krstos', 6)
('ynebr', 4)
('ymHeru', 2)
('wengEl', 2)
('ed`hnene', 3)
('may', 1)
('na`ebyeki', 2)
('hebo', 1)
('PETros', 1)
('eqabE', 1)
('Igztne', 2)
('ytwaHyu', 1)
('ze`ITan', 2)
('IgziI', 1)
('ed`hane', 2)
('temeherke', 1)
('ElsabET', 4)
('tIzazu', 2)
('beSelet', 1)
('sfaH', 2)
('ebuke', 1)
('bela`IlEke', 1)
('Im`Ser', 1)
('mntni', 3)
('tese`ete', 2)
('wsTe', 12)
('Imdr', 2)
('neSerwo', 2)
(`Snsa', 2)
('yebset', 3)
('nefsye', 2)
('Ysuqeni', 1)
('Imantu', 2)
('leequyaSike', 1)
('ka`Ibe', 1)
('menu', 2)
('w', 1)
('memkre', 2)
('Kbr', 5)
('Kemebo', 1)
('y`suqeni', 2)
('bEto', 2)
('bezeymeSI', 1)
('lebzu`han', 2)
('ski', 2)
('sbuH', 1)
('Imtekle', 2)
('emkurab', 1)
('wsTEta', 1)
('bewaIyke', 1)
('ybElo', 7)
('yhuda', 3)
('Zntu', 2)
('lb', 2)
('konet', 2)
('bemenfes', 1)
('Imerdaihuni', 1)
('eHzab', 1)
('Imu', 2)

('weladite', 14)
('itkl`omu', 1)
('zerIye', 2)
('Imz', 1)
('Brhan', 1)
('mudayu', 2)
('enkerwo', 2)
('lhiqan', 1)
('Edom', 1)
('sebI', 9)
('`selestu', 2)
('Lene', 1)
('yfEwso', 1)
('elbo', 2)
('ye`eTn', 2)
('weboe', 2)
('gEgay', 1)
('Ilenehebu', 1)
('m`Iraf', 1)
('Peralyu', 2)
('eziz', 1)

APPENDIX -III: Sample of stop word lists

enti	zku	Inbele	boto	sfn	nahu
entmu	newa	emTa	bomu	mnt	`adi
enetin	neyu	keme	botomu	Isfnt	SbahemEha
wItu	neya	eme	bon	mntat	zelfe
wIton	neyeke	sobe	boton	leliha	Iwe
wItomu	neyeki	Inze	bke	lelike	emEn
ylti	neye	Ime	bekmu	leliki	eman
ene	neyumu	ela	bki	lelye	InqWa`I
nhne	neyun	Igzio	bkn	lelihon	Isme
kiyaye	neyekmu	wey	bye	lelikmu	emTane
kiyake	neyekn	elE	bne	leline	ekonu
kiyaki	neyene	eh	elbo	zieye	Inte
keyahu	dibe	ey	elbotu	ziene	Ile
keyaha	mel`Ilte	`hedg	elbomu	Intiene	ew
kiyane	la`Ile	oho	elbati	Intieke	wemime
kiyakmu	tahte	zsku	elbon	EtE	Inze
kiyakn	methte	zktu	elbke	Ifo	eyat
kiyahomu	wsTe	Intakti	elbkmu	maIzE	eytE
kiyahon	wsaTE	Intku	elbki	zye	beIfo
lelihu	ma`Ikele	Ilktu	elbye	hye	baHtu
lelihomu	beynene	Ilku	elbne	kulenE	Inga
zntu	beInte	Imantu	Iwe	dhre	Indai
zati	Im	lotu	enga	yeman	`In`I
Ilontu	Imne	kulu	Inbi	Segam	Tqe
Ila	bEza	kula	Inbiye	lefE	nstit
Ilu	hyente	kulkmu	Inbi	welefE	HdaT
Ilemenu	`hebe	kulomu	Inbiye	yom	Hqe
menu	mengele	kulon	In	tmalm	`Smite
Ilantu	gizE	kulkn	Tqe	gE`sem	qdm
zktu	meTene	el	sey	yIzi	da`Imu

bhil	bebeynatihou	mel`Ite	kiyake	zntu	neyu
wetre	bebeynatihu	enti	kiyaki	zati	neya
zelfe	bebeynatiha	entmu	keyahu	Ilontu	neyeke
Isku	bebeynatihon	enetin	keyaha	Ila	neyeki
ne`a	bebeynatine	wItu	kiyane	Ilu	neye
heb	bebeynatikmu	wIton	kiyakmu	Ilemenu	neyumu
enbi	ebeynatikn	wItomu	kiyakn	menu	neyun
Intieye	kulu	yIti	kiyahomu	Ilantu	neyekmu
Ilieye	Iske	ene	kiyahon	zktu	
kalI	nke	nhne	lelihu	zku	
baHtit	Kiyaki kme	kiyaye	lelihomu	newa	

APPENDIX IV: SAMPLE OF EMPLOYED STOP WORD LISTS

Words	frequencies	Words	frequencies
ዝንቱ 'zntu'	**2**	ጥቀ 'Tqe'	**1**
አንቲ 'enti'	**20**	መንገለ 'mengele'	**1**
ከመ 'keme'	**17**	ኀበ '`hebe'	**7**
ወእቱ 'wItu'	**16**	ጊዜ'gizE'	**1**
አነ 'ene'	**1**	እንተ 'Inte'	**6**
በእንተ 'beInte'	**5**	ላዕለ 'la`Ile'	**2**
የማን 'yeman'	**1**	እለ 'Ile'	**8**
እስመ 'Isme'	**19**	እሉ 'Ilu'	**1**
ወስጠ 'wsTe', '	**1**	ኩሉ 'kulu'	**8**
እመ 'Ime'	**3**	ኪያኪ 'kiyaki'	**2**
ናሁ 'nahu'	**1**	ዲበ 'dibe'	**4**
አልቦሙ'elbomu'	**1**	እማንቱ 'Imantu'	**1**
ዚአየ 'zieye'	**2**	እንዘ 'Inze'	**9**
ሜኑ 'menu'	**3**	እንበለ 'Inbele'	**2**
እምኅ 'Imne'	**1**	ኩሉሙ'kulomu'	**1**
አመ 'eme'	**6**	አልብየ 'elbye'	**4**
ሎቱ 'lotu'	**2**	ድህረ 'dhre'	**1**
ይእቲ 'yIti'	**1**		

APPENDIX V: THE SAMPLE TEXT OF THE RESEARCH WITH ITS TRANSLATED VERSION

ወንጌል ዘሉቃስ ምዕራፍ 1 ክፍል በእንተ ዘየብሰት እዴሁ

ወእምዝ በካልእት ሰንበት በአምኩራብ ወመህሮሙ ወህሎ ህየ ብእሲ ዘየብሰት እዴሁ እንተ የማን ወይትዐቀብዎ ጸሐፍት ወፈሪሳውያን ነአመ ይፈወሶ በሰንበት ከመ ይርከቡ ምክንያተ በዘያስተዋድይዎ ወወእቱሰ የአምሮሙ ዘይሐልዩ ወይቤሎ ለወእቱ ብእሲ ዘየብሰት እዴሁ ተንስእ ወቁም ሜእከለ ወተንስአ ወቆመ ወነጸሮ ኩሎሙ በመዓት ይቤሎ ለወእቱ ብእሲ ስፋሕ እዴከ ወሰፍሓ ወሐይወት እዴሁ ወኮነት ከመ ከልእታ ወእመንቱሰ ዐብዱ ፈድፋደ ወተሣርሩ በበይናቲሆሙ ዘከመ ይሬስይዎ ለእግዚእ ኢየሱስ

ወእምዝ ኮነ በወእቱ መዋእል ዐርገ እግዚእ ኢየሱስ ወስተ ደብር ይጸሊ ወያሌሊ በጸለት ጎበ እግዚአብሔር እሉ እምንቱ ስምዖን ዘተሰምየ ጴጥሮስ ወእንድርያስ እኁሁ ወያእቆብ ወልደ እልፍዮስ ወስምኦን ዘይብልዎ ቀናኢ ወይሁዳ ዘያእቆብ ወይሁዳ አስቆርቃዊ ዘዐለዎ ወዘአግብአ ወወረደ ምስሌሆሙ ወቆመ በገዳም ውብዙኃነ ወ ሰብእ እምአርዳኢሁኒ ውብዙኃን ጥቀ እምሕዝብ ዘእምኩሉ ይሁዳ ወእምኢየሩሳሌም ወእምጳራልያስ ወጸራልዩ ወእምጢሮስ ወሲዶና እለ መጽኡ ይስምዕዎ ወይትፈወሱ እምደዌሆሙ ወእለሂ አጋንንት እኩያን የሐይው ወኩሎሙ አሕዘብ ይፈቅዱ ይግስስ እስመ ኃይል ይወፅእ እምኔሁ ወያሐይዎሙ ለኩሎሙ ቡርክት አንቲ እምአንስት ወቡሩክ ፍሬ ከርስኪ አ ሜርያም ድንግል ወላዲተ አምላክ ዘእንበለ ርኩስ ሰረቀ ለነ እምኔኪ ፀሓየ ጽድቅ ወአቅረበነ ታሕተ ክንፊሁ እስመ ወእቱ ፈጠረነ ሰአሊ ለነ ቅድስት ለኪ ለባሕቲትኪ አ እግዝትነ ወላዲተ አምላክ እመ ብርሃን አንቲ ናዐብየኪ በስብሐት ወበወዳሴ

wengEl zeluqas m`Iraf kfl beInte zeyebset IdEhu weImz bekalIt senbet beemkurab wemeheromu wehelo heye bIsi zeyebset IdEhu Inte yeman weyt`eqebwo SeHeft weferisawyan neIme yfEwso besenbet keme yrkebu mknyate bezeyastewadywo wewItuse yeemromu zeyHElyu weybElo lewItu bIsi zeyebset IdEhu tensI wequm maIkele wetense weqome weneSiro kulomu beme`at ybElo lewItu bIsi sfaH IdEke wesefHa weHeywet IdEhu wekonet keme kellta weImuntuse `ebdu fedfade wetemakeru bebeynatihomu zekeme yrEsywo leIgzIl iyesus weImz kone bewItu mewall `erge IgzIl iyesus wste debr ySeli weyalEli beSelet `hebe IgziebHEr Ilu Imntu sm`on zetesemye PETros weIndryas I`huhu weyaIqob welde Ilfyos wesmon zeyblwo qenai weyhuda zeyaIqob weyhuda esqortawi ze`elewo wezeegbo wewerede mslEhomu weqome begedam webzu`hanewe sebI Imerdaihuni webzu`han Tqe ImHzb zeImkulu yhuda weImiyerusalEm weImPeralyas wePeralyu weImTiros wesidona Ile meSu ysm`Iwo weytfewesu ImdewEhomu wellehi egannt Ikuyan yeHeyw wekulomu eHzab yfeqdu ygss Isme `hayl ywe`SI ImnEhu weyaHeywomu lekulomu burket enti Imenst weburuk frE kerski o maryam dngl weladite emlak zeInbele rkus sereqe lene ImnEki `SeHaye Sdq weeqrebene taHte knfihu Isme wItu feTerene seeli lene qdst leki lebaHtitki o Igztne weladite emlak Ime brhan enti na`ebyeki besbHet webewdasE

Declaration

I, the undersigned, declare that this thesis is my original work, has not been presented for any other universities. All sources of materials and advices used for the thesis have been fully acknowledged.

Declared by:

Name: Abebe Belay

Signature: _____

Date: July 9/2010

Confirmed by Advisor:

Name: _____

Signature: _____

Date: _____