Addis Ababa University
College of Natural and Computational Sciences
School of Information Science

Morpheme-Based Bi-directional Ge'ez -Amharic Machine Translation

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Science

By:
Tadesse Kassa
Addis Ababa, Ethiopia

Advisor: Million Meshesha (PhD)
Addis Ababa, Ethiopia

October, 2018

Dedication

I dedicate this work to my mother Tiruwork Abdi, my wife Wegayew Kelemu, to my children Tsgazab, Egezeharya and Zeraburuk Tadesse.

I would like also to dedicate to the Ethiopian Orthodox Tewahedo Church and her scholars for their dedication and faithfulness to preserve and hand down the language along with all resources written in it to this generation!

እግዚኦ መኑ ይኅድር ውስተ ጽላሎትከ ፤ ወመኑ ያጸልል ውስተ ደብረ መቅድስከ ።
ዘየሐውር በንጹሕ ወይገብር ጽድቀ ፤ ወዘይነብብ ጽድቀ በልቡ ።
ወዘኢጓሕለወ በልሳኑ ወዘኢገብረ እኩየ ዲበ ቢጹ ፤ ወዘኢያጽአለ አዝማዲሁ ።
ወዘምኑን በቅድሜሁ እኩይ ወዘያከብሮሙ ለፈራህያነ እግዚአብሔር ፤ ዘይምሕል ለቢጹ ወኢይሔሱ ።
ወዘኢለቅሐ ወርቆ በርዴ ወዘኢነሥአ ሕልያነ በላዕለ ንጹሕ ፤ ዘይገብር ከመዝ ኢይትሀወክ ለዓለም ።
መዝሙር ፲፭፤ ፩-፭

Addis Ababa University
College of Natural and Computational Science
School of Information Science

Morpheme-Based
Bi-directional Ge'ez -Amharic
Machine Translation

Signature for Approval

Name	Signature	Date
Million Meshesha (PhD), Advisor	_____	_____
Wondwossen Mulugeta (PhD), Examiner	_____	_____
Solomon Tefera (PhD), Examiner	_____	_____

Acknowledgment

Above all I would like to thank the almighty God, who gave me the opportunity and strength to achieve whatever I have achieved so far. I would like to express my gratitude to all the people who supported and accompanied me during the progress of this work. Special thanks go to Bahitawi Hailu in every aspects of my life. Bahitawi this is your result.

First, I would like to express my deep-felt gratitude to my advisor, **Dr. Million Meshesha**, whose excellent and enduring support shaped this work considerably and made the process of creating this work an invaluable learning experience.

Second, I would like to thank Addis Ababa University, College of Natural and Computational Science School of Information Science sponsoring me in Msc in Information Science

Memehire Ephrem Taglo for helping to understand the grammar of Geez. W/o Mulunesh and Firewote in writing Kidan and Liton. W/t Berekti for writing wedasemaryam, yewedesewa melahekete, Anketse-Berhan. Ato Tsegaye Andargie, helping in crawling the dataset from websit. Ato Amanuel Lemma in supporting mosses installation and shell scripting. Ato Wondimageghu Tsegaye in supporting python programming and shell programming over all. Ato Michael Melese, Ato Berihun Addase and Ato Habete Abera in showing how to use morfessor and providing comment regarding scripting code and research.

My sincere thanks go to my spiritual fathers, Bahitawi Hailu and Aba Aserate, Aba Mezemure, and Aba Hayle Michael, your praise, love of education, precious advices and motivation always push me forward. Your spirit will be with me forever.

Finally, to finish this program, the share of my wife Wegayew Kelemu is unlimited by taking care of our children and me. My little babies Tsegazab, Egezharya and Zehrahbruk this is your prayers result, I want to thank you from the bottom of my heart. All my family especial my mother Tiruwork Abdi and all my brothers and sisters I thank you.

Abstract

This study aims to explore the effect of morpheme level translation unit for bi-directional Ge'ez-Amharic machine translation. Using word as a translation unit is a problem in statistical machine translation while conducting translation between two morphologically rich languages such as Ge'ez and Amharic. At word level, data scarcity and unavailability of well prepared corpus is a challenge for under resourced language. And, at word level, it is difficult to manage many forms of a single word, not specific and lacks consistency. At morpheme level sub parts of words are specific, easy to manage specific parts and has consistency our many words of the same class.

To conduct the experiment, parallel corpus was collected from online sources. Such Online sources include Old Testament of Holy bible and anaphora (or Kidase). The corpus include manually prepared bitext from Wedase Maryam, Anketse Berhane, yewedesewa melahekete, Kidan and Liton. To make the corpus suitable for the system, different preprocessing tasks such as tokenization, cleaning and normalization have been done. The data set contains a total of 13,833 simple and complex sentences, out of which 90% and 10% are used for training and testing, respectively. To build a language model for both languages we used 12, 450 parallel sentences. For both statistical and rule-based approachs we used Mosses for translation process, MGIZA++ for alignment of word and morpheme, morfessor and rules were used for morphological segmentation and IRSTLM for language modeling. After preparing and designing the prototype and the corpus, different experiments were conducted.

Experimental results showed a better performance of **15.14%** and **16.15%** BLEU scores using morpheme-based from Geez to Amharic and from Amharic to Geez translation, respectively. As compared to word level translation there is on the average **6.77%** and **7.73%** improvement from Geez-Amharic and Amharic-Ge'ez respectively. This result further shows that morpheme-level translation performs better than word-level translation. As a result, using morpheme as a translation unit we conducted further experiment using unsupervised and rule-based morpheme segmentation approaches. Accordingly, the performance of rule-based morphological segmentation is better than unsupervised with an average BLEU score of **0.6%** and **1.27%** for Ge'ez to Amharic and Amharic to Ge'ez respectively.

Alignments of Amharic and Ge'ez text have shown correspondence, such as one-one, one-to-many, many-one and many-many alignment. In this study, many-to-many alignment is the major challenge. So further research is needed to handle many-to-many, word order and morphology of the two languages.

Key word: SMT; morpheme level alignment; morfessor; Amharic; Geez

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ALPAC - Automatic Language Processing Advisory Committee

BLEU - BiLingual Evaluation Understudy

EOTC - Ethiopian Orthodox Tewahedo Church

FDRE – Federal Democratic Republic of Ethiopia

FVSO – Verb- Subject-Object

IRSTLM –Institute of Research

LM Language Model

MT - Machine translation

RBMT - Rule Based Machine Translation

SMT - Statistical Machine Translation

SOV - Subject-Object-Verb

SVO - Subject-Verb-Object

Chapter One

Introduction

1.1. Background

Machine translation (MT) is a technology that enables the use of computers to automate the process of translating from one language to another. Translation, in its full generality, is a difficult, fascinating, and intensely human endeavor, as rich as any other area of human creativity [1].

The translation of natural languages by machine, first dreamt of in the seventeenth century, has become a reality in the early [2]. The history of machine translation is traced from the pioneers and early systems of the 1950s and 1960s, the impact of the Automatic Language Processing Advisory Committee (ALPAC) report in the mid-1960s, the revival in the 1970s, commercial and operational systems of the 1980s, and research during the 1980s [2] [3].

Machine Translation has different advantages [4]. The first one is currently time is a crucial factor, machine translation can save the day. Individuals are not expected to spend hours poring over dictionaries to translate the words. Instead, a software can translate the content quickly and provide quality output to the user immediately. The speed of translation by machine is exponentially faster than that of humans. On an average, human can translate around 2,000 words a day [2]. Multiple translators can be assigned to a given project to increase translation output, but it is not-comparable with the speed of machine translation. Machine translation can generate thousands of words with in a minute [5].

The second advantage of machine translation is that it is comparatively cheap. Initially, it might look like an unnecessary investment but in the long run it is a very small cost considering the return on investment it provides. This is because the use of the expertise of a professional translator, he/she will charge on a per page basis which is going to be extremely costly while this will be cheap in the case of MT. Thirdly, confidentiality is another advantage that makes machine translation favorable. Giving sensitive data to a translator might be risky while with machine translation information is protected. The role humans in postediting of the machine translation output is unreplaceable.

Finally, a machine translator usually translates text with which it is trained. The same is true for professional, so there is no such major concern while a professional translator specializes in one field.

MT approaches includes rule based, corpus based and hybrid [1]. Rule-Based Machine Translation, also known as Knowledge-Based MT, is a general term that describes machine translation systems based on linguistic information about source and target languages. Corpus-based MT Approach, also referred as data driven machine translation, is an alternative approach for machine translation to overcome the problem of knowledge acquisition problem of rule-based machine translation. Corpus Based Machine Translation uses, a bilingual parallel corpus to obtain knowledge for new incoming translation. Statistical techniques are applied to create models whose parameters are derived from the analysis of bilingual text corpora. Example-based machine translation (EBMT) is one of the example of corpus-based machine transaltion, characterized by its use of bilingual dictionary with parallel texts as its main knowledge, in which translation by correlation is the main idea. By taking the advantage of both corpus based and rule-based translation methodologies, hybrid MT approach is developed, which has a better efficiency in MT systems [1]. For under-resourced languages such as Ge'ez and Amharic with limited or no linguistics resources, statistical approach is recommended [1].

1.2. Morpheme, word, phrase and sentence

Morpheme is the minimal meaningful unit in a word. The concept of word and morpheme are different, and a morpheme may or may not stand alone. One or several morphemes compose a word. As stated in [6] [7], there are four types of morphemes:

- ✓ Free morphemes: can appear with other lexemes such as town and dog; for example, town hall or dog house or they can stand alone, i.e. "free". They are meaningful when used alone.
- ✓ Bound morphemes: appear only together with other morphemes to form a lexeme. Bound morphemes in general tend to be prefixes (un-, dis-), suffixes (-ing, -ed, -es), infix (**bleep** in fivebleepmile) and circumfix (**em- -en** in embiggen, embolden and embrighten).
- ✓ Derivational morphemes can be added to a word to create (derive) another word: the addition of "-ness" to "happy" for example, gives "happiness". They carry semantic information. Word class will change.

- ✓ Inflectional morphemes modify a word's tense, number, aspect, and so on, without deriving a new word or a word in a new grammatical category (as in the "dog" morpheme if written with the plural marker morpheme "-s" becomes "dogs"). They carry grammatical information.

Word is a single distinct meaningful element of speech (phonologically) or writing (orthographically), used with others (or sometimes alone) to form a sentence and typically shown with a space on either side when written or printed [8].

Phrase is a small group of words standing together as a conceptual unit, typically forming a component of a clause. Phrase is a group of words that express a concept and is used as a unit within a sentence [9]. A Phrase is separate by punctuation mark [10].

A sentence is a group of words that are put together to mean something. A sentence is the basic unit of language which expresses a complete thought. Sentence is a set of words that is complete, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clauses [8]. Morphemes, word, phrase and sentence are among the different translation unit [10] [11].

Machine translation has its own challenges even if it is active current research area [1]. Several well-known problems are, fundamentally, problems of scarce bitext. The first challenge in MT is translation of low-resource language pairs. The most straightforward example of scarce bitext covers most of the world's language pairs. The second one is translation across domains. Translation systems are not robust across different types of data, performing poorly on text whose underlying properties differ from those of the system's training data. The third challenge is translation into morphologically rich languages. Finally, translation of speech**.** Much of human communication is oral. Even ignoring speech recognition errors, the substance and quality of oral communication differs greatly from that found in most bitext [12].

According to Okpor [13], an important new development for MT in the last decade has been the rapid progress that has been made towards developing speech to speech machine translation. Once thought simply too difficult, improved speech-analysis technology has been coupled with innovative design to produce many working systems, albeit still experimental, which suggest that this may be the new growth area for MT research

1.3. Ge’ez and Amharic Languages

Ethiopian is one of the country in Africa that have its own Fidel or Letter and Numbers. This scripting method is the identity of the country not only in African but also in the international Arena. The word Ge’ez means first in the Alphabet, first in reading style and first in Zema (Gloss) teaching of the Ethiopian orthodox Tewahedo Church. Ge’ez (**ግዕዝ**) is an ancient South Semitic language and is a member of the Ethiopian Semitic group. The language originated in southern regions of Eritrea and the northern region of Ethiopia in the Horn of Africa. It later became the official language of the Kingdom of Aksum and Ethiopian imperial court [14].

Today, Ge'ez remains only as the main language used in the liturgy of the Ethiopian Orthodox Tewahedo Church, the Eritrean Orthodox Tewahedo Church, the Ethiopian Catholic Church, the Eritrean Catholic Church, and the Beta Israel Jewish community [15].

As presented in Appendix **XII**, these days, Ge’ez is being researched and taught in Ethiopia, European[1] and United States of America Universities[2]. The Holy Trinity Spiritual College in Ethiopia is teaching Ge’ez language at Diploma Level. It is also being taught by Ethiopian Orthodox Tewahedo Church schools called **አብነት ትምህርት**. Abune Gorgorios Academy is the only academy that teaches Ge’ez as a subject from Kinder Garden to Preparatory in a well-organized manner. On the other hand, language teaching center and Online Ge’ez schools also working on Ge’ez language. The one that is the source of Ge’ez language is the Ethiopian Orthodox Tewahedo Church that is teaching Ge’ez in traditional schools that exists inside and outside the country.

From the above explanation Ge’ez language is becoming well-known by local and international community; as a result of which there is an increase in the number of Ge’ez language speakers from time to time. Hence, an attempt is made in this study to design a bi-directional machine translation from Ge’ez to Amharic and vice versa.

[1] http://www.geeskaafrika.com, https://www.borkena.com
[2] https://www.washington.edu, https://myplan.uw.edu)

In Ethiopia, Amharic (the main lingua franca of modern Ethiopia) and other local languages, such as Tigrinya and Tigre are closely related to Ge'ez, with at least four different configurations proposed. Ge'ez is the root language for Ethiopian Semitic languages such as Amharic, Tigrinya and Tigre.

However, some linguists do not believe that Ge'ez constitutes the common ancestor of modern Ethiopian languages, but that Ge'ez became a separate language early on from some hypothetical, completely unattested language and can thus be an extinct sister language of Tigre and Tigrinya [16]. The foremost Ethiopian experts such as Amsalu Aklilu point to the vast proportion of inherited nouns that are unchanged, and even spelled identically in both Ge'ez and Amharic [17].

Amharic is the official working language of the Federal Democratic Republic of Ethiopia and is estimated to be spoken by well over 20 million people as a first or second language [18]. Amharic is the second most spoken Semitic language in the world (after Arabic). Today it is probably the second largest language in Ethiopia (after Oromo, a Cushitic language) and possibly one of the five largest languages on the African continent. Following the Constitution drafted in 1993, Ethiopia is divided into nine independent regions, each with its own regional working language. Amharic is the working language of different regional states including Amhara regional state, Addis Ababa and Southern Nations, Nationalities and peoples.

Ge'ez script is an alpha syllabary script also called "**Abugida**", in which a character represents a consonant and a vowel combination. This is different form alphabetic script where a character represents one sound either a consonant or a vowel. The alphabet of Amharic script are unique scripts acquired from Ge'ez and use an alpha syllabary writing system where the consonant and vowel are combined to form a single symbol. Thus, once a person knows all the alphabets, he/she can easily read and write both Ge'ez and Amharic.

Script in Ge'ez and Amharic includes 26 and 34 basic alphabets (called 'Fidel'), each having seven forms created by fusing a consonant for an alphabet, yielding 182 and 238 distinct characters respectively and other additional forms are derived from the basic alphabets like ቈ ቑ ቒ ቓ ቔ from ቀ, ኰ ኵ ኲ ኳ ኴ from ከ, ጐ ጒ ጓ ጔ ጕ from ጎ and ጘ ጙ ጚ ጛ ጜ from ገ. Modern Ge'ez and Amharic are written from left to right. Before the 4th century it was written from right to left [16].

The syntactic structure is formed by combining different word classes in sequence [9]. The usual word order of Amharic is Subject-Object-Verb (SOV) whereas Ge'ez follows Subject-Verb-Object (SVO) word order for declarative sentences. In Ge'ez, other orders are possible like VSO, and SOV.

For example, the Amharic equivalent for the Ge'ez sentences with SVO **"ውእቱ መጸአ እምቤቱ"** [weetu metsa embetu], VSO **"መጸአ ውእቱ እምቤቱ"** [metsa weetu embetu] and SOV **"ውእቱ እምቤቱ መጸአ"** [weetu embetu metsa] is **"እሱ ከቤቱ መጣ"** [esu kbetu meta] meaning "He came from his home" where **"እሱ** [esu]" is the subject of the Amharic sentence equivalent to **"ውእቱ** [weetu]" in Ge'ez, **"ከቤቱ** [kbetu]" is the object of the Amharic sentence equivalent to **"እምቤቱ** [embetu]" in Ge'ez, and **"መጣ [**meta]" is the verb of the Amharic sentence which is equivalent to **"መጸአ** [metsa]". But usually pronouns are not omitted both in Ge'ez and Amharic sentences rather it become part of the verb when they used as a subject **"መጸአ እምቤቱ** [metsa embetu]" equivalent to **"ከቤቱ መጣ** [kbetu meta]".

Both Amharic and Ge'ez have a complex morphology. The word formation for instance, involves different formations including prefixation, infixation, suffixation, and reduplication. Most function words in Amharic and Ge'ez such as Conjunction, Preposition, Article, Pronominal affixes, Negation markers are bound morphemes which are attached to the content words, resulting in complex words composed of several morphemes [19]. Morphologically complex languages also tend to display a rich system of agreements between the syntactic part of a sentence like nouns, verbs, person, number, gender, fine and place. This increases the complexity of word generation.

In addition, the baseline phrase-based translation approach has limited success on translating between languages with very different syntax and morphology, especially when the translation direction is from a language with fixed word structure to a highly inflected language [1]. In addition, the rich morphology of a highly inflected language permits a flexible word order, thus making difficult to model long range word order differences between languages. When both the source and the target languages are morphologically rich, difficulty in translation also gets complex [20]. There are two main points to improve on: morphological translation equivalence and long range reordering [20]. Translating the correct surface form realization of a word is dependent not only on the source word-form, but it also depends on additional morpho-syntactic information.

1.4. Statement of the Problem

Ge'ez is an ancient language and many manuscripts are already archived by Ethiopian Orthodox Church as well as by the National Archival agency. Ge'ez had been known as being used in Ethiopia since the 4th century and as a spoken language close to a thousand years and had been serving as official written language practically up to the end of 19th century [14].

Since currently there are a lot of historical, cultural and religious documents available in Geez language, there is a need to translate the manuscripts to Amharic and other Ethiopian Languages to make the decoded knowledge accessible to every especially Amharic users. On the other hand, as discussed earlier, Ge'ez as a language being researched and taught in different Universities around the word in terms of accessing the decoded knowledge. Indirectly, Ge'ez language speakers are being created therefore, there is also a need to translate Amharic documents to Geez language.

Some attempts are done by EOTC (Ethiopian Orthodox Tewahedo Church) and individuals to translate manually some of the religious manuscripts, law and philosophical works. The problem observed in manual translation are time taking, resource intensive, and linguistic knowledge of the language is mandatory. Machine translation, although it has its own challenges, can improve performance and reduce cost. Though there are advancement in applying MT for different languages pairs, it is still in its infant stage for our local languages.

These days Geez language is on revival; different Universities in the country and internationally start offering Geez as a course and a subject. This also necessitates transaltion of documents from Amharic to Ge'ez. As a matter of fact, there are few researches made on MT in Ethiopian languages. Most of these works attempts to pair local language with English, such as Amharic [21] , Afaan Oromoo [22] [19] [23], Tigrigna [24] [25], and Ge'ez [15].

However, Dawit [15], conducted an experiment on Ge'ez to Amharic language pair by using statistical MT approach. As noted by the researcher, word level translation process is challenged by many forms of a single word, due to morphological richness of the two languages where a single word in any of the two languages composed of many sub-words or morphemes.

Also the same affixes (prefixes and suffixes) exists in different words, which is not specific, unmanageable and inconsistent at word level. Another challenge is the unavailability of well-prepared parallel corpus for the machine translation task.

Since for morphologically rich languages it is not possible to cover all the words that exists in the language for translation, there is a need to experiment morpheme based translation.

At morpheme level, morphemes are specific, easy to manage and consistent as well as easy to overcome the data scarcity of the languages [4]. When translating across these pair of languages, morphological changes result in large numbers of out-of-vocabulary (OOV) terms between training and test sets leading to reduced BLEU scores in evaluation [26]. It is therefore, the main aim of this study to undertake morpheme-based bi-directional Ge'ez-Amharic automatic machine translation.

To this end, this study attempted to answer the following research questions:

- ✓ What are the suitable approach for morpheme-based corpus preparation?
- ✓ To what extent does morpheme-based translation improves the performance of the translation result?

1.5. Objective of the study

1.5.1. General Objective

The general objective of this research is to design morpheme-based bi-directional machine translation for Ge'ez-Amharic textual documents.

1.5.2. Specific Objectives

To achieve the general objective of the research, the following specific objectives are formulated:

- ✓ To review Literature to identify surface approaches and technologies for statistical machine translation and rule-based.
- ✓ To prepare data set for experimentation.
- ✓ To identify the syntactic relationship between Ge'ez and Amharic languages.
- ✓ To design an optimal language and transaltion model.
- ✓ To evaluate the performance of the prototype.

1.6. Scope and limitation of the Study

Machine translation has different approaches such as, example-based approach, rule-based approach, statistical approach and hybrid approach. In this study, statistical and rule-based machine translation approaches are compared. Statistical approach is economically wise since it doesn't need linguist professionals but if it is morpheme based, it requires this knowledge.

On the other hand rule–based approach needs linguistic knowledge of both languages. The translation process is done by using parallel corpus of paired language. In this study we used free morpheme, bound morpheme (prefix, suffix and circumfix) morpheme types.

Bi-directional Ge'ez-Amharic, machine translation is designed to translate a sentence written in Ge'ez text into Amharic text and vice versa. The source of the data set includes Old Testament Holy Bible, Wedase Maryam, Kidase, Kidan, Liton and which include Ge'ez and Amharic version and simple sentences. These sources are selected because they are available, and they are parallel corpus which is suitable for SMT.

Because of unavailability of standardized corpus (corpus ready for MT research purpose) and balanced corpus (in terms of discipline) the data set prepared in this study focus on sources that are parallel textual data, because of which most of the data we used for training and testing are from religious documents.

1.7. Significance of the study

The beneficiaries of this research include the Society, translators and scholars. The society that able to understand Amharic benefited in getting resources that are written in Ge'ez such as history, philosophy, laws, tradition, and religion and so on. Especially the history of Ethiopia is almost being written in Ge'ez understanding this is not only essential for Ethiopian but also the rest of the world. It is also vital for us since in one or another way different document are translated from other languages such as Arabic, Greek. For translator it is also helpful in a way that to produce draft translation for post editing. The rate of machine translation is exponentially faster than that of human translation [10]. The main significance of this research work is the following; the first importance is reaching under resourced languages; by translating the different valuable publications; for example, from Ge'ez to Amharic it is possible to address information need of Amharic language speakers. The second importance is it solves language barriers between individuals to read and understand different publications.

The third importance is it helps for designing cross-language information retrieval to translate the documents the users are searching for and/or the query pose by the users. It also have academic significance in motivating researcher while conducting MT between local languages looking morpheme is another option.

1.8. Methodology of the study

Research methodology is a way to systematically plan for solving the research problem [27]. It may be understood as a science of studying how research is done scientifically. The advantage of knowing the methodology of the study before doing the experiment is to reason out what, how and why the methods or the techniques are selected for the experiment to know the risks for conducting the research in detail.

1.8.1. Research Design

To conduct the research, we followed experimental research design. To explore morphemes and words based on SMT and rule-based approach, different experiments were done. Experimental research investigates the possible cause-and-effect relationship by manipulating independent variables to influence the dependent variable(s) in the experimental group, and by controlling the other relevant variables, and measuring the effects of the manipulation by some statistical means [28]. Steps in Experimental Research include the following [28], devising alternative hypotheses/research questions, designing crucial experiments with alternative possible outcomes, each of which exclude one or more possible hypotheses and finally conducting the experiment, get a clean result and measure the performance of bi-directional Geez to Amharic MT.

1.8.2. Data Collection and Preparation

The data set, was collected from Old Testament Holy bible from sources https://www.ethiopicbible.com ,https://www.stepbible.org and http:// www. tau.ac.il/ ~hacohen/ Biblia .html and simple sentences adapted from [29], to perform the experiments. The reason to select these sources of data for corpus preparation is, because, it is easily accessible from the web and they are parallel corpus which is suitable for SMT and rule-based approach. Manually prepared data set like Wedase Maryam, Anketse Berhan, yewedesewa melahekete, Kidan and Liton were written manually by secretary with no fee. Anaphora's or Kidase, were collected from http://ethiopianorthodox.org. We also prepare suffixes and prefixes with the help of professional.

A total of 14, 412 parallel sentences were collected out of which 579 removed being repted in both language, through cleaning. Size of the corpus for the experiment is 13,833, prepared from the above-mentioned source of corpus. The reason why we select more corpus from Old Testament Holy bible is because of the availability of large amount of parallel textual corpus with more coverage of the domain. Tokenization and normalization are used as preprocessing activities. Tokenization is a task of separating out words from running text. Example I'am, need to separate into two words I and am. Normalization dealing with nonstandard words. Non-standard words include number, acronyms, abbreviations, and so on. For example, "March 31" needs to be pronounced "March thirty-first", not "March three one"; "$ 1 billion" needs to be pronounced one billion dollars, with the word dollars appearing after the word billion.

1.8.3. Implementation Tools

The basic tool used for accomplishing the machine translation task is Moses; free available open source software which is used for statistical machine translation and integrates different toolkits, which are used for translation purpose. These toolkits include IRSTLM for language model, Moses Decoder for translation and MGIZA++ for word and morpheme alignment. Python programming language is used as a tool for preprocessing and rule-based segmentation in Ubuntu Environment.

Since the purpose of the study is designing morpheme-based Geez-Amahric MT, we used two approaches in morphological segmentation. The first one is unsupervised morphological segmentation using morfessor. Morfessor is a family of probabilistic machine learning methods for finding the morphological segmentation from raw text data [31].

The other is rule-based morphological segmentation. For organizing rules we use Python programming language. We used morfessor and python due to, we are familiar with them, and easy to use in text processing researchers.

1.8.4. Evaluation Procedure

Machine translation systems are evaluated by using either human or automatic evaluation method. Since human evaluation method is time consuming and inefficient with respect to automatic evaluation method, we used BLEU score metrics to evaluate the performance of the system.
BiLingual Evaluation Understudy (BLEU) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another [32]. Quality is the correspondence between a machine's translation output and that of a human translated output.

The basic idea behind BLEU is, if the machine translation output closer to human translation output it is considered as better translation [32]. BLEU was one of the metrics to achieve a high correlation with reference translation and remains one of the most popular automated and inexpensive metrics used in different researches for evaluation purpose.

1.9. Thesis organization

This thesis is organized in to six chapters, the first chapter discuss about introduction, Ge'ez and Amharic language, statement of the problem, objective of the study, scope and limitation of the study, methodology followed including research design, data collection and preparation, Implementation tools and MT Evaluation procedure.

The second chapter presents literature review which focus on approach of machine translation, alignment and the effects of alignment on statistical machine translation, and different tools used for corpus alignment and related works related with this study.

The third chapter deals with an over view of Ge'ez language and its relationship with Amharic language and discussion of relationship between Amharic and Ge'ez Language.

Chapter four discuss about designing processes of the prototype including, corpus preparation, types of corpus used for the study, corpus alignment, and briefly discuss about the proto type of the system.

Chapter five deals with experimentation of the study which include different experiments and the results of the experiments with interpretation of findings.

The last, chapter six deals with conclusion of the findings and recommendations for further works.

Chapter Two

Literature Review

2.1. Overview of Machine Translation

The history of machine translation is traced from the pioneers and early systems of the 1950s and 1960s, the impact of the Automatic Language Processing Advisory Committee (ALPAC) report in the mid-1960s, the revival in the 1970s, the appearance of commercial and operational systems in the 1980s, research during the 1980s, new developments in research in the 1990s, and the growing use of systems in the past decade [3] [2]. These resulted in the birth of modern Machine translation.

Machine translation (MT), can be defined as translation of information from one natural language source language to another language target language using computerized systems; automatic or semi-automatic [33]. It is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.

Due to the advent of Computer and the internet the world is becoming together to one [13]. Thus, the knowledge, culture, tradition, history, religious, philosophy documents of one country language can be translated to another language and the rest of the world through Machine translation. To create a paperless working environment translation plays a great role and to make accessible the document of one language in another language. Sharing of Knowledge is also possible besides facilitating easy communication. No more being language barrier for Communications in any way.

2.2. Approaches of Machine Translation

MT systems can be classified according to their core methodology in to two main paradigms; the rule-based approach and the corpus-based approach [13]. In the rule-based approach, human experts specify a set of rules to describe the translation process, so that an enormous amount of input from human experts is required. On the other hand, under the corpus-based approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. Integration of both rule-based and Corpus based MT systems results in the Hybrid Machine Translation Approach.

There are two process of translations that are uni-directional and bi-directional process [30]. Uni-directional works only in one direction, which is first the system (language model and translation model) train by using the data set in one direction from source to target language, and the translation process also done in one direction only from source to target language. In bi-directional, the translation process is done in both direction from source language to target language and form target language to source language [23].

2.2.1. Rule-Based Machine Translation (RBMT) Approach

Rule-Based Machine Translation (RBMT), also known as Knowledge-Based Machine Translation or Classical Approach of MT, is a general term that denotes machine translation systems based on linguistic information about source and target languages basically retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. Having input sentences (in some source language), an RBMT system generates them to output sentences (in some target language) based on morphological, syntactic, and semantic analysis of both the source and the target languages involved in a concrete translation task [1] [4].

The basic principles of RBMT methodologies is to apply a set of linguistic rules in three different phases [1]: analysis, transfer and generation. Therefore, a rule-based system requires: syntax analysis, semantic analysis, syntax generation and semantic generation. The main approach of RBMT systems is based on linking the structure of the given input sentence with the structure of the demanded output sentence, necessarily preserving their unique meaning. Speaking in general terms, RBMT generates the target text given a source text following the steps shown in figure 2-1 below.

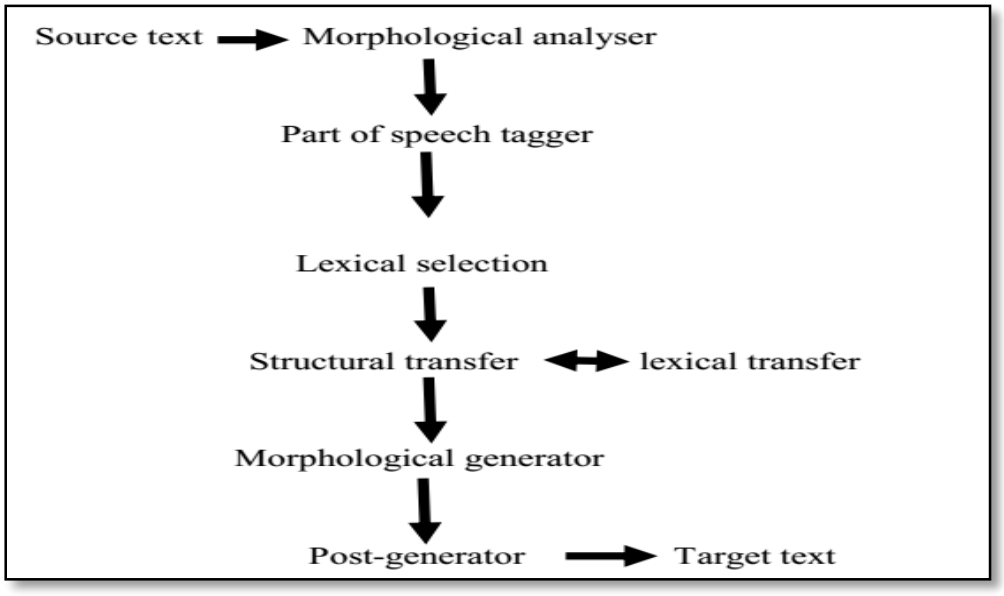


Figure 2-1 Architecture of RBMT Approaches

There are three different approaches under the rule-based machine translation approach [1], such as Direct, Interlingua and Transfer-Based Machine Translation approaches. They differ in the depth of analysis of the source language and the extent to which they attempt to reach a language-independent representation of meaning or intent between the source and target languages, though they all belong to the RBMT.

2.2.1.1. Direct Machine Translation

Direct Machine Translation Approach is the oldest and less popular approach. Direct translation is made at the word level. Machine translation systems that use this approach can translate a source language (SL) directly to target language (TL). Words of the SL are translated without passing through an additional/intermediary representation. The analysis of SL texts is oriented to only one TL. Direct translation systems are basically bilingual and uni-directional. Direct machine translation (DMT) approach needs only a little syntactic and semantic analysis. SL analysis is oriented specifically to the production of representations appropriate for one TL. DMT is a word-by-word translation approach with some simple grammatical adjustments. As shown in figure 2-2 below major tasks in direct machine translation include the following: Shallowest morphological analysis, Lexical transfer using bilingual dictionary, Local reordering and Morphological transfer [1] [4] [13].

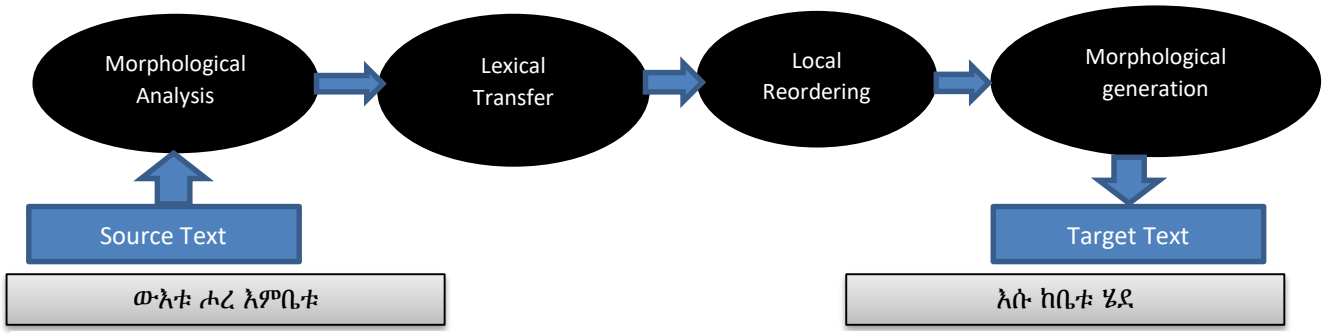


Figure 2-2 Major tasks in Direct Machine Translation approach

2.2.1.2. Interlingua Machine Translation

The failure of the first-generation systems led to the development of more sophisticated linguistic models for translation. There was increasing support for the analysis of source language texts into intermediate representation. A representation of its "meaning" in some respect which could form the basis of generation of the target text. Interlingua machine translation is one instance of rule-based machine-translation approaches.

In this approach, the source language, i.e. the text to be translated, is transformed into an Interlingua language, i.e. a "language neutral" representation that is independent of any language. The target language is then generated out of the Interlingua [1] [34].

2.2.1.3. Transfer-based Machine Translation

Transfer-based approach uses an intermediate representation that captures the structure of the original text to generate the correct translation. In transfer-based approach first the input text is parsed and then apply rules to transform the source language parse into a target language parse. The process of transfer-based translation involves: analysis, transfer and generation. Transfer bridges the gap between the output of the source-language parser and the input to the target language generator. Transfer based need rules for: syntactic transfer, Semantic transfer, and lexical transfer [35] [1].

Syntactic transfer rules will tell us how to modify the source parse tree to resemble the target parse tree. Semantic transfer using semantic role labeling. Lexical transfer rules based on a bilingual dictionary. The dictionary can be used to deal with lexical ambiguity

2.2.2. Corpus-based Machine Translation Approach

Rule-based approaches have been the dominant paradigm in developing MT systems. Such approaches, however, suffer from difficulties in knowledge acquisition to meet the wide variety and time-changing characteristics of the real text. To attack this problem, some statistical translation models and supporting tools had been developed in the last few years.

However, a simple statistical model often results in a large parameter space and thus requires a large training corpus. Therefore, it is required to introduce language models that take advantages of well-justified linguistic knowledge to make stochastic MT systems practical [36].

Corpus based machine translation, also referred as data driven machine translation, is an alternative approach for machine translation to overcome the problem of knowledge acquisition problem of rule-based machine translation. Corpus Based Machine Translation (CBMT) uses bilingual parallel corpus to obtain knowledge for new incoming translation. This approach uses a large amount of raw data in the form of parallel corpora. This raw data contains text and their translations.

These corpora are used for acquiring translation knowledge. Corpus based approach is further classified into the following two sub approaches: Statistical Machine Translation and Example-based Machine Translation Approach [13].

Statistical machine transaltion focus on the result, not the process. The correspondence between the words in the source and the target strings is described by alignments that assign target word positions to each source word position. The probability that a certain target language word will occur in the target string is assumed to depend basically only on the source words aligned with it [37].

2.2.2.1. Example-based Machine Translation Approach

The essence of EBMT, called “machine translation by example guided inference, or machine translation by the analogy principle” [38], is succinctly captured much-quoted statement:

> Man does not translate a simple sentence by doing deep linguistic analysis, rather, Man does translation, first, by properly decomposing an input sentence into certain fragmental phrases ..., then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence.

The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference [38].

In EBMT a set of phrases in the Source language and their corresponding translations in the Target language are given in example database. The MT system uses these examples to translate new similar SL phrases into the TL. The basic premise is that, if a previously translated phrase occurs again, the same translation is likely to be correct again.

The three main components of EBMT:

- Matching the SL input against the example database
- Alignment/Adaptation – Selecting the corresponding fragments in the TL.
- Recombination (target sentence generation or synthesis) Recombining the TL fragments to form a correct text.

Example:

- የመጽሃፉ ዋጋ ከ500 ብር በላይ ነው -> The price of the book is more than 500 Birr
- የቤቱ ዋጋ ርካሽ ነው -> The price of the house is cheap

Based on the above example translations, the following translation can be done

- የቤቱ ዋጋ ከ500 ብር በላይ ነው -> The price of the house is more than 500 Birr

EBMT is an attractive approach to translation because it avoids the need for manually derived rules. However, it requires analysis and generation modules to produce the dependency trees needed for the examples database and for analyzing the sentence. Another problem with EBMT is computational efficiency, especially for large databases, although parallel computation techniques can be applied [13].

2.2.2.2. Statistical Machine Translation

The goal of translation as the production of an output that maximizes some value function that represents the importance of both faithfulness and fluency. SMT is an approach that builds probabilistic models of faithfulness and fluency, and combine these models to choose the most probable translation. The product of faithfulness and fluency is used as a quality metrics in SMT for source and target language [4] [1].

$$\textbf{Best transaltion } \hat{\mathbf{T}} = \textbf{argmax}_{\mathbf{T}} = \textbf{faithfulness}(\mathbf{T},\mathbf{S})\textbf{fluency}(\mathbf{T})$$

It is possible to make this analogy perfect and formalize the Bayesian Noisy channel model for machine translation. First, let assume every source language string $G=g_1, g_2, g_3, \ldots\ldots g_m$. We want to translate into target language. In probabilistic model the best Amharic sentence $\hat{A} = a_1, a_2, a_3, \ldots\ldots a_l$ is the one whose probability $P(A|G)$ is the highest [1] [4]. Such as in the noisy channel we can rewrite this via Bayes rule:

$$\hat{A} = \mathbf{argmax}_A \, P(A|G)$$
$$\hat{A} = \mathbf{argmax}_A \, \frac{P(G|A)P(A)}{P(G)}$$
$$\hat{A} = \mathbf{argmax}_A P(G|A)P(A)$$

We can ignore the denominator $P(G)$ inside the **argmax** since we are choosing the best target sentence for a fixed foreign sentence G and hence $P(G)$ is a constant. The resulting noisy channel equation shows that we need to components: A Translation Model $P(G|A)$ and a language Model$P(A)$.

2.2.2.3. Hybrid Machine Translation Approach

By taking the advantage of both statistical and rule-based translation methodologies, a new approach was developed, called hybrid-based approach, which has proven to have better efficiency in MT systems [36]. At present, several governmental and private sectors use this hybrid-based approach to develop machine translation from source to target languages, which is based on both rules and statistics. The hybrid approach can be used in many ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system. This technique is better example-based MT and has more power, flexibility, and control in translation.

2.3. Architecture of Statistical Machine Translation

In SMT approaches there are three components: decoder, language model and translation models [1]. The goal of language modeling is to assign n-gram (unigram, bigram…) to a sentence of target language, which is a monolingual. On the other hand, translation model is bilingual probability which is computed from the source and target languages. For the source language sentence to get well translated into target language we have to select one with highest probability in target language [1].

The overall Architecture of Statistical Machine Translation is shown in figure 2-3 below [23]**.** As you can see, from the figure an input for the system is the source text. Language model, decoder and translation model acts on the source text and finally produce a target text as output.

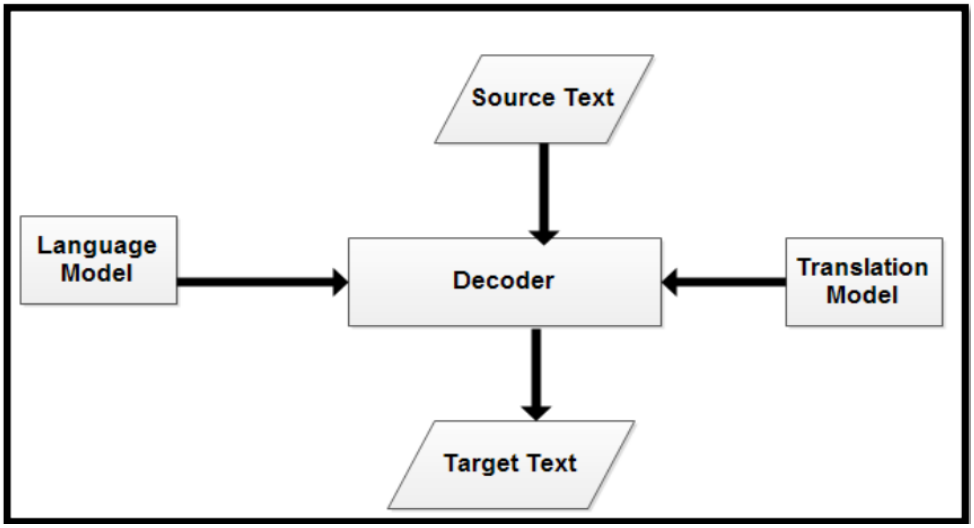


Figure 2-3 General Architecture of Statistical Machine Translation Adapted form [23]

Source and target Text: source text is a text for source language that is initializer for machine translation process to start. Target text is a text that we are going to translate to it. For example, if the translation performed from Ge'ez text to Amharic text, Ge'ez text is source text and Amharic is target text.

Language model:

A statistical language model is a probability distribution over sequences of words. Given such a sequence with length $\boldsymbol{m}$, it assigns a probability, $P(w_1, w_2, w_3 \ldots \ldots \ldots w_m)$ to the whole sequence. Having a way to estimate the relative likelihood of different phrases is useful in many natural language processing applications, especially ones that generate text as an output [32].

The intuition of the *N*-gram model is that instead of computing the probability of a word given its entire history, we **approximate** the history by just the last few words [1]. To achieve this, we apply the Markov assumptions which says that the probability of a word depends only on the previous words.

Markov models are the class of probabilistic models that assume that we predict the probability of some future unit without looking too far in to the past. Based on it different kinds of N-gram probability exists such as **Unigram**, **bigram** (looks one word in to the past), **trigram** (looks two words in to the past) and in general **N-gram** (looks N-1 words in to the past) [1].

The N-gram model performs well, for the corpus with simple sentences with the unigram, bigram and trigram models since the words in the sentence are not that long. Yet a problem exists if the sentences are too long, and the solution would be smoothing which is avoiding zero probability.

Which means by avoiding zero probability is no matter how long the decimal gets, it shouldn't be approximated to zero. Based on this method language model calculate the probabilities of N-grams which is used by decoder [1] [4].

Translation Model: To build a translation model as mentioned earlier, we should have a source language sentence (E.g. Ge'ez (G)) and target language sentence (E.g. Amharic (A)) of parallel corpus. Therefore, the job of the translation model is to assign a probability that A generates to$G$. As mentioned above, for a given source and target sentences **G** and **A**, it is the way sentences in **G** get converted to sentences in **A** which is denoted by [1] [4]:

$$\boldsymbol{P(A|G)} = \left(\frac{\boldsymbol{Count(A,G)}}{\boldsymbol{Count(G)}}\right)$$

The above equation may be difficult to achieve, if the sentences are too long. To overcome this problem the sentence is decomposed into words and sub-words called morpheme, as in language modeling [4].

$$\mathrm{p(G|A)} = \sum_{\mathrm{X}} \mathrm{p(G,X|A)}$$

The variable X represents alignments between the individual chunks in the sentence pair where the chunks in the sentence pair can be morphemes or words or phrases. In morpheme-based translation, the fundamental unit of translation is a morpheme. Phrase-based translations, most commonly used, translates whole sequences of words, where the lengths may differ in which blocks are not linguistic phrases but, phrases found using statistical methods from corpus.

Decoding: Third component of the SMT system is decoder. The main purpose of decoder is searching a best translation sentence, for the source sentence (either Ge'ez or Amharic) from the target sentence (either Amharic or Ge'ez)**,** according to the product of translation and language models.

It looks up all translations of every source morphemes, words, phrases, using word or phrase translation table and recombine the target language phrases that maximizes the translation model probability multiplied by the language model probability. From Ge'ez to Amharic translation $P(a|g)$ =$\underset{\boldsymbol{g}}{\mathbf{argmax}}\,(\boldsymbol{p(g|a)} * \boldsymbol{p(a)})$**.**

Also for translating Amharic to Ge'ez $P(g|a)$= $\underset{\boldsymbol{a}}{\mathbf{argmax}}\,(\boldsymbol{p(a|g)} * \boldsymbol{p(g)})$. By following the above procedures the decoder perform the translations of the input text for both languages.

Finally, the decoder produces the best translation of the source language text according to the product of the translation and the language models. Finding the sentence which maximizes the translation and language model probabilities is a search problem, and decoding is thus a kind of search [1]. Decoders in MT are based on best-first search, a kind of heuristic or informed search; these are search algorithms that are informed by knowledge from the problem domain. Best-first search algorithms select a node n in the search space to explore based on an evaluation function f (n). MT decoders are variants of a specific kind of best-first search called A∗ search [4].

Major components of statistical machine translation: Statistical machine translation is an approach that tries to generate translations using statistical methods based on bilingual text corpora. Statistical machine translation has three components [1].

Translation model, language model and decoder. Figure 2-4 below shows the components of the approach:

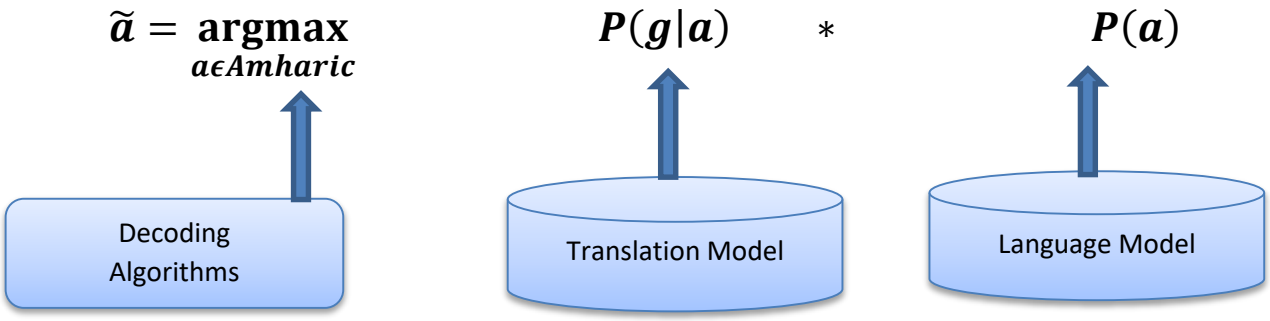


Figure 2-4 Components of Satirical Machine transaltion

If we want to translate a sentence ($\boldsymbol{g}$) in the source language ($\boldsymbol{G}$) to a sentence ($\boldsymbol{a}$) in the target language($\boldsymbol{A}$), the noisy channel model describes the process in the following ways: For example, the translated sentence ($\boldsymbol{g}$) must first considered in language($\boldsymbol{A}$), as some sentence ($\boldsymbol{a}$), during communication ($\boldsymbol{a}$) was corrupted by the channel to($\boldsymbol{g}$).

Now, assume that each sentence in($\boldsymbol{A}$), is a translation of ($\boldsymbol{g}$) with some probability, and the sentence that we choose as the translation ($\boldsymbol{X}$) is the one that has the highest probability. Let the source and target language be Ge'ez and Amharic texts. Then $\boldsymbol{p(a|g) = argmax *} \frac{\boldsymbol{p(g|a)*p(a)}}{\boldsymbol{P(g)}}$ Where $\boldsymbol{p(g|a)}$ depends on one language model (types of the sentences found in language$\boldsymbol{A}$) and second translation model (the way sentence $\boldsymbol{E}$ converted to sentence in$\boldsymbol{G}$).

Derivation of Bayes rule:$p(a|g) = (\frac{p(g|a)*p(a)}{p(g)})$ where $\boldsymbol{g}$ and $\boldsymbol{a}$ are source and target texts respectively.

$argmax * p(a|g) = argmax * (\frac{p(g|a)*p(a)}{p(g)})$ By combining the questions, we gate

$X = argmax * (\frac{p(g|a)*p(a)}{p(g)})$ Which is used by the decoder for translation process.

Challenges of Statistical Machine Translation Approach

There are different challenges that SMT has been confronting during transaltion. Some of them are discussed below [35].

Sentence Alignment: In parallel corpora single sentences in one language can be found translated into several sentences in the other and vice versa. Sentence aligning can be performed through the Gale-Church alignment algorithm [39].

Statistical Anomalies: Real-world training sets may override translations of, say, proper nouns. An example would be that "I took the train to Berlin" gets miss-translated as "I took the train to Paris" due to an abundance of "train to Paris" in the training set.

Data Dilution: This is a common anomaly caused when attempting to construct a new statistical model (engine) to represent a distinct terminology (for a specific corporate brand or domain). Training sets used from alternative sources to the specific brand to compensate for a limited quantity of brand specific corpora may 'dilute' brand terminology, choice of words, text format and style.

Idioms: Depending on the corpora used, idioms may not translate "idiomatically".

Different word orders: Word order in languages differ. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence and one can talk, for instance, of SVO or VSO languages. There are also additional differences in word orders, for instance, where modifiers for nouns are located, or where the same words are used as a question or a statement. Corpus creation can be costly for users with limited resources.

The results are unexpected. Superficial fluency can be deceiving. Statistical machine translation does not work well between languages that have significantly different word orders (e.g. Japanese and European languages). The benefits are overemphasized for European languages.

2.4. Alignment in MT

The growing availability of bilingual, machine-readable texts has stimulated interest in methods for extracting linguistically valuable information from such texts [40]. A parallel segmentation of the two texts, typically into small logical units such as sentences, such that the n^{th} segment of the first text and the n^{th} segment of the second are mutual translations known as alignment [41].

Current word alignment models for statistical machine translation do not address morphology beyond merely splitting words. However, current alignment models do not consider the morpheme, the smallest unit of syntax, beyond merely splitting words. Since morphology has not been addressed explicitly in word alignment models, researchers have resorted to tweaking SMT systems by manipulating the content and the form of what should be the so-called "word".

Since the word is the smallest unit of translation from the standpoint of word alignment models, the central focus of this research is on translating morphologically rich languages (Ge'ez and Amharic) by decomposing of morphologically complex words into tokens of the right granularity and representation for machine translation [42]. Morpheme is the focus of this study as a translation unit.

Sentence alignment represents the basis for computer-assisted translation, terminology management, word alignment and cross linguistic information retrieval [43]. Sentence alignment is the problem of, given a parallel text, finding a bipartite graph matching minimal groups of sentences in one language to their translated counterparts.

Because sentences do not always align 1-to-1, the sentence alignment task is non-trivial [44]. Sentence alignment means identifying which sentence in the target language is a translation of which one in the source language [45]. Automatic sentence alignment methods typically face two kinds of difficulties called robustness and accuracy [41].

For any statistical machine translation system, the size and domain of the parallel corpus used strongly influences the quality of translations produced [46]. Sentence-aligned parallel bilingual corpora have proved very useful for applying machine learning to machine translation, but they usually do not originate in sentence aligned form. This makes the task of aligning such a corpus of considerable interest, and several methods have been developed to solve this problem. Ideally, a sentence alignment method should be fast, highly accurate, and require no special knowledge about the corpus of the two languages [47]. Based on the above concepts sentence alignment of parallel corpus affect the performance of the machine translation especially on statistical machine translation. Following the standard alignment models of Brown et al. [48], we assume one-to-many alignment for both words and morphemes. A word alignment $\boldsymbol{a_w}$ is a function mapping a set of word positions in a source language sentence to a set of word positions in a target language sentence [42].

A morpheme alignment a_m is a function mapping a set of morpheme positions in a source language sentence to a set of morpheme positions in a target language sentence. A morpheme position is a pair of integers (j, k), which defines a word position j and a relative morpheme position k in the word at position j [42].

2.4.1. Alignment Tools

Parallel corpora are usually a collection of documents which are translations of each other. To be useful in NLP applications such as word alignment or machine translation, they first must be aligned at the sentence level [39]. There are different tools and algorithms used for aligning corpus for different purpose for text processing [39]. The common tool is MIGIZA++ [49].

MGIZA++ is a software based on the famous word-alignment software GIZA++. Since GIZA++ is a signal-processing software and the processing of GIZA++ is time-consuming, MGIZA++ modify the structure of GIZA++ and then support the multi-thread architecture.

GIZA++ is part of the statistical machine translation toolkit used to train IBM Model 1 to Model 5 [40] and the Hidden Markov Model. It is part of the SMT toolkit EGYPT which was developed by the SMT team during the summer workshop in 1999 at the Center for Language and Speech Processing at Johns Hopkins University [50]. Lexical translation is simple model for machine translation that is based solely on, the translation of words in isolation. This requires a dictionary that maps words form one language to another [4] [51] [47] [52].

IBM Translation Model

Consider all statistical translation models are based on the idea of a word alignment. A word alignment is mapping between the source words and the target words in the set of parallel sentences.

The IBM models offer principled probabilistic formulation and (mostly) tractable inference. There are five IBM models namely IBM Model 1, to IBM 5 [40].

IBM Model 1

It is the simplest of all the other models. It uses Lexical translation probabilities and the notion of alignment allows us to define a model that generates many different translations for a sentence, each with different probabilities. Given source language Ge'ez and target language Amharic.

The goal is Ge'ez to Amharic translation. Let $\boldsymbol{m}$ and $\boldsymbol{l}$ is the length of Ge'ez and Amharic sentence respectively. IBM model $\boldsymbol{p(g|a)}$ directly with no intermediate structure.

A critical idea in IBM model was to define the idea of alignment between source and target languages. An Alignment $\boldsymbol{a}$ identities which Amhari

c word each Ge'ez word originated from.

Formally, an alignment a is:

$\{a_1, a_2, a_3, \dots\dots, a_m\}$ Where $a_j \ \epsilon\ 0,1,2, \dots\dots l$

$g_1, g_2, g_3, \dots\dots\dots\dots\dots, g_m$ $\quad source\ language$

$\boldsymbol{a_1, a_2, a_3}, \dots\dots\dots\dots\dots, \boldsymbol{a_l}$ $\quad target\ language$

For Amharic word there are $((l+1)^m)$ possible alignments. Consider the example given in figure 2-5 where Amharic sentence is the source language and Ge'ez sentence is target language both with five words length

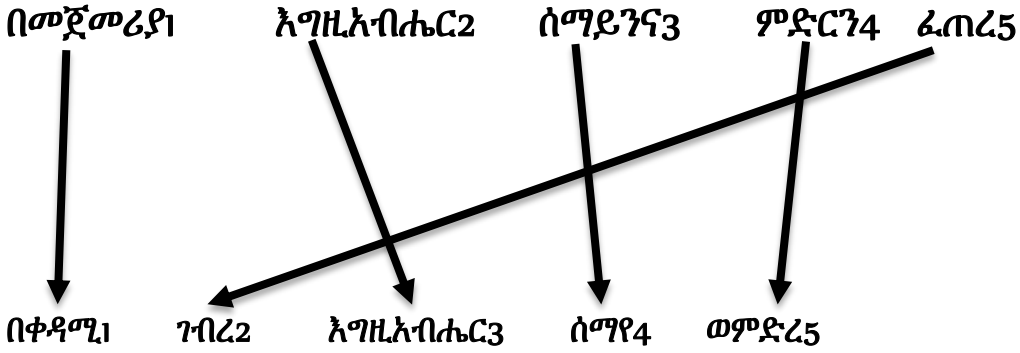


Figure 2-5 Alignment Example

The relationship between alignment and translation can be expressed as follows: These two models $p\ (a|a,m)$ and $p(g|a,a,m)$ are used to compute alignments and translation probabilities of IBM Model 1. All alignments $\boldsymbol{a}$ are equally likely. The generative process to generate a Ge'ez string g from Amharic string$\boldsymbol{a}$ [1]**.**

- Step 1: pick an alignment a with probability of

$$p(\boldsymbol{a}|a,m) = \frac{1}{(1+l)^m}$$

- Step 2: pick the Ge'ez words with the translation probabilities

$$p(g|a,\boldsymbol{a},m) = \prod_{j=1}^{m} t\left(g_j \middle| \boldsymbol{a}_{a_j}\right)$$

The result:

$$p(g,a|\boldsymbol{a},m) = p(a|\boldsymbol{a},m) * p(g|a,\boldsymbol{a},m)$$

$$p(g,a|\boldsymbol{a},m) = \frac{1}{(1+l)^m} * \prod_{j=1}^{m} t\left(g_j \middle| \boldsymbol{a}_{a_j}\right)$$

For the above example:

$l = 5, m = 5$

$\mathbf{a} = \{\mathbf{1}, \mathbf{5}, \mathbf{2}, \mathbf{3}, \mathbf{4}\}$

$$p(g|a,a) = t(\text{በቀዳሚ}|\text{በመጀመሪያ}) * \mathrm{t}(\text{እግዚአብሔር}|\text{እግዚአብሔር}) * \mathrm{t}(\text{ሰማየ}|\text{ሰማይንና}) * \mathrm{t}(\text{ወምድረ}|\text{ምድርን}) * \mathrm{t}(\text{ገብረ}|\text{ፈጠረ})$$

$$p(\boldsymbol{a}|a,m) = \frac{(1)}{(6)^5}$$

IBM Model 1 is weak in terms of conducting reordering or adding and dropping words. In most cases, words that follow each other in one language would have a different order after translation, but IBM Model 1 treats all kinds of reordering as equally possible.

Another problem while aligning is the fertility (the notion that input words would produce a specific number of output words after translation). In most cases one input word will be translated into one single word, but some words may produce multiple words or even get dropped (produce no words at all). The fertility of word models addresses this aspect of translation.

While adding additional components increases the complexity of models, the main principles of IBM Model 1 are constant. Nowadays, the original IBM models are rarely used for translation, but they are used to recover the alignment.

IBM Model 2

In IBM Model 1, we do not have a probabilistic model for alignment aspect of translation. Consequently, according to IBM Model 1 the translation probabilities for the following two alternative translations are the same.

IBM Model 2 addresses the issue of alignment with an explicit model for alignment based on the positions of the input and output words. The translation of a foreign input word in position *i* to an English word in position *j* is modeled by an **alignment probability distribution.**

$$a(\boldsymbol{i}|\,\boldsymbol{j}, \boldsymbol{l}, \boldsymbol{m})$$

Where $i = index\ of\ Amahric\ word$

$J = index\ of\ Geez\ word$

$l = length\ of\ Geez\ sentence$

$m = length\ of\ Amharic\ Sentence$

IBM Model 2 is a two-step translation process such as lexical translation and an alignment step: as shown in figure 2-6 below.

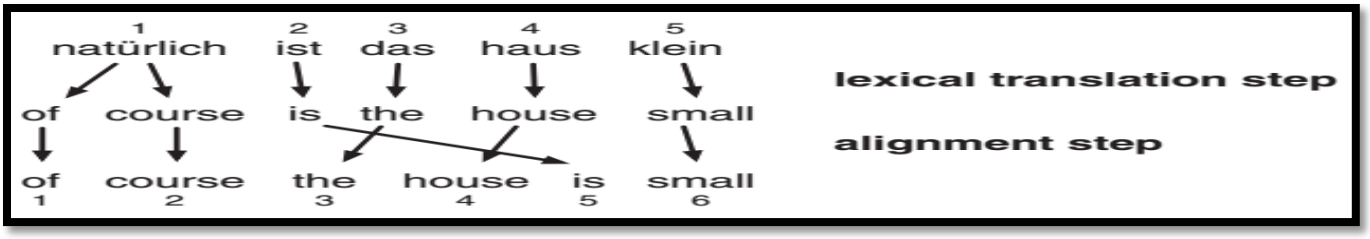


Figure 2-6 Lexical translation and alignment probability using IBM model 2

Generative processes for translating from Ge'ez to Amharic

- Step 1: pick an alignment $a = \{a_1, a_2, \ldots\ldots a_m\}$ with the probability of

$$p(a|\boldsymbol{a}, m) = \prod_{j=1}^{m}(a_j|j, l, m)$$

- Step 2: Pick the Ge'ez word with the probability of

$$p(g|a, \boldsymbol{a}, m) = \prod_{j=1}^{m} t\left(g_j\middle|\boldsymbol{a}_{a_j}\right)$$

Finally, $p(g, a|\boldsymbol{a}, m) = p(a|\boldsymbol{a}, m) * p(g|a, \boldsymbol{a}, m)$

$$\boldsymbol{p(g, a|a, m)} = \prod_{\boldsymbol{j=1}}^{\boldsymbol{m}}\left(\boldsymbol{a_j}\middle|\boldsymbol{j, l, m}\right) * \boldsymbol{t}\left(\boldsymbol{g_j}\middle|\boldsymbol{a_{a_j}}\right)$$

Note that the alignment function a maps each Amharic output word j to a foreign input position $a(j)$ and the alignment probability distribution is also set up in this reverse direction. The two steps are combined mathematically to form IBM Model 2 [1]:

$$\boldsymbol{p(a, a|g)} = \epsilon\prod_{\boldsymbol{j=1}}^{\boldsymbol{m}}\left(\boldsymbol{a_j}\middle|\boldsymbol{g_{a_j}}\right) * \boldsymbol{a(a_j|j, l, m)}$$

IBM Model 3

A single word in the source language may not be translated into a single word in the target language. For each source language word$(\boldsymbol{w_i})$,$(\boldsymbol{\varphi}|\boldsymbol{w_i})$) probability distribution indicates how many $\boldsymbol{\varphi} = \boldsymbol{0, 1, 2}, \ldots$ output words it usually translates to. Fertility deals explicitly with dropping input words by allowing$\varphi = 0$. We could model the fertility of the NULL token in the same way as for all the other words by the conditional distribution$\boldsymbol{n(\varphi|NULL)}$. However, the number of inserted words clearly depends on the sentence length, so we choose to model **NULL insertion** as a special step. After the fertility step, we introduce one NULL token with probability $p1$ after each generated word, or no NULL token with probability$p_0 = 1 - p_1$.

The addition of fertility and NULL token insertion increases the process in **IBM Model 3** to four steps [40] in figure 2-7.

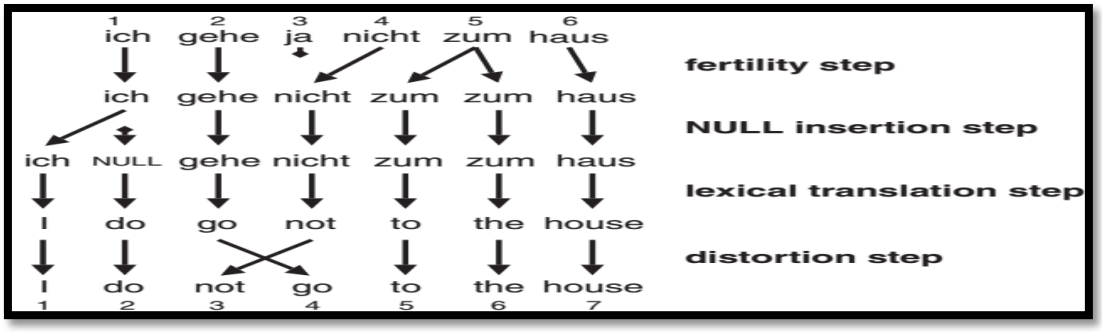


Figure 2-7 Alignment probability using 4 steps IBM model 3

The last step is called distortion instead of alignment because it is possible to produce the same translation with the same alignment in different ways. Mathematically, IBM Model 3 can be expressed as:

$$p(S|E,A) = \prod_{i=1}^{I} \varphi_i ! \, n(\varphi|e_j) * \prod_{j=1}^{J} t\left(f_i \middle| e_{a_j}\right) * \prod_{j:a(j)\neq 0}^{J} d(j|a_j, I, J) * \binom{J-\varphi_0}{\varphi_0} p_0^{\varphi_0} p_1^{J}$$

Where φ_i represents the fertility of e_i, each source word S is assigned a fertility distribution n, I and J refer to the absolute lengths of the target and source sentences, respectively.

Model 3 is already a powerful model for statistical machine translation that accounts for the major transformations in a word-based translation process: translation of words (T-table), reordering (distortion), insertion of words (NULL insertion), dropping of words (words with fertility 0), and one-to-many translation (fertility).

IBM Model 4

The set of distortion probabilities for each source and target position (i.e., the probability of a word in the source sentence change its position in the target sentence). As opposed to Model 2 which does absolute reordering, model 4 does relative reordering.

IBM Model 5

According to IBM model 4, it is possible that multiple output words may be placed in the same position. In other words, some impossible alignments have positive probability according to the model. Model 5 fixes this problem and eliminates deficiency. It also resolves the problem of multiple tableaux for the same alignment.

In general, IBM models use a modeling technique called the **noisy channel model**, which allows them to break up the translation task into a translation model and a language model, which ensures fluent output.

IBM Model 1 uses only lexical translation probabilities, Model 2 adds an **absolute alignment model,** Model 3 adds a **fertility** model, Model 4 replaces the absolute alignment model with a **relative alignment model**, and Model 5 fixes a problem with **deficiency** in the model (assigning probability mass to impossible alignments). One important concept introduced by the IBM models is the **word alignment** between a sentence and its translation. The task of word alignment is interesting for a variety of uses. The quality of word alignment can be measured with the **alignment error rate** (AER). One method to improve word alignment is the **summarization** of IBM model alignments.

2.5. Morphological Segmentation

Morphological segmentation is an important sub-task in many natural language processing (NLP) applications, aiming to break words into meaning-bearing sub-word units called morphemes [53] [54]. Numerous methods in NLP, information retrieval, and text mining make use of word-level information. However, since the number of word forms in a language is often infinite, morphological preprocessing may be vital for such methods to generalize to new forms [54]. Morphological segmentation may allow us to break them down into more familiar units that have been observed before in the data.

2.5.1. Segmentation tools

Morfessor is an unsupervised data-driven method for the segmentation of words into morpheme like units [49]. The general idea behind the Morfessor model is to discover as compact a description of the input text data as possible. Substrings occurring frequently enough in several different word forms are proposed as morphs and the words are then represented as a concatenation of morphs, e.g., 'hand, hand+s, left+hand+ed, hand+ful'.

From the alignment tools mentioned above we used MGIZA++ and morfessor for word level, morpheme level alignment and used for finding the morphological segmentation from raw text data respectively because, these tools go with our objective and they are current tools used in SMT research area.

In the theory of linguistic morphology, morphemes are the smallest meaning-bearing elements of language. Any word form can be expressed as a combination of morphemes, as for instance the following English words: 'arrange+ment+s, foot+print, mathematic+ian+'s, un+fail+ing+ly' [49].

For this research we used morfessor as a segmentation tool to segment corpus for both language prepared. The segmentation process uses corpus as an input and sets of morpheme-like structure called **morph** as output.

2.5.2. Identifying Morphemes

Morfessor Baseline takes a corpus as input and segments its words into a set of morphs without labeling them [55]. The morfessor algorithm is based on the Maximum Aposteriori estimate. The algorithm is looking for a much that has the highest probability in the given the corpus:

$$M^* = argmax_M P(M|Corpus) = argmax_M P(Corpus|M) * P(M) \quad \text{.....2.5.1}$$

The Maximum Aposteriori Estimate consists of two parts:

Where $P(Corpus|M) =$ the *maximum likelihood* estimate of the corpus conditioned on the given model of language.

$P(M)$ = the probability of the model of language.

The model consists of the lexicon of morphs and a description of how the morphs can be combined, the grammar:

$$P(M) = P\big((L, grammar)\big) \quad \text{.....2.5.2}$$

Where $L = \{\mu_1, \mu_2, \dots\dots\dots, \mu_{|L|}\}$ is the morph lexicon.

The Morfessor Baseline model does not consider any contextual information for morphs: it assumes that a morph is as likely to be used no matter what morphs precede or follow it. Thus, there is no grammar as such and the model probability is just the probability of the lexicon:

$$P(M) = P(L) \quad \text{.....2.5.3}$$

The probability of the lexicon is calculated as the probability of coming up with morphs:

$$P(L) = |L|!\, P\left(properties(\mu_1), prooperties(\mu_1), \dots\dots, properties\big(\mu_{|L|}\big)\right) \quad \text{.....2.5.4}$$

Where the properties of an individual morph within the paradigm of this algorithm is nothing but its frequency and its form, a string of characters. Assuming independence of strings and frequencies.

$$P\left(properties(\mu_1), \dots\dots, properties\big(\mu_{|L|}\big)\right) = P\left(f_{\mu_1}, \dots f_{\mu_{|L|}}\right) P\left(s_{\mu_1}, \dots\dots\dots s_{\mu_{|L|}}\right) \quad \text{.....2.5.5}$$

To estimate probability distribution of the morph frequencies Morfessor Baseline uses the non-informative prior:

$$P\left(f_{\mu_1}, \dots f_{\mu_{|L|}}\right) = \frac{1}{\binom{N-1}{|L|-1}} \quad \text{.....2.5.6}$$

Where $N = \sum_{J=1}^{|L|} f_{\mu_j}$ (number of morph tokens in the corpus).

It is also assumed that all the morphs are independent from each other:

$$P\left(s_{\mu_1}, \ldots\ldots\ldots s_{\mu_{|L|}}\right) = \prod_{k=1}^{|L|} P\left(s_{\mu_k}\right) \quad \text{2.5.7}$$

and all the characters within the morph are also independent:

$$P\left(s_{\mu_k}\right) = \prod_{k=1}^{l_k} P(C_{ik}) \quad \text{2.5.8}$$

Where $s_{\mu_k} = C_{1k}, \ldots C_{l_k}$, and $P(C_{ik})$ is the character probability distribution over the alphabet estimated by counting its frequency in the corpus.

The probability of a morph being of a length assumed to be exponentially distributed:

$$P(l) = \left(1 - P(\#)\right)^l P(\#) \quad \text{2.5.9}$$

Where # is a special end-of-morph character.

With all the independence assumption mentioned above the probability of the corpus given the model is the product of probabilities of all the morph tokens:

$$P\left((Corpus|M)\right) = \prod_{j=1}^{W} \prod_{k=1}^{n_j} P\left(\mu_{jk}\right) \quad \text{2.5.10}$$

Where W is the number of tokens in the corpus and $P(\mu_i)$is estimated by counting its frequency:

$$P(\mu_i) = \frac{f_{\mu_j}}{\sum_{j=1}^{|L|} f_{\mu_j}} \quad \text{2.5.11}$$

The algorithm uses the following data structure [55].

1. Every word type is assigned a binary tree, which is referred to as a split tree; the word itself is the root of the tree. If the word is not split its split tree consists of just the root. Otherwise, the word is split in two; the segments are the children; each segment may also be split in two and so on. The leaves of the split tree are the morphs.
2. The data structure contains all the split trees such that the nodes are shared between the trees. Thus, each node is present in the structure only once; each non-leaf node has two children; any node can have any number of parents.
3. Each node is associated with its frequency (occurrence count in the corpus). The frequency of each node is exactly the sum of frequencies of all its parents.
4. The set of leaves of this structure is the morph lexicon.

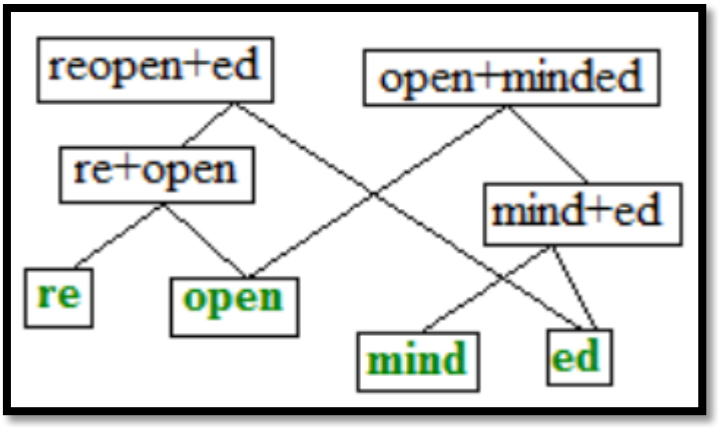


Figure 2-8 The Morfessor Baseline data structure containing the split trees of the words

2.6. MT Evaluation

Evaluating the quality of a translation is an extremely subjective task, and disagreements about evaluation methodology are rampant. Two types of raters exist in MT; namely, human and automatic raters [1] [4].

Human raters

The most accurate evaluations use human raters to evaluate each translation along each dimension. For example, along the dimension of fluency, we can ask how intelligible, how clear, how readable, or how natural is the MT output (the target translated text). There are two broad ways to use human raters to answer these questions [4].

One method is to give the raters a K-point scale, for example from 1 (totally unintelligible) to 5 (totally intelligible) and ask them to rate each sentence or paragraph of the MT output. We can use distinct scales for any of the aspects of fluency, such as clarity, naturalness, or style. The second class of methods relies less on the conscious decisions of the participants. For example, we can measure the time it takes for the raters to read each output sentence or paragraph. Clearer or more fluent sentences should be faster or easier to read.

A similar variety of metrics can be used to judge the second dimension, fidelity. Two common aspects of fidelity which are measured are adequacy and informativeness [1]. The adequacy of a translation is whether the translated text contains the information that existed in the original. Adequacy is measured by using raters to assign scores on a scale. If we have bilingual raters, we can give them the source sentence and a proposed target sentence, and rate, perhaps on a 5-point scale, how much of the information in the source was preserved in the target.

If we only have monolingual raters, but we have a good human translation of the source text, we can give the monolingual raters the human reference translation and a target machine translation, and again rate how much information is preserved. The informativeness of a translation is a task-based evaluation of whether there is sufficient information in the MT output to perform some task.

For example, given multiple-choice questions about the content of the material in the source sentence or text, the raters answer these questions based only on the MT output. The percentage of correct answers is an in formativeness score. Another set of metrics attempt to judge the overall quality of a translation, combining fluency and fidelity. For example, the typical evaluation metric for MT output to be post-edited is the edit cost of post-editing the MT output into a good translation. For example, one can measure the number of words, the amount of time, or the number of keystrokes required for a human to correct the output to an acceptable level.

Fidelity and fluency are two major dimensions while evaluating a SMT systems. SMT can be evaluated using Human Rater and automatically [1]. Human evaluations of machine translation are extensive but expensive. Human evaluations can take months to finish and involve human labor that cannot be reused. **Automatic Evaluation BLEU**

While humans produce the best evaluations of machine translation output, running a human evaluation can be very time-consuming, taking days or even weeks. It is useful to have an automatic metric that runs relatively frequently to quickly evaluate potential system improvements [32].

There are different types of heuristic methods, such as BLEU, NIST, TER, Precision and Recall, and METEOR [1]. All heuristic methods except Bleu requires human transaltion and time-consuming. In BLEU each MT output is evaluated by a weighted average of the number of *N*-gram overlaps with the human translation.

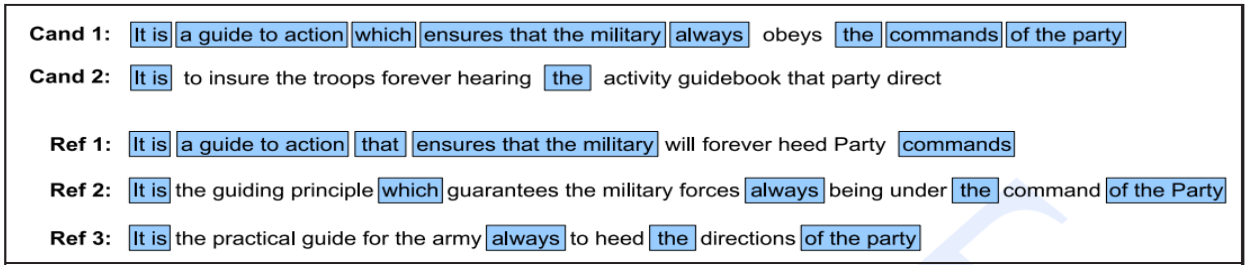


Figure 2-9 Intuition for BLEU: one of two candidate translations of a source sentence language shares more words with the reference human translations [1]

The Bleu score is computed, starting with just unigrams. BLEU is based on precision. A basic unigram precision metric would be to count the number of words in the candidate translation (MT output) that occur in some reference translation and divide by the total number of words in the candidate translation.

If a candidate translation had 10 words, and 6 of them occurred in at least one of the reference translations, we would have a precision of 6/10 = 0.6. There is a flaw in using simple precision: it rewards candidates that have extra repeated words.

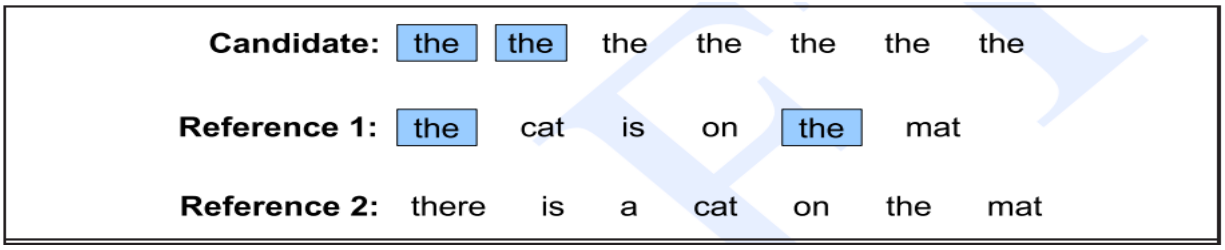


Figure 2-10 A pathological example showing why Bleu uses a modified precision metric

Figure 2-10 shows an example of a pathological candidate sentence composed of multiple instances of the single word. Since each of the 7 (identical) words in the candidate occur in one of the reference translations, the unigram precision would be unreasonably high (7/7)!

To avoid this problem, Bleu uses a modified N-gram precision metric. We first count the maximum number of times a word is used in any single reference translation. The count of each candidate word is then clipped by this maximum reference count. Thus, the modified unigram precision in the example in figure 2-10 would be 2/7, since Reference 1 has a maximum of 2 **the**'s.

To compute a score over the whole test set, Bleu first computes the N-gram matches for each sentence and add together the clipped counts over all the candidates' sentences and divide by the total number of candidate N-grams in the test set. The modified precision score is thus:

$$\boldsymbol{p_n} = \frac{\left(\sum_{C\in\{Candidates\}}\sum_{n-gram\in C} \boldsymbol{Count_{Clip}}(n-gram)\right)}{\sum_{C'\in\{Candidates\}}\sum_{n-gram'\in C'} \boldsymbol{Count}(n-gram')} \quad [1]$$

Bleu uses unigram, bigrams, trigrams, and often quad grams; it combines these modified *N*-gram precisions together by taking their geometric mean. In addition, BLEU adds a further penalty to penalize candidate translations.

2.7. Related works

This section discusses related works done in Machine translation using different approaches and methodologies by foreign and local researchers. The researchers are related to our study:

2.7.1. International languages

(a) Morphology-Aware Statistical Machine Translation Based on Morphs Induced in an Unsupervised Manner

The research was conducted by Sami Virpioja and its friends [56] at Helsinki University of Technology in Finland. As described by the researchers, Statistical machine translation was applied to the direct translation between eleven European languages, all those present in the Europarl corpus.

An impressive number of 110 different translation systems were created, one for each language pair. Koehn discovered that the most difficult language to translate from to is Finnish. Finnish is a non-Indo-European language and is well known for its **extremely rich morphology**. As verbs and nouns can, in theory, have hundreds and even thousands of word forms, data scarcity and out-of-vocabulary words present a huge problem even when large corpora are available.

It appears that especially translating into a morphologically rich language poses an even bigger problem than translating from such a language. The study also showed that English, which has almost exclusively been used as the target language, was the easiest language to translate into. Thus, it is natural to suspect that English as a target language has biased SMT research.

The researchers apply a method of unsupervised morphology learning to a state-of-the-art phrase-based statistical machine translation (SMT) system. In SMT, words are traditionally used as the smallest units of translation. Such a system generalizes poorly to word forms that do not occur in the training data. This is problematic for languages that are highly compounding, highly inflecting, or both. An alternative way is to use sub-word units, such as morphemes.

Morfessor is used to find statistical morpheme like units (called morphs) with the aim of reducing the size of the lexicon and improve the ability to generalize. Translation and language models are trained directly on morphs instead of words. The approach is tested on three Nordic languages (Danish, Finnish, and Swedish) that are included in the Europarl corpus consisting of the Proceedings of the European Parliament.

The state-of-the-art smoothing technique is modified Kinser–Ney interpolation. Word-based n-gram models are unsuitable for languages of rich morphology. They were using three types of language models to model the target language in our translation tasks. The two base-line models, tri-gram and quad-gram models, are trained with the SRI Language Modeling toolkit. The third is a variogram model trained with the VariKN Language Modeling toolkit.

Experiments are run on the Moses systems on all six language pairs and with both word tokens and morph tokens. Quantitative evaluation is provided with BLEU scores. To attain the objective of the research, the data were selected for our experiments consists of the proceedings of European Parliament from 1996 to 2001 in 11 languages, of which the Nordic languages Danish (da), Finnish (fi) and Swedish (vs.). All three pairs of the sentence-aligned bi-texts were preprocessed by removing XML-tags, conversion of some special characters and lowercasing all characters. The corpora were divided into training, development and test sets.

Morph segmentations were trained with Morfessor using the training sets. The segmentation models produced were utilized to segment the development and test sets. At this point, two data sets were created for each alignment pair: one with the original word tokens and the other with morph tokens. The training sets were used for language model training, and the development sets for parameter tuning. Additional filtering for the training data was performed by the Moses cleaning script, which removed sentence alignments when either part had no tokens or too many tokens or the ratio of tokens in the two languages was not appropriate. Such sentence pairs were selected into the test set in which both sentences had at least 5 words and at most 15 words. Depending on the language pair, the filtered test set had 10, 700–12, 900 sentences. Of this set, we used only the 1000 first sentences for the evaluation.

The results so far were quite interesting as such, but our main result is the comparison of the word and morph-based approaches. For this they were using those language models and maximum phrase lengths that have worked best on average, i.e., 4-gram models for both words and morphs, and a maximum phrase length of 7 for words and 10 for morphs.

Although the BLEU scores for word-based and morph-based translation are very close, the morphs do not outperform the standard word approach in their experiments.

(b) Deeper than Words: Morph-based Alignment for Statistical Machine Translation

This article which is written by Mark Fishel [57] at the University of Tartu in Estonia. He introduces a novel approach to alignment for statistical machine translation. The core idea is to align sub-word units or morphs, instead of word forms.

As indicated in the article word-based and phrase-based statistical machine translation ignores possible morphological relatedness of the words. This is more of a problem for inflectional languages, the richer their morphology, the larger the training corpus must be to cover most of the possible word forms. To solve this problem researchers came with two approaches of using morphological analysis and using unsupervised morphology. In most cases morphological analysis is used to segment the words or otherwise augment the text with morphological information. Also, recently an alternative approach of using unsupervised morphology for the same task has been introduced.

The problem with all previous work is that all preprocessing is language-specific. The recent advances no longer depend on linguistic tools, but still deduce segmentations that are language-specific, ignoring the bilingual nature of the task at hand. As indicated by the researcher the deduction of morphology is integrated with SMT training. The paper focuses on a one-sided approach, where the morphs of one language are aligned to words of the other one.

As indicate by the researcher parallel corpus is used for the source and target language of which source language is highly inflectional language such as Turkish or Finnish and target language is English or Chinese. Standard word alignment learning techniques, like the IBM models were used to align each source language word form with all its substrings. However here the alignment search space is constrained, unlike the word to word case: the selected morphs cannot intersect and must cover all the word forms.

The researcher was using Joint Learning for an Asymmetric Alignment probability for both source and target language and vice versa, to maximize the jointly maximizing the alignment probabilities. Searching for the Optimal Alignment is also the other methods used to find an alignment a for a sentence pair (e, f) with a maximum joint probability.

2.7.2. MT for Afaan Oromo Language

(a) English – Afaan Oromo Machine Translation: An Experiment Using Statistical Approach

Sisay [19] conducted a research that attempts to apply statistical machine translation approach so as to design English-to-Afaan Oromo machine translation system.

Monolingual and Parallel corpus used for the experiment was collected form governmental a non-government organization documents which exists on the web such as Constitution of FDRE (Federal Democratic Republic of Ethiopia), Universal Declaration of Human Right, proclamations of the Council of Oromia Regional State, religious documents, and other documents as these are the already translated and available documents. Then the corpus divided into 9^{th} of it for training and 1^{th} for testing the MT system. The corpus used for the experiment were preprocessed using Perl script which includes tasks like apostrophe, sentence aligning, tokenization, lowercasing and truncating long sentences that take the alignment to be out of optimality were done by those scripts. The size of the monolingual which is Afaan Oromo 62,300 sentences and bilingual corpus of 20,000 were used for conducting the experiment of which 90% and 10% used for training and testing the MT system respectively.

The experimentation of statistical machine translation of English to Afaan Oromo was conducted and a score of 17.74% was found. Although Afaan Oromo is among resource-scarce languages of the world, the result of this experiment shows that the amount of data available can be used as a good starting point to build machine translation system from English to Afaan Oromo. The researcher also recommends a lot to do on translation between the two languages so as to enhance translation accuracy make real.

(b) Bidirectional English – Afaan Oromo Machine Translation Using Hybrid Approach

The research was conducted by Jabesa Daba in 2013 for partial fulfilment of degree of MSC in computer science from Addis Ababa University, with purpose of using hybrid approach to develop a bidirectional English-Afaan Oromo Machine translation system. He conducted the experiment with previously work done by Sisay [19] which is having BLEU score of 17.74% not satisfactory and due to unidirectional problems, that is English to Afaan Oromo.

The researcher uses Hybrid approach which is the combination of corpus-based approach and rule-based approach requires the availability of bilingual parallel corpus. Parallel corpus collected from

different domain including the Holy Bible, the Constitution of FDRE, and the Criminal Code of FDRE, international conventions, Megeleta Oromia and a bulletin from Oromia health bureau. A monolingual Afaan Oromo and English corpus collected from the web. After the corpus collected it passes through preprocessing activities such as tokenization, True-casing and cleaning were used. For the experiment purpose freely available software like IRSTLM toolkit, GIZA++, and Moses for the statistical part and Python programming language for the rule part were used.

A total of 3000 English–Afaan Oromo parallel sentences for training and testing the system was used in two experiments namely Experiment I and Experiment II. From the total of 3000 parallel sentences, 2, 900 parallel sentences were used for training whereas the rest were used for testing the system. Statistical and Hybrid approach were used for Experiment I and Experiment II and Experiment II respectively.

The result of experiment I, the BLEU score methodology recorded result shows 32.39% for English to Afaan Oromo translation and 41.50% for Afaan Oromo to English translation. The result of experiment II BLEU score methodology shows that 37.41% for English to Afaan Oromo translation and 52.02% for Afaan Oromo to English translation.

As mentioned by the researcher the reason difference between both the records in the two experiments were that there is a difference between feminine and masculine representation in English and Afaan Oromo languages. The researcher concluded that hybrid approach was better than statistical approach based on the two experiments conducted for English Afaan Oromo language pair.

(c) Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation

The thesis was conducted by Yitayew Solomon in 2017 for partial fulfilment of the degree of MSc in Information Science from Addiss Ababa University, with the purpose of using statistical machine translation approach, exploring an optimal alignment for bidirectional English-Afaan Oromo MT Systems. For the researcher to have such an objective was, the research done by Sisay Adugna [19] and Jebesa Daba [22] score poor performance of BLEU score is 17% and 37% respectively, this is due to the alignment quality of the prepared data due to the unavailability of well-prepared corpus for the MT task for English to Afaan Oromo

Statistical machine translation and experimental research approach were used. FDRE criminal code, FDRE constitution; Megeleta Oromia, Holy Bible and simple sentences were used as data set or corpus for the experiments. To build the translation model, 6400 parallel sentences and 19300 and 12200 sentences, to build language model for both English and Afaan Oromo languages were used respectively. Randomly, for training 90% and 10 % testing of corpus size were used. 700 simple and 5700 complex sentences with a total of 6400 sentence used.

Moses for Mere Mortal used for statistical machine translation and integrates different toolkits which used for translation purpose such as IRSTLM for language model, Decoder for translation, MGIZA++ for word alignment. Hunalign, Anymalign and MGIZA++ where software tools, used for sentence, phrase and word level alignment respectively. BLEU score was used to evaluate the MT system. Preprocessing tasks sentence splitting, margining and true casing used to make ready the corpus for the experimentation purpose.

Six experiments were done by the researcher to select the optimal alignment quality for English to Afaan Oromo where, Experiment I and II for word level alignment, Experiment III and IV for phase level alignment and experiment V and VI for sentence level alignment.

Word level alignment when the max phrase length is 4 and min is 1which record 21% and 42% BLEU score from English-Afaan Oromo and From Afaan Oromo-English respectively. Phrase level alignment when the max phrase length is 16 and min is 4 which record 27% and 47% BLEU score from English-Afaan Oromo and From Afaan Oromo-English respectively.

Sentence level alignment when the max phrase length is 30and min is 20 which record 18% and 35% BLEU score from English-Afaan Oromo and From Afaan Oromo-English respectively.

An optimal alignment is phrase level alignment when the max phrase length is 16 and min is 4 which record 27% and 47% BLEU score from English-Afaan Oromo and from Afaan Oromo-English respectively.

Finally, the researcher recommends, better results can be achieved by using the corpus with proper alignment used for training the system. So, by increasing the size of the training data set that properly aligned at phrase level one can develop a better bi-directional English-Afaan Oromo machine translation.

2.7.3. MT for Tigrigna language

(a) English -Tigrigna Factored Statistical Machine Translation

The research was conducted by Tariku Tsegaye in 2014 for partial fulfillment of Degree of MSc from AAU with the theme of integrating Linguistic features to develop English to Tigrigna SMT System [24]. The researcher produced this theme due to there is no machine translation work done and to translate documents from English to Tigrigna, to address it to be addressed by the users of the language.

The researcher was using 31, 256 English and 31, 234 Tigrigna sentences for the experiment conducted in three corpus types. The monolingual raw data Tigrigna were collected form http://www.voanews.com/ and the Bible and bilingual raw data from bible. Sentence level segmentation and tokenization preprocessing tasks in addition to cleaning were done. Text and POS tagged Monolingual Tigrigna data were used to build the language model using SRILM toolkit. MGIZA++ for word alignment and mosses were used.

Three types of experiments were conducted namely baseline experiments, experiments with segmentation and using factored based experiments model in three different corpus type called baseline, segmented and factored respectively. The BLEU score experiment result using three corpora was 21.04 %, 22.65% and 16.5% for baseline, segmented and factored respectively.

The results of the three experiments were scored with two types of references namely segmented and unsegment. The result obtained shows that the system translates the words with a maximum accuracy of 21.04% using baseline, 22.65% using Segmented and 16.5% using factored translation system using un-segmented and segmented references.

Finally, the author recommended due to the unavailability of a full morphological analyzer for Tigrigna, the segmentation performed is using a stemmer. A complete morphological analyzer and segmented should be developed to obtain optimal result in segmented and factored translation systems.

(b) Bidirectional Tigrigna – English Statistical Machine Translation

This thesis was conducted by Mulubrahan Hailegebreal in 2017 for partial fulfillment for the Degree of MSc in Information Science from AAU with the aim “investigate the development of a bidirectional Tigrigna–English machine translation system using statistical approach” [25]. The

researcher believes that to make the documents written in both language English and Tigrigna available to the international and local community is vital in addressing the language barrier thereby reducing the effect of digital divide.

The study was using statistical machine translation approach which needs parallel corpus. Corpus data were collected from the Holy Bible, Constitution of the FDRE, and simple sentences which organized into five different corpora. Baseline SMT, morph-based and Phase-based experiments conducted in each five corpora namely **Corpus I, Corpus II, Corpus III, Corpus IV and Corpus V**. IRSTLM, GIZA++, Morfessor 1.0 and BLEU were used to build the language modeling, word alignment, segmentation purpose and automatic evaluation technique respectively. For all the corpus data 90% training and 10 % testing were used. The experiments were conducted by using three different systems called Baseline SMT, Morph-based System, and Post-Processed Segmented System respectively with similar corpora.

For Tigrigna–English language pair the experiment result shows that, post processed segmented system performs better than the other two. Due to morphology, the researcher obtained better translation accuracy in each experiment, when Tigrigna and English used as a source and target sentences respectively. Accordingly, the result obtained from the post processed experiment using corpus II has outperformed the other, and the result obtained has a BLEU score of 53.35 % for Tigrigna – English and 22.46 % for English – Tigrigna translations.

Finally, the researcher recommends that, segmentation of only preposition and conjunctions has led to a huge gain in BLEU score. Supervised segmentation of other derivational and inflectional morphs of Tigrigna language may lead to further improvement of the translation quality. This can be an area of study towards improving performance of a translation system for this language pair.

2.7.4. MT for Amharic language

(a) Preliminary experiments on English-Amharic Statistical Machine Translation

These preliminary experiments were conducted by Mulu and Laurent [58]. The main objective of the research was the need to begin empirical researches towards developing English-to-Amharic statistical machine translation. As mentioned in the article rule-based approach yet not recommended to be used for under resourced languages like Amharic due to the different linguistic knowledge, rules and resources required.

To meet their goal, the total corpus size of 632 Parliamentary corpora of which 115 had been used for the experiment. The experiment had been conducted using 18,432 English-Amharic sentence pairs extracted from these corpora to measure the accuracy of the translation system. To make ready the corpus for the experiment some preprocessing had been conducted which include text conversion, trimming, sentence splitting, sentence aligning and tokenization. The process of trimming is performed before and after aligning at document level. Hunalign had been used as a sentence aligner.

Out of the total 90% or 16,432 randomly selected sentence pairs had been used for training while the remaining 10% or 2,000 sentence pairs were used for tuning and testing. Thus, the preliminary experiment was developed using a total of 18,432 English-Amharic bilingual parallel and 254,649 monolingual corpora. There were different software resources used for the experiment in general integrated with MOSES like SRLIM to build the language model, Giza++ for building translation model and BLEU metric for evaluating the performance of the MT system.

When the researchers evaluate their MT system for English-to-Amharic SMT the baseline phrase-based BLEU score results 35.32% translation had been achieved. The preliminary experiment result shows that the EASMT can translate the basic meaning of the English sentence when translating into Amharic sentence. However, there are some strong as well as weak points in performance of the EASMT.

Keeping the storing side, to address problems like non-translated words, wrongly translated, insertion, deletion, alignment problem, preposition usage, and morphological errors they had used word segmentation on the Target side is vital.

According to these results, more experimentation and research is required to further improve the translation accuracy of the EASMT. The experiment done so far is encouraging as the translation is done from less inflected English language to a morphologically rich language Amharic.

(b) Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus

The thesis was Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Science, conducted by Eleni Teshome with the aim of using constrained corpus to design and develop English Amharic MT system which is bi-directional [21]. The reason

that initiate researcher was unavailability of Machine translation application at hand for time being used by people of both language users, for translating English to Amharic and vice versa.

As indicate in title of the research statistical machine translation approach was followed for the study. SMT needs monolingual as well as bilingual corpus for the experiment to be carried out. Accordingly, 1020 simple sentences manually prepared and 1951complex sentences from public Procurement Directive 414 and 1537 from bible, was collected. Before conducting the experiment, she made the corpus suitable for by doing preprocessing such as tokenization, true case and cleaning. Two corpora were prepared namely Corpus I for simple sentences and Corpus II for complex sentences to meet the aim. Both the corpus classified into training and testing sets with the rate of 90% and 10% for Corpus I and 98% and 2% for Corpus II respectively.

The researcher sees the result from two perspectives namely the accuracy and the time it takes to translate a sentence. The following findings were presented from the experiment.

Experiments results were recorded for all translation. The results obtained were accurate using BLEU Score methodology and preparing a questionnaire. The result obtained for the simple sentence using BLEU Score had an average of 82.22% accuracy for the English to Amharic, 90.59% for the Amharic to English and for the complex sentences, the result acquired was approximately 73.38% for the English to Amharic, 84.12% for the Amharic to English.

The results obtained from the questionnaire method, the accuracy from English to Amharic was 91% and from Amharic to English was 97% for the simple sentences and from English to Amharic was 87% and from Amharic to English was 89% for the complex sentences. And the maximum time taken for each translation to be carried out is 17 microseconds and 4.987 seconds, for the simple sentences and complex sentences respectively.

The result recorded was somehow high because the test set taken was from the corpus itself and the whole corpus was used for language modeling.

Finally, the researcher recommends, Morphological analyzers and synthesizers should be developed for Amharic and used for the translation purpose. This method decreases the size of the corpora to be used which is a magnificent idea since the language is very complex; it breaks it into pieces and makes it easier to be translated.

(c) Incremental Learning of Affix Segmentation

Wondwossen Mulugeta, Michael Gasser, and Baye Yimam [59] conducted the research with the aim of, to incrementally learn and segment affixes, using generic background knowledge and supervised machine learning approach. As described in the article, Amharic is semantic language with very complex inflectional and derivational verb morphology that need segmentation of affixes into valid morphemes.

The main reasons for conducting this research was for continuation of previous work namely, applying a machine learning approach to learn morphological rules for Amharic verbs using Inductive Logic Programming. In the research it is possible to detach affixes attached to stem and analyze the internal stem structure of the verb. As described in the article limitation of the work concerns words made up of the stem and more than one adjacent prefix or suffix; in those cases the system fails to segment the affixes.

The research describes that Amahric verbs can take up to four prefixes and up to five suffixes, and the affixes have an intricate set of co-occurrence rules. The researchers conducted necessary related work review for meeting their objective. As an approach Inductive Logic Programming and Incremental learning process was used. The researchers describe that Incremental learning use of less complex structures to be learned at early stages and move on to more complex and sophisticated structures using knowledge of previous structures as a basis. Incremental learning process was implemented using Inductive Logic Programming (ILP) which is a machine learning approach that learns rules from positive and negative examples. As described in the article the first step in the segmentation process is to detach the affix from the main stem.

Three major background predicate were rules learned through ILP namely, set_affix, template and feature. The first predicate set_affix: uses a combination of multiple ‘split’ operations to identify the prefixes and suffixes attached to the input word, the second one template, used to extract the valid template for Stem and the final one feature used to associate the identified affixes and root CV pattern with the known grammatical features from the example.

The finding of the experiment shows that the Inductive Logic Programming can also be used not only for simple morphology but also complex languages with more sophisticated background

predicates and more examples. Precision and recall is used to measure the effectiveness of the system. The system is able to do the segmentation with 0.94 Precision and 0.97 Recall rates.

(d) Ge'ez to Amharic Automatic Machine Translation: A statistical Approach

Dawit [15] to investigate Ge'ez to Amharic automatic machine translation using satirical machine translation. As stated by Dawit, the research came with the aim of addressing the Amharic speakers to get the knowledge that is decoded in Ge'ez is mandatory using automatic translation techniques.

As a research methodology, the researcher used using qualitative Experimental method to investigate the effect of variables such as normalization, corpus and test split options on the Statistical Machine Tarnation result. The researcher perform literature review on synthetic structure study for both language Geez and Amharic, in order to understand the Interlingua structures, morphological characteristics and foresee their impact on the translation. The data used for the research experiment were found from both online and manually prepared. The online document were accessed from https://bible.org/sites/bible.org/resources/foreign/amharic/ for Amharic language and https://www.tau.ac.il/~hacohen/Biblia.html for Ge'ez document.

The data collected were in HTML, MS-word, MS-Publisher and MS-Excel format. To make all this format suitable for the experiment, the researcher merge all documents to Ms.-Word format and align to verse/sentence level, cleaned for noisy characters and converted to plain text in UTF-8 format. Even if inherently data in both language were verse level aligned, but the researcher align sentences manually which is misaligned at verse and sentence level. Language expert also used for cross checking of the correct alignment of the corpus. The data set used by the researcher were biblical data. The source language is Ge'ez and the target language is Amharic.

As described by the researcher, the bilingual data used for the experiment include Old Testament Holy Bible, Genesis, Exodus, Leviticus, Numbers, Deuteronomy, Joshua, Judith, Ruth and Psalms and some religious books like Wedase Mariam and Arganon were used. 12, 860 parallel sentence were used for both language. In the same way the monolingual data used for the target language were includes all the New Testament of the Holy Bible of which, 26, 818 sentence.

Regarding the organization of the data, out of the bilingual data, 90% for training and 10% for testing were used for experiment. Moses decoder, IRSTLM, GIZA++ and BLEU were used to build translation model, language model, Word alignment and evaluation of the Ge'ez to Amharic MT system respectively. The Parallel corpus used for the experiment was sentence level aligned.

As the researcher indicate in the architecture of the SMT, monolingual data passed through only tokenization whereas bilingual data passed through both tokenization and cleaning.

As described by researcher, the translation result got high score, when the testing data taken from psalm as a whole and low when the testing data contains sentences from the praise of Saint Mary and part of the Bible using 10-flod cross validation. The result show inconsistence.

Due to this, the researcher also check the performance of the system after splitting the each book of the Bible in to training and testing set. With this he got consistence in the result of the SMT system performance. Dawit prove that increasing the data set of the target language and normalizing it increase the performance of the SMT system.

The researcher after conducting experiment, average translation accuracy of BLEU score 8.26. With the use sufficiently large parallel Ge'ez-Amharic corpus collection and language synthesizing tool, it is possible to develop a better translation system for the language pairs.

Finally, the researcher suggested, Ge'ez and Amharic are related but morphology rich languages as well limited researches have been done on the morphological segmentation and synthesizing of the two languages. The development of the languages' morphological synthesizers and segmenting tools can help for better performance. The researcher recommends extension of this research using the different morphological segmentation and synthesizing mechanisms.

Research Gap

As to the researcher knowledge there is only one study conducted to deal with Geez-Amharic unidirectional statistical machine translation. The study used word as a transaltion unit. As described by Dawit the performance of Geez-Amharic SMT affected greatly due to morphologically richness of both languages. Therefore, he recommends the need for further study to design better translation unit that takes into account morphological richness of the languages. Hence because of the availability of specific, consistent morphemes in a given language, it is better to use morpheme as a transaltion unit, especially for morphological rich languages. Accordingly the aim of this study is to experiment morpheme based bi-directional Ge'ez-Amharic Machine transaltion languages.

Chapter Three

Ge’ez and Amharic Language

3.1. Writing systems

Writing is a method of representing language in visual or tactile form. Writing systems use sets of symbols to represent the sounds of speech and may also have symbols for such things as punctuation and numerals. There are six different types of writing systems or scripts namely, Alphabets (English, Russian, Greek), Abjads (Arabic, Hebrew), Abugidas or alpha syllabaries (Devanagari, Thai, Ge’ez, Amharic), Featural alphabets (Hangul), Syllabaries (Japanese, Cherokee), and Logographic systems (E.g., Chinese characters) [60] [61].

An abjad and an abugida were used to write Ge’ez language. The abjad, used until c. 330 AD, had 26 consonantal letters. Vowels were not indicated [9]. The Ge'ez abugida developed under the influence of Christian scripture by adding obligatory vocalic diacritics to the consonantal letters. The diacritics for the vowels, u, i, a, e, ə, o, were fused with the consonants in a recognizable but slightly irregular way, so that the system is laid out as a syllabary. The original form of the consonant was used when the vowel was ä (/ə/), the so-called inherent vowel. The resulting forms are shown below in their traditional order. For some vowels, there is an eighth form for the diphthong -wa or -oa, and a ninth for-yä [14]. Before the first Patriarch for Ethiopian Aba Frimentatos, Ge’ez was written from right to left but now it is written from left to right [16]. **አዐሊዳ** And **ሀሁ** are the two types of Ge’ez alphabet arrangement called previous and current. The writing system used for Amharic language is Abugida or (alphasyllabary).

In Amharic there are 34 basic alphabets or Fidel of which 26 is derived from Ge’ez. The remaining 8 of them where by modifying 8 Ge’ez Fidel’s; namely, **ሰ** to **ሸ** ፣ **ተ** to **ቸ** ፣ **ነ** to **ኘ** ፣ **ከ** to **ኸ** ፣ **ዘ** to **ዠ** ፣ **ደ** to **ጀ** ፣ **ጠ** to **ጨ** and **ፀ** to **ፘ**.

As it is described in the above paragraph, to modify character they were using -, and o. Also, Amharic has taken the entire derived alphabet from Ge’ez. The current writing direction for both Ge’ez and Amharic is from left to right.

3.2. Syntax

The usual word order of Amharic is Subject-Object-Verb (SOV). However, if the object is tropicalized it may precede the subject (OSV). Noun phrases are head-final with adjectives and other modifiers preceding their nouns. Prepositions, postpositions or a combination of both are used to indicate syntactical relations, revealing the mixture of Semitic and Cushitic traits [6]. Whereas, the syntax of Ge'ez follows **SVO**, **VSO** and **OVS**.

	ግዕዝ	ካዕብ	ሣልስ	ራብዕ	ሐምስ	ሳድስ	ሳብዕ
፩	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
፪	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
፫	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
፬	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
፭	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
፮	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
፯	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
፰	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
፱	ጎ	ጐ	ጒ	ጓ	ጔ	ጎ	ጐ
፲	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
፳	የ	ዩ	ዪ	ያ	ዮ	ይ	ዮ
፴	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
፵	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
፶	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
፷	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
፸	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
፹	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
፺	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
፻	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
፼	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
፽	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
፾	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
፿	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
ᎀ	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
ᎁ	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
ᎂ	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
ᎃ	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ᎄ	ጎ	ጐ	ጒ	ጓ	ጔ	ጎ	ጐ
ᎅ	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
ᎆ	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
ᎇ	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
ᎈ	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
ᎉ	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
ᎊ	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
ᎋ	የ	ዩ	ዪ	ያ	ዮ	ይ	ዮ
ᎌ	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
ᎍ	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ᎎ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
ᎏ	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
᎐	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
᎑	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
᎒	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
᎓	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

(a) (b)

ሏ	ሗ	ሟ	ሧ	ሯ	ሷ	ቋ	ቧ	ቷ	ኗ	ኧ	ኳ	ዃ	ዟ	ዷ	ጓ	ጧ	ጷ	ጿ	ፇ	ፏ	ፗ

(c)

Table 3-1 Ge'ez Script Arrangement (a) Previous Ge'ez Script (b) Current Ge'ez Script (c) Derived Ge'ez Script

As it is presented in Table 3-1 (a) and (b), 7*26 = 182 basic letters exists in Ge'ez language with two arrangement Previous and Current. Table 3-1 (c) shows derived letters of Ge'ez language from the basic letters.

Amharic language takes all the basic and derived alphabets of Ge'ez besides adding letters in presented in the Table 3-2 (a). Therefor the total number of basic Amharic Alphabet is 7*34 = 238. That of the delivered Amahric scrip in Table 3-2 (b).

	ግዕዝ	ካዕብ	ሣልስ	ራብዕ	ሐምስ	ሳድስ	ሳብዕ
ⶀ	ⶐ	ⶑ	ⶒ	ⶓ	ⶔ	ⶕ	ⶖ
ⶁ	⶘	⶙	⶚	⶛	⶜	⶝	⶞
ⶂ	⶟	ⶠ	ⶡ	ⶢ	ⶣ	ⶤ	ⶥ
ⶃ	ⶦ	⶧	ⶨ	ⶩ	ⶪ	ⶫ	ⶬ
ⶄ	ⶭ	ⶮ	⶯	ⶰ	ⶱ	ⶲ	ⶳ
ⶅ	ⶴ	ⶵ	ⶶ	⶷	ⶸ	ⶹ	ⶺ
ⶆ	ⶻ	ⶼ	ⶽ	ⶾ	⶿	ⷀ	ⷁ
ⶇ	ⷂ	ⷃ	ⷄ	ⷅ	ⷆ	⷇	ⷈ

(a)

ⷉ	ⷊ	ⷋ	ⷌ	ⷍ	ⷎ	⷏	ⷐ

(b)

Table 3-2 Amharic Script (a) added script, (b) Derived script

3.3. Ge'ez Numerals

Geez is also have its own numerals for designating numbers. Amharic language also takes these numbers as it is. These numbers are used in Ethiopian yearly calendar. Table 3-3 below show the Geez and Amharic numerals.

-	ⶀ	ⶁ	ⶂ	ⶃ	ⶄ	ⶅ	ⶆ	ⶇ	ⶈ	ⶉ
አልቦ	አሐዱ	ክልኤቱ	ሠለስቱ	አርባዕቱ	ሐምስቱ	ስድስቱ	ስብዓቱ	ስመንቱ	ተሰዓቱ	አሠርቱ
0	1	2	3	4	5	6	7	8	9	10
ⷑ	ⷒ	ⷓ	ⷔ	ⷕ	ⷖ	⷗	ⷘ	ⷙ	ⷚ	
20	30	40	50	60	70	80	90	100	1000	
እስራ	ሠላሳ	አርብዓ	ሃምሳ	ስድሳ	ሰብዓ	ሰማንያ	ተሰዓ	ምዕት	እልፍ	
ⷛ	ⷜ									
አእላፋት	ትልፊታት									
1000 000	100 000 000									

Table 3-3 Ge'ez and Amahric numerals

3.4. Similar Letters (ተመኳሳይያን)

They are letters that have similar sounds. Even though they are having similar sound, the letters are different in shape orthographically. These are described below in Table 3-4.

Sound	Letters
hä	**ሀ፣ሐ፣ኀ**
sä, śä	**ሰ፣ ሠ**
ʾä , ʿä	**አ፣ዐ**
ṣä, ẓä	**ጸ፣ፀ**

Table 3-4 similar letters in Ge'ez and Amharic

Letters	Known Name	Reason
ሀ	ሃሌታው "ሀ"	Since it is the beginning of the Ge'ez word **ሃሌ**
ሐ	ሐመሩ "ሐ"	Since it is the beginning of the Ge'ez word **ሐመር**
ኀ	ብዙኃኑ "ኀ"	When the word **ብዙኃን** written it is used.
ሰ	እሳቱ "ሰ"	When the Ge'ez word **እሳት** written it is the one used.
ሠ	ንጉሡ "ሠ"	When the Ge'ez word **ንጉሡ** written it is the one used.
አ	አልፋው "አ"	The word **አልፋ** is always written using it
ዐ	ዐይኑ "ዐ"	The shape is like Eye and the Ge'ez word **ዐይን** is written using it
ጸ	ጸሎቱ "ጸ"	The Ge'ez word **ጸሎት** written it is the one used.
ፀ	ፀሐዩ "ፀ"	The shape is like sun and used to write the Ge'ez word **ፀሐይ**

Table 3-5 Similar Letters, Their Known name and reason

3.5. Word Classes

Grammar (ሰዋሰው) Structure for Ge'ez and Amharic

In linguistics, **grammar or ሰዋሰው** is the set of structural rules governing the composition of clauses, phrases, and words in any given natural language such as Ge'ez and Amharic. The term refers also to the study of such rules, and this field includes phonology, morphology, and syntax, often complemented by phonetics, semantics, and pragmatics [8].

For many people, words are the center of language. This comes as no surprise if we consider that the most obvious, concrete, and recognizable parts of any language are its words or its lexicon. In any given language, there are tens of thousands of words, although most speakers know and use only a relatively small number of them [8]. Each word that we use for speech as well as writing has its own part of speech. Based on parts of speech a word of grammarians classified words in to eight major part in both Ge'ez and Amharic [14] [10].

These are Nouns/**ስም**, Verbs/**ግሶች**, Adjectives/**ቅጽሎች**, Adverbs/**ተውሳከ ግሶች,** Pronoun/**ተውላጠ ስሞች**, Preposition/ **መስተዋድዶች**, Conjunction/**መስተጻምር** and Interjection/**ቃል አጋኖ**.

However, many grammar texts prefer to think of parts of speech in terms of **form** and **structure** classes [8]. The form classes are composed of the major parts of speech: Nouns/ስም, Verbs/ግሶች, Adjectives/ቅጽሎች, and Adverbs/ተውሳከ ግሶች. These are the words that carry the content or meaning of a sentence. The **structure** class words are composed of the minor parts of speech: Pronoun/ተውላጠ ስሞች, Preposition/ መስተዋድዶች, Conjunction/መስተጻምር and Interjection/ቃል አጋኖ. These words serve primarily to indicate grammatical relationships and are frequently referred to as structure words.

Content words, such as nouns, verbs, adjectives, and adverbs, are words that carry lexical or content meaning. These **major** class words are also referred to as **open word classes**. Structure words, such as prepositions, pronouns, conjunctions, and determiners, are words that show grammatical relationships within sentences. These **minor** class words are referred to as **closed word classes**. Speakers are endlessly creating new Ge'ez and Amharic open words, especially nouns and verbs. Therefore, the major word or form classes are called open word classes because new words enter the language constantly.

Closed word classes are among the most common and frequently used Ge'ez and Amharic words. These classes are considered "closed" for several reasons [8] [10]. First, they consist of small numbers of words that change very little over long periods of time and that have been in the English language for centuries. They include: Prepositions, determiners, coordinators and pronouns. Second, words in the closed classes are fixed and invariant, meaning that they do not have other forms. There is only one form for the preposition "in". In contrast, open class words can have different forms because they can take different beginnings and/or endings. The noun, dog, for instance, can take the plural and possessive endings (dogs or dog's); the verb walk can take three different endings (walked, walks, walking); and the adjective tall can take two different endings (taller, tallest). Third, these words occur only in a narrow range of possible positions within a sentence, and they must always accompany content words. There is no flexibility in word order. The word "the" always precedes a noun. It cannot follow a noun. We cannot say "dog the", but must say "the dog". Finally, closed word classes have little lexical or semantic function. The job of these words is to show the relationships between the different parts of sentences. Therefore, we must know the grammar (or **ሰዋሰው**) of Ge'ez and Amharic to answer the main objective of the research which is morpheme based.

3.5.1. Major Parts of Speech

3.5.1.1.Noun (ስም)

Noun is a name that represents a person, places, animal, thing, feeling and idea. In Ge'ez and Amharic there are different types of nouns in general concert and abstract, common and proper, collective and countable and countable noun. Most nouns in both ends with the sixth letter, sadese Fidel. It doesn't mean that it never ends by other letters or Fidel. In Amharic noun have suffix -ኦች/ዎች (Plural marker) በግ→በጎች/በግዎች፣ ኡ (used to show already known nouns) በግ+ኡ → በጉ and ኤ (to show subject or possession) በግ+ኤ→ በጌ.

In general in Ge'ez language there two ways of forming plural forms of a nouns. These are the following:

1. Pattern replacement: ደብር dabr --------አድባር adbar ሀገር to አህጉር ፤ ቤት to አብያት፣
2. Addition of an ending: አመት ---------- አመታት

Plurals formed by pattern replacement are often referred to as "**broken" plurals"** or "**internal**" plurals; those formed by adding suffixes, as " **external'** plurals. [62]. The two endings used to form external plurals are **-ãn** (አነ) and -ãt (አት). **-ãn** is, for the most part, restricted to nouns denoting male human being.

Most Ge'ez nouns form their plural form using broken plural or internal plural ways. In Ge'ez we use **አ ፤ አ……ት ፤ ን ፣ ው ፣ት ፣** to inflect a singular noun to Plural.

Ge'ez			**Amahric**		
Using	Original word	Inflicted to	Using	Original word	Inflicted to
አ	ልብ	**አ**ልባብ	ኦች	ልብ	ልቦች
አ……ት	ባሕር	**አ**ብሕርት		ባሕር	ባሕሮች
	ገበር	**አ**ግበርት	ዎች	ባርያ	ባርያዎች
	ጥብ	**አ**ጥበት		ጡት	ጡቶች
	ነቅዕ	**አ**ንቅዕት		ምንጭ	ምንጮች
ት	ገዳም	ገዳማት	**ት**/ኦች	ገዳም	ገዳማት/ገዳሞች
	እም	እማት	ኦች	እናት	እናቶች
ል	ኪሩብ	ኪሩቤል	ል	ኪሩብ	ኪሩቤል
	ሱራፌ	ሱራፌል		ሱራፌ	ሱራፌል
ን	ጻድቅ	ጻድ**ቃ**ን	**ን**/ኦች	ጻድቅ	ጻድ**ቃ**ን/ጻድ**ቃ**ኖች
ው	እጉ	አ**ጎ**ው		ውንድም	ውንድሞች
	አብ	አ**በ**ው		አባት	አባቶች

Table 3-6 Example of infliction in numerals in Ge'ez and Amharic

Ge’ez plural formation nouns can occur by changing the Fidel to ራብዕ ፣ ሳድስ. For example **አን (ሳድስ) ቀ (ግዕዝ)** ጽ to **አና (ራብዕ) ቅ (ሳድስ)** ፣ ደብተራ to ደባትር ፣ መክሊት to መካልይ ፣ መድሎት to መዳልው

We can also form plural noun the end of the noun is the Fidel ሣልስ ፣ ኀምስ ፣ ሳብዕ using ወ and የ as ራብዕነት (**ዊ** and **ያ**) with ት and with ን. For example ደቄ to ደቂያት ፣ አረጋዊ to አረጋዊያን ፣ ቅዳሴ to ቅዳሴያት and ዘማሪ to ዘማሪያን.

3.5.1.2. Adjectives (ቅጽል)

An adjective is a word that describes, identifies, or further defines noun or pronoun. Nouns tell about things nature, but adjectives tell about things behavior or characteristics, like shape, size, color, type, property [9]. Different types of adjectives in Ge’ez and Amharic based on property, size, shape, color, nation or nationality. They differ from noun by their usage. In Amharic adjectives repeat themselves to indicate plural number, for example ጥቁር -- ጥቋቁር ፣ ቀይ -- ቀያይ ፣ ነጭ -- ነጫጭ.

In Amharic suffix ‘**-ኢት**’ is used to show feminine gender and no gender suffix for Masculine. ቆንጆ is used for both gender but if we add the suffix ‘**-ኢት**’ →ቆንጆ + ኢት → ቆንጇት፡፡ In Ge’ez and Amharic the suffix “**ዊ**” and “**ይ**” being add in noun used to expresses belongingness of a person to a specific nation and to express the nationality of nationality of a person.

ኢትዮጵያ (ስም) --- ኢትዮጵያይ **(** ቅጽል **)** --- ኢትዮጵያዊ **(** ቅጽል **).**

In both language “**-ያን**” used to create the plural form of an adjectives ---ኢትዮጵያው**ያን.** The suffix “**-ት**” used to show the feminine ኢትዮጵያዊ**ት.** To make it plural Amharic and Ge’ez used suffix **“-ያት”** ኢትዮጵያው**ያት.** Note that the suffix “**ዊ**” changed into “**ው**” to express plural form in both case.

The use of adjectives in sentence in both language are not the same [16]. In Ge’ez language adjectives are used before and after noun where as in Amharic adjectives are used before noun language. For example,

ፍንዋን እደው ይነግሩ መልእክተ ------- **የተላኩ** ወንዶች መልእክት ይናገራሉ ፡፡

እደው **ፍንዋን** ይነግሩ መልእክተ ------- **የተላኩ** ወንዶች መልእክት ይናገራሉ ፡፡

In Ge’ez [14] [16]and Amharic [10] there are different types of adjectives. As it is depicted in table 3.7 we can inflect adjectives in Ge’ez to plural number by prefixing “እለ ፣ አ” at the beginning, suffixing “ን ፣ ዊ/ይ ፣ ያ ፣ ት ፣ ሙ ፣ ው …” at the end. On the other hand we can inflict adjectives in Amharic to form plural using “የ” and “እነ” prefixes and “ያን” and “ኦች” as a suffixes.

Ge'ez					Amahric			
No	Adjective				Adjectives			
	Singular	Plural	prefix	suffix	Singular	Plural	Prefix	Suffix
1	ፍንው	ፍንዋን		…ን	የተላከ	የተላኩ	የ……	
2	ሰማያዊ	ሰማያውያን		…ያን	ሰማያዊ	ሰማያውያን		...ያን
3	መኑ	እለመኑ	እለ…		ማን	እነማን	እነ……	
4	ውእቱ	ውእቶሙ		…ሙ	እሱ	እነሱ		
5	ሀገር	አህጉር	አ…		ሀገር	አገሮች		....አች
6	አብ	አበው		…ው	አባት	አባቶች		....አች

Table 3-7 Ge'ez and Amharic adjective suffix and Prefix

We found adjective in Ge'ez from ቀዳማይ አንቀጽ primary (Past), ካልአይ አንቀጽ secondary, and ሣልሳይ አንቀጽ tertiary verbs [14]. Adjectives in Amharic can be formed in several ways [10]: they can be based on nominal patterns, or derived from nouns, verbs and other parts of speech. Adjectives can be nominalized by way of suffixing the nominal article (see Nouns above). Amharic has few primary adjectives. Some examples are dägg 'kind, generous', dəda 'mute, dumb, silent', biče̱a 'yellow'.

3.5.1.3. Verb (ግስ)

Verb is a word used to describe an action, state, or occurrence, and forming the main part of the predicate of a sentence [9]. In Ge'ez and Amharic there are two types of verbs regular and irregular verbs based on the affix used to form. Regular verbs are main verbs that have four types; namely, (ቀዳማይ/ኀላፊ) past tense/perfect, (ካልአይ/የአሁንና እና የመጻኢ) present and future /imperfect, command and (ዘንድ) to verbs. (ትዕዛዝ) Command and (ዘንድ) to verbs are the same.

Perfect verbs show already past or completed action, which include past perfect, past continuous, past participle with relative pronoun ዘ (of). Whereas the imperfect one includes present, continuous and future action. The end of all perfect verbs is the first order while all imperfect verbs ends with the 6th order when the noun is (ውእቱ) he. Morphology of verbs starts with perfect verbs. To change imperfect verbs, it has its own rules which is expressed by the root verbs (ግስ አርእስት) [16] [11].

Root verbs (ግስ አርእስት)

Root verbs are those that leads the time behavior and using their morphology style. Other similar verbs also follows the style of root verbs morphology. Those root verbs are regular verbs. Root Verbs in Ge'ez either eight or seven, each having their own characteristics [7].

ተራ ቁጥር	የግእዝ ግስ አርእስት	የፊደል ደረጃ	Pronunciation (አነባበብ)
1	ቀተለ	(ቀ)ግእዝ (ተ)ግእዝ (ለ)ግእዝ	/ḱətələ/
2	ቀደሰ	ግእዝ ግእዝ ግእዝ	/ḱəddəsə/
3	ገብረ	ግእዝ ሳድስ ግእዝ	/gəbɨrə/
4	አእመረ	ግእዝ ሳድስ ግእዝ ግእዝ	/ʔəʔmərə/
5	ባረከ	ራብዕ ግእዝ ግእዝ	/barəkə/
6	ሤመ	ኀምስ ግእዝ	/śemə/
7	ብእለ	ሳድስ ሳድስ ግእዝ	/bɨhɨlə/
8	ቆመ	ሳብዕ ግእዝ	/ḱomə/

Table 3-8 Root/Main Verbs in Ge'ez

Table 3-8 depict the main root Verbs of Geez. Verbs morphology in Ge'ez starts with perfect/past tense and continuous to the future. The morphology starts with the pronoun (ውእቱ) he. To inflect verb in Ge'ez we need to know the root verb of the verb we need to inflect from the table 3-8. For example, ወደሰ ፣ ነጸረ are family of **ቀደሰ** since the middle sound of each word need to geminate. And accordingly, the morphology is conducted after knowing the family of the verb. Verbs in Ge'ez and Amharic languages are source for morphology of adjective, root or main verb and noun.

No	Perfect /Past		imperfect					
			Present/ Future		Command/ Verb to Be		ንዑስ አንቀጽ	
	Ge'ez	Amharic	Ge'ez	Amharic	Ge'ez	Amharic	Ge'ez	Amharic
1	**ቀተለ**	ገደለ	**ይ**ቀትል	ይግደላል	**ይ**ቅትል	ይግደል	**ቀቲል/ቀቲሎት**	መግደል
2	**ቀደሰ**	አመሰገነ/ለየ	**ይ**ቄድስ	ይቀድሳል	**ይ**ቄድስ	ይቀድስ	ቀድሶ/ቀድሶት	ማመስገን
3	**ባረከ**	ባረከ	ይባርክ	ይባርካል	ይባርክ	ይባርክ	ባርኮ/ባርኮት	መመረቅ/ማመስገን

Table 3-9 Root verb of Ge'ez and Amharic

Active (**ገቢር**/) and passive **(ተገብሮ/)** voice are the two types of Verb [11]. Active voice verb should have a subject and an object. አብርሃም **ወለደ** ይስሐቅ ።። **ወለደ** is connect a subject አብርሃም and object ይስሐቅ.

Passive voice verbs when the subject of the sentence is acted on by the verb and further divided into passive voice verb that add a prefix **'ተ'** for example ይስሐቅ **ተ**ወልደ at the beginning of the verb and that doesn't add **'ተ' መጽአ ብእሲ.**

Ge'ez and Amharic verbs have two main characteristics; namely, how they are written and usage of affixes. In Ge'ez language verbs are written using alphabets or Fidels Ge'ez **(ግዕዝ),** rabe **(ራብዕ),** hamese **(ሐምስ),** sadese **(ሳድስ),** and Sabe **(ሳብዕ).** This is based on the first script of the verb. Verbs in Ge'ez never start with kahbe and Salese.

The five primary anktse of Ge'ez, in teachers of Ge'ez they are called **መራሁተ ግስ.** Unlike Ge'ez, Amharic use only Ge'ez script to write verbs. Other scripts are not used. For example, **ሰበረ ፣ ወፈረ ፣ ፈረደ ፣ ገደለ ፣ ጨፈረ ፣ ወሰደ ፣ ዘመረ ፣ ሰገደ ፣ ጨለጠ** and so on.

The other property is using of affixes [prefix, suffixes, infixes, and circumfix]. In both language verbs are using affixes for inflectional morphology. Affixes are morphemes that are sub words of a word. Based on affixes usage two types of morphemes exists. The one that inflect verbs in number, gender, tense and if the newly formed word class is same as that the first such a morpheme is called **Inflectional Morphemes**. **Derivational morphemes** are responsible not only for the formation of new word but also the word class of the new word also different from that of the previous one. Let us discuss each of the types of affixes in both language.

ዝርዝር / Suffixes

In grammar of Ge'ez and Amharic, these are morphemes that are suffixed at the end of verbs to show Number (Singular or Plural), Gender (Masculine or feminine), nearness or farness, either by mentioning in script or by changing the sound, to indicate the subjectivity or objectivity. Suffixes means indicator, pointer and shape /script/ sound. 10 and 8 pronouns exist in Ge'ez and Amharic respectively. For example, Ge'ez **(ቀተለ)** and Amharic **(ገደለ)** [63]**.**

There are two types of suffixes. Verbal suffixes and **ነባራዊ/ የነባር ቃል/.** Verbal suffixes also group in to subjective Zmde and objective Bahd suffixes. Suffixing morphemes at the end of a verb to indicate only the subject Gender, Number and nearness or farness called subjective suffixes [63] [9]. If the morpheme is mentioned with script it is called subjective Zmde (ዘመድ) suffix and if it is not mentioned Subjective Bahd (ባዕድ) suffix. As you see from the table those **ውእቱ ፣ ውእቶሙ and ውእቶን** in Ge'ez and **እሱ ና እነሱ** in Amharic are subjective Zmde **(ዘመድ)** suffix and the rest are Subjective Bahd **(ባዕድ)** suffix.

Every Verbs in Ge’ez and Amharic inflected using pronouns and suffixes of each language.

Ge’ez			Amharic			English	
Pronoun	ግሱ/Verb	ዝርዝር/Suffixes	Pronoun	ግሱ/Verb	ዝርዝር/suffixes	Pronoun	verb
አነ	ቀተልኩ	-ኩ	እኔ	ገደልኩ	-ኩ	I	I Killed
አንተ	ቀተልከ	-ከ	አንተ	ገደልክ	-ክ	You	You killed
አንቲ	ቀተልኪ	-ኪ		ገደልሽ	-ሽ		
ውእቱ	ቀተለ	ግዕዝ ድምጽ (ኧ)	እሱ	ገደለ	ግዕዝ ድምጽ (ኧ)	He	He killed
ይእቲ	ቀተለት	-ት	እሷ	ገደለች	-ኧች	She	She killed
ንሕነ	ቀተልነ	-ነ	እኛ	ገደልን	-ን	We	We killed
አንትሙ	ቀተልክሙ	-ክሙ	እናንተ	ገደላችሁ	-ኣችሁ	They	They killed
አንትን	ቀተልክን	-ክን					
ውእቶሙ	ቀተሉ	-ኡ (ካዕብ ድምጽ)	እነሱ	ገደሉ	-ኡ (ካዕብ ድምጽ)		
ውእቶን	ቀተላ	-ኣ (ራብዕ ድምጽ)					

Table 3-10 Ge’ez and Amharic Subjective Suffix

In both language verbs can be inflected either by sound or by adding suffixes. As depicted Table 3-10, to inflect a verb in Ge’ez the first person and second person, the last character of a word changes its sound to **SADES** and add the suffix. Amharic first person and second person (**አንተ** and **አንቺ**) when verbs are inflected the last character of a verb is changes its sound to SADES and add the suffix.

Objective suffixes are those that are added in addition to subjective suffixes at the end of the verb to show the object’s Number, Gender, nearness or farness, either mentioned in script/Fidel/character or sound [10] [11]. Accordingly, verbs in Ge’ez and Amharic with primary Anketse inflected up to 80/96 and 50 respectively. To inflect verbs in both language using objective suffixes, pronouns in the same person can’t demonstrate with the same person pronouns. Table 3-11 below shows objective suffix using in Ge’ez አነ **(ቀተልኩ)** and in Amharic እኔ（**ገደልኩ)**.

Pronoun	Ge’ez	Subjective Suffixes	Objective suffixes	Pronoun	Amharic	Subjective Suffixes	Objective suffixes
አንተ	ቀተልኩከ	-ኩ	-ከ	አንተ	ገደልኩክ	ኩ	ክ
አንቲ	ቀተልኩኪ	-ኩ	-ኪ	አንቺ	ገደልኩሽ	ኩ	ሽ
ውእቱ	ቀተልከዎ	-ክ	-ዎ	እሱ	ገደልኩት	ኩ	
ይእቲ	ቀተልክዋ	-ክ	-ዋ	እሷ	ገደልኳት	ኳ	ት
አንትሙ	ቀተልኩክሙ	-ኩ	-ክሙ	እናንተ	ገደልኳችሁ	ኳ	ችሁ
አንትን	ቀተልኩክን	-ኩ	-ክን				
ውእቶሙ	ቀተልክዎሙ	-ክ	-ዎሙ	እነሱ	ገደልኳቸው	ኳ	ቸው
ውእቶን	ቀተልክዎን	-ክ	-ዎን				

Table 3-11 Ge’ez and Amharic Objective Suffix inflection

Objective Suffixes used to indicate the object whereas Subjective suffixes is to indicate subject [1]. Verbs that are inflected using subjective suffix only is called **YEWA** verb while inflected using both subjective and objective suffixes called **MESERI** verb. Singular and double styles are the two types for YEWA and MESERI verbs respectively.

<u>አስራው/Prefixes</u>

These are prefixes used to inflect the present and the future form of the verb in Ge'ez and Amharic [10]. For example, using the third person **ውእቱ/እሱ**/he. In Ge'ez, the use of prefixes **(አ)** for first person singular **(አነ)** and **(ነ)** for first person plural **(ንሕነ). (ተ)** for all second person pronouns **(አንተ፣ አንቲ፣ አንትሙ፣ አንትን)** and for feminine 3rd person singular **(ይእቲ).** Use () for all 3rd person pronouns except **(ይእቲ).** In Amharic use (አ) for all first-person pronouns **(እኔ፣ እኛ), (ተ)** for all 2nd person pronouns including 3rd person **(አንተ፣ አንቺ፣ እናንተ፣ እሷ)** and (የ) for all 3rd person except **(እሱ፣ እነርሱ).**

	Ge'ez					Amharic			
Pronoun	**Past**	**Present**	**Future**	**Command**	**Pronoun**	**Past**	**Present**	**Future**	**Command**
አነ	ቀተልኩ	እቀትል	እቅትል	እቅትል	**እኔ**	ገደልኹ	እገድል	እገድል ዘንድ	ልግደል
አንተ	ቀተልከ	ትቀትል	ትቅትል	ቅትል	**አንተ**	ገደልኽ	ትገድል	ትገድል ዘንድ	ግደል
አንቲ	ቀተልኪ	ትቀትሊ	ቀተልኪ	ቅትሊ	**አንቺ**	ገደልሽ	ትገድሊ	ትገድሊ ዘንድ	ግደዪ
ውእቱ	ቀተለ	ይቀትል	ይቅትል	ይቅትል	**እሱ**	ገደለ	ይገድል	ይገድል ዘንድ	ይግደለ
ይእቲ	ቀተለት	ትቀትል	ትቅትል	ትቅትል	**እሷ**	ገደለች	ትገድል	ትገድል ዘንድ	ትግደል
ንሕነ	ቀተልነ	ንቀትል	ንቅትል	ንቅትል	**እኛ**	ገደልን	እንገድል	እንገድል ዘንድ	እንግደል
አንትሙ	ቀተልክሙ	ትቀትሉ	ትቅትሉ	ቅትሉ	**እናንተ**	ገደላችኹ	ትገድሉ	ትገድሉ ዘንድ	ግደሉ
አንትን	ቀተልክን	ትቀትላ	ትቅትላ	ቅትላ					
ውእቶሙ	ቀተሉ	ይቀትሉ	ይቅትሉ	ይቅትሉ	**እነሱ**	ገደሉ	ይግደሉ	ይገድሉ ዘንድ	ይግደሉ
ውእቶን	ቀተላ	ይቀትላ	ይቅትላ	ይቅትላ					

Table 3-12 Amharic and Ge'ez Prefixes to show perfect tense

As described in the Table 3-12, it shows the infliction of verbs **ቀተለ** and **ገደለ** for Ge'ez and Amharic verbs respectively. In Ge'ez language **ውእቱ** is the third person singular male gender indictor whereas **ውእቶሙ** is third person plural and male gender indicator. In the same way **ይእቲ** is the third person singular female gender indicator whereas **ውእቶን** is a third person plural female gender indicator. In case of Amharic language, when we translate Ge'ez pronoun to Amharic **ውእቶሙ** and **ውእቶን**, in case of both gender is **እነሱ.** In Ge'ez the pronoun **አንትሙ** and **አንትን** is the plural form of pronoun **አንተ and አንቲ** respectively. When they are translated in to Amharic **(አንትሙ** and **አንትን)** they take "**እናንተ**" meaning of pronoun.

According to Desta [7] and Yitayal [6] the prefixes morphological analysis of Ge'ez Verbs. They identify lists of subjective markers prefixes that include **-ክ**/-kä/, **-ክሙ**/-kmu/, **-ኪ**/-ki/, **-ክኑ**/-kn/, **-አ**/-ä/, **-ኡ**/-u/, **-አት**/-ät/, **-ኣ**/-a/, **የ**/-yä/, **-ኦ**/-o/, **-ሙ**/-mu/, **-ን**/-n/, **-እ**/-ə/, **-ኡ**/-u/ and **-ኢ**/-i/. and also list of Ge'ez verbs prefixes **አ**-/ä-/, **እ**-/ə-/, **አስተ**-/ästä-/, **ኢ**-/i-/, **ና**-/na-/, **አስተ**-/nastä-/, **ኑ**-/n-/, **ነት**-/nt-/, **ታ**-/ta-/, **ታስተ**-/tastä-/ , **ተ**-/tä-/, **ት**-/t-/, **ትት**-/tt-/, **ያ**-/ya-/, **ያስተ**-/yastä-/, **ይ**-/y-/, and **ይት**-/yt-/. The objective markers suffixes includes the following **-ኒ**/-ki/, **-ን**/-nä/, **-ክሙ**/-kä/, **-ክ**/-kä/, **-ክሙ**/-kmu/, **-ኪ**/-ki/, **-ክኑ**/-kn/, **-ክዎ**/-kwo/, **-ሁ**/-hu/, **-ኦ**/-o/, **-ዎ**/-wo/, **-የ**/-yo/, **-ዎሙ**/womu/, **-ክሙ**/kmu/, **-ሙ**/-mu/, **-ሆሙ**/-homu/, **-የሙ**/-yomu/, **-ዋ**/-wa/, **-ሃ**-ha/, **-ኣ**/-a/, **-ያ**/-ya/ , **-ዎኑ**/-won/, **-ሆኑ**/-hon/, **-ኑ**/-n/, and **-የ'ን**/-yon/.

3.5.1.4. Adverb (ተውሳከ ግስ)

It is a word used to describe the property of a verb. In Ge'ez and Amharic there are different types of adverb [11] [10]. Adverbs position in Amharic is always used before the verb it describes while in Ge'ez the position of the adverb is before and after the verb just like that of the adverb.

3.5.1.5. The Stems of Verbs (አዕማደ ግስ)

According to Ethiopian scholars [63] [10], stems are also called **አዕማደ** /ʔəʕimad/ 'pillars', or shaft that support the roof of building. They are pillars or bases of verbs that support the conjugations of verbs. These scholars believe that Ge'ez [14] [16] [11] and Amharic [10] have five stem patterns which all are independent of each other.

- ✓ Perfective stems ገቢር/ማድረግ ዐምድ ,
- ✓ Causative stems አገብሮ/ማስደረግ ዐምድ,
- ✓ Causative-reciprocal stems አስተጋብሮ/አደራራጊ ዐምድ
- ✓ Reflexive stems ተገብሮ/መደረግ ዐምድ
- ✓ Reciprocal stems ተጋብሮ/መደራረግ ዐምድ

Each of the above pillars of verbs in Ge'ez and Amharic have **prefixes** with clear example shown in table 3-13 below using words **ቀተለ** and **ገደል**

No	**Stems of Verbs (አዕማደ ግስ)**	**Ge'ez**	**Amharic**
1	Perfective stems **ገቢር/ማድረግ ዐምድ**	ቀተለ	ገደለ
2	Causative stems አገብሮ/ማስደረግ ዐምድ	**አቅተለ**	**አስግደለ**
3	Causative-reciprocal stems አስተጋብሮ/አደራራጊ ዐምድ	**አስተቃተለ**	**አገዳደለ**
4	Reflexive stems ተገብሮ/መደረግ ዐምድ	**ተቀትለ**	**ተገደለ**
5	Reciprocal stems ተጋብሮ/መደራረግ ዐምድ	**ተቃተለ**	**ተጋደለ**

Table 3-13 stems of verbs of Ge'ez and Amharic

3.5.2. Minor Parts of Speech

3.5.2.1. Pronoun (መራሕያን/ተውላጠ ስም)

A **pronoun** is a word that substitutes for a noun or noun phrase. Unlike other languages Ge'ez have 10 pronouns [11] [14], Amharic 9 [10] and English 7 [8]. Pronouns (**መራሕያን**) in both langauge can be used as instead of noun, verb to be and adjectives. They are the main component to not only to understand but also to indicate direction how to use the language. In both there are different types of pronoun namely personal, reflexive, relative, reciprocal, demonstrative, interrogative, indefinite, and possessive pronoun. When word is inflected or derived using pronoun.

Personal (ምድብ ተውላጠ ስም) Pronoun

In Ge'ez and Amharic pronouns can be classified as singular and plural, masculine and feminine, and near and far. These are shown in the table 3-14 below:

	Pronouns (መራሕያን/ ተውላጠ ስም)			ጾታ			የቁጥር መጠን	
	ግእዝ	አማርኛ	English	ተባዕታይ/ወንድ/ Masculine	አንስታይ/ሴት/ Feminine	የወል common gender	ብዙ/ plural	ነጠላ/ singular
1st Person (ቀዳማይ/አንደኛ) መደብ	አነ	እኔ	I			✓		✓
	ንሕነ	እኛ	We			✓	✓	
2nd Person (ካልኣይ/ሁለተኛ) መደብ	አንተ	አንተ	You	✓				✓
	አንቲ	አንቺ			✓			✓
	አንትሙ	እናንተ		✓			✓	
	አንትን				✓			
3rd Person (ሣልሳይ/ሦስተኛ) መደብ	ውእቱ	እሱ	He/It					✓
	ይእቲ	እሷ	She/It		✓			✓
	ውእቶሙ	እነሱ	They				✓	
	ውእቶን	እርሶ ወይም አንቱ			✓		✓	

Table 3-14 Ge'ez and Amharic Pronouns

We cannot talk about grammar without pronoun. Since a pronoun tells about category of person (1st, 2nd and 3rd), gender (Masculine and Feminine) and place (near and far). As you see from the Table 3-14 in Amharic **እናንተ** used to as to express both Masculine and Feminine pronouns in 2nd person pronoun and **እርሶ ወይም አንቱ** used to express our respect to those are older than the speaker, due to this it is called respect pronoun.

In Ge'ez, for each Gender in 2nd and 3rd personal pronoun plural form each have Masculine and feminine form. In Amharic the plural form of '**አንተ**' and '**አንቺ**', '**እሱ**' and '**እሷ**' are '**እናንተ**' and '**እነርሱ**' respectively for both Masculine and feminine.

Pronoun in Ge'ez and Amharic can be used being Subject in leading the sentence as singular and plural, near and far, and Masculine and feminine. Example : As singular **አነ መጻእኩ → እኔ መጣሁ →** came and plural **ንሕነ መጻእነ → እኛ መጣን።።** we came. As near **አነ ሀለውኩ → እኔ አለሁ →** I existed and far **አንተ ሀለውከ → አንተ አለህ →** you existed. **ውእቱ ሀለወ → እሱ አለ →** He Existed. As Masculine **አንተ መጻእከ → አንተ መጣህ →** He came. And feminine **ይእቲ መጻአት → እሷ መጣች →** she came. When they act like a pronoun, they indicate morphology type of the noun. For example, in table 3-15 using **ቀደሰ/አመሰገነ/ ቀዳማይ አንቀጽ** (Past Tense) depicted.

N o	Ge'ez	Amharic
1	አነ ቀደስኩ	እኔ አመሰገንኩ
2	አንተ ቀደስከ	አንተ አመሰገንክ
3	አንቲ ቀደስኪ	አንቺ አመሰገንሽ
4	ውእቱ ቀደሰ (ግዕዝ ድምፅ)	እሱ አመሰገነ(ግዕዝ ድምፅ)
5	ይእቲ ቀደሰት	እሷ አመሰገነች
6	ንሕነ ቀደስነ	እኛ አመሰገንን
7	አንትሙ ቀደስክሙ	እናንተ አመሰገናችሁ
8	አንትን ቀደስክን	
9	ውእቶሙ ቀደሱ (ካዕብ ድምፅ)	እነሱ አመሰገኑ (ካዕብ ድምፅ)
10	ውእቶን ቀደሳ (ራብዕ ድምፅ)	

Table 3-15 Ge'ez and Amharic suffix

3.5.2.2. Demonstratives (አመልካች)

A demonstrative pronoun stands in for a person, place or thing and can function as a subject, an object or an object of the preposition. It is used before a verb of the sentence not before a noun. In Ge'ez and Amharic, the following pronouns exists.

Near					Far				
Gender	Singular		Plural		Gender	Singular		plural	
	Ge'ez	Amharic	Ge'ez	Amharic		Ge'ez	Amharic	Ge'ez	Amharic
Male	ዝንቱ /ዝ/	ይህ፣ ይህ ነው ፣ ይህውና፣ይህው This	እሎ/እሉ /እሎንቱ/	እነዚህ These	Male	ዝኩ፣ ውእቱ ዝክቱ ፣ ዝስኩ	ያ፣ያውና፣ ያነው That	ውእቶሙ፣ እልኩ፣እልክቱ	እነዚያ ፣ እነዚያው ፣ እነዚያ ናቸው Those
Female	ዛቲ/ዛ/	ይቺ፣ ይቺው፣ ይቺውና፣ ይቺ ናት This	እሎን፣እላ፣ እላንቱ		Female	ይእቲ፣ እንታክቲ ፣ እንትኩ	ያቺ፣ ያቺውና፣ ያቺው That	ውእቶን፣ እልኮን፣ እልክቶን	

Table 3-16 Demonstrative Pronoun in Ge'ez and Amharic

For Example, **ዝንቱ** ውእቱ ወልድየ፡፡ **ይህ** ልጄ ነው፡፡

3.5.2.3. Possessives (አገናዛቢዎች)

Possessive pronouns are words that demonstrate ownership or possession. Possessive pronouns show that something belongs to someone or something. In Ge'ez and Amharic the following are possessive markers or suffixes. The table below show possessive suffixes of Ge'ez and Amharic with example.

	Singular			Plural		
	Ge'ez	Amharic	English	Ge'ez	Amharic	English
1st Person	ዚአየ	የእኔ	mine	ዚአነ	የእኛ	ours
2nd Person	ዚአከ	የአንተ	Yours	ዚአክሙ	የእናንተ	Yours
	ዚአኪ	የአንቺ		ዚአክን		
3rd Person	ዚአሁ	የእሱ	his	ዚአሆሙ	የእነሱ	theirs
	ዚአሃ	የእሷ	hers	ዚአሆን		

Table 3-17 possessive pronoun in Ge'ez and Amharic

When Ge'ez pronouns are used as **verb to be** each pronoun express their own meaning as translated into Amharic. For example, look at **ወእቱ**: - when it is used in a sentence it may have one of the following meaning in Amharic. It may be **ነው ፣ ኸነ ፣ ነበር ፣ ኖረ ፣ ይኑር ፣ ነሽ ፣ ነኝ ፣ ነኸ ፣ ናችኸ ፣ ነበራችኸ፡፡**

ሕዝቅኤል ነብይ ውእቱ፡፡

ሕዝቅኤል ነብይ ነው፡፡

In the same way other shows the following.

Pronoun	Meaning When Translated into Amharic
ይእቲ	ናት፣ነበረች
ውእቶሙ	ናቸው፣ኾኑ፣ነበሩ፣ኖሩ፣ይኑሩ
ውእቶን	ናቸው፣ነበሩ
አንተ	ነህ፣ነበርክ፣ኾንክ፣ኖርክ፣ኑር
አንቲ	ነሽ፣ነበርሽ፣ኾንሽ፣ኑሪ
አንትሙ	ናችኹ፣ሆናችኹ፣ነበራችኹ፣ኑሩ
አንትን	ናችኹ፣ሆናችኹ፣ነበራችኹ፣ኑሩ
ንሕነ	ነን፣ኾንን፣ነበርን፣ኖርን፣እንኑር
አነ	ነኝ፣ኾንኩ፣ነበርኩ፣ልኑር

Table 3-18 meaning of Ge'ez pronouns when use as verb to be

3.5.2.4. Conjunction (አያያዥ)

Conjunction is a word used to connect clauses or sentences or to coordinate words in the same clause. In Ge'ez ወ ፤ አው ፤ and ዳዕሙ ፣ አላ ፣ ባሕቱ and in Amharic እና ፣ ወይም ፣ and ነገር ግን are conjunction used. ወ in Ge'ez has 27 meaning. The most commonly used meaning of ወ used as እና ፣ ም. When it conjugates two name it has meaning of እና/ና. For example,

ማርያም ወማርታ መጹ።።

ማርያምና ማርታ መጡ።።

When it conjugates three names it has a meaning of ም. For example,

አብርሃም ወይስሐቅ ወያዕቆብ።።

አብርሃም እና ይስሐቅ ያዕቆብም።።

The other conjunction pronoun is አው meaning ወይም. For example,

አመተ ማርያም አወ አብዲሳ።።

አመተ ማርያም ወይም አብዲሳ።።

3.5.2.5. Punctuation Marks

In Ge'ez there is no question mark whereas Amharic has. The interrogative is placed at the end of a word or the sentences. It is pronounced with a low level and the style of pronunciation by itself also shows an interrogation. In most cases, Ge'ez interrogatives are preceded by a radical which has the same order to the interrogative. For example, ሁ፣ ኑ፣ ኡ፣ ኢ፣ ት፣ ኣ፣ ኣይት? ሶበኑ? (When?) ፣ ተአምሩኑ? (Do you know?) ፣ አንትሙሁ (are you?) ፣ ተአምረኒኢ (do you know me?).

Sometimes, Ge'ez interrogatives are placed with possession or definite articles. For example, ወርቁኑ "is that his gold?" or "is he/it the Gold?" ቤተ መቅደስኑ? (By the temple?

3.6. Morphology

Morphologically, languages are often characterized along two dimensions of variation. The first is the number of morphemes per word, ranging from isolating languages in which each word generally has one morpheme, to polysynthetic languages in which a single word may have very many morphemes.

The second dimension is the degree to which morphemes are segmentable, ranging from **agglutinative languages** like Turkish have relatively clean boundaries, to **fusion languages** like Russian, in which a single affix may conflate multiple morphemes, like -om in the word stolom,

(table-SG-INSTRDECL1) which fuses the distinct morphological categories instrumental, singular, and first declension.

Both languages are fusion languages and the number of morphemes for a single word is one or greater than one. Beside this Ge'ez and Amharic exhibit such character that the performance of the SMT system difficult. Inflectional morphemes include the grammatical functions of the word. These are number, tense/aspects, possession and comparison [11].

Number: - Ge'ez and Amharic has singular and plural numbers. The number marker in Ge'ez and Amharic usually exists noun, adjectives, and verb conjunctions. It exists in either of prefix, infix, suffix and super-fix. The number markers in pronouns, demonstratives, prepositions are the same but numbers in nouns are complex with exception of every conjunction. In Ge'ez, *-yan, -an, yat,* and *-at* are suffix plural number marks in Ge'ez.

Gender: -in Ge'ez the gender markers are not limited. They may vary from time to time accordingly to the part of speech. The gender markers are the feminine markers. Gender is distinguishable in both singular and plural. Gender is nouns, adjectives, some adverbs, prepositions, demonstratives, possessive, verbs are marked by the following -ኣት- at plural, -ተ - as person profile as personal suffix, -ን - in pronoun plural, -ኢ- in pronoun possessive, and aspect… ሃ - as objective markers in personal names in possession preposition. ኣ -in gerund, infinitive and derivational morphemes

3.7. Challenges of Ge'ez and Amharic during machine transaltion

There are different challenges observed in a machine transaltion system between Ge'ez and Amahric. The major once are accessed below.

Morphological challenges

Translating between two morphologically rich languages poses challenges in analysis, transfer and generation. The complex morphology induces an inherent data scarcity problem, and the limitation imposed by the dearth of available parallel corpora is magnified. Both Ge'ez and Amharic are ploy syntactic languages which is the number of morphemes per word is not always one. Most of the research conducted in SMT are using morphologically rich language as a source language and target language is morphologically poor. The performance scored was good, this due to one to many alignments between source and target language.

Syntactic challenges

Syntactically, languages are perhaps most saliently different in the basic word order of verbs, subjects, and objects in simple declarative clauses. Amharic syntactically sentence is organized using SOV word order. Which implies the object come after the object and before the verb. Where Ge'ez follows **SVO** or **VSO** and **OVS**. This makes the translation most challenging. Beside the corpus is organized in either of it of mixed word order. Therefore, syntax is another challenge.

Alignment Challenge

Alignment is also another challenge, which plays a critical role in statistical machine translation. Alignment is critically being a challenge if SMT is conducted between two morphologically rich languages. Different types of alignment exist in a sentence namely one to one, one to many, many to one and many to many see figure 3-1.

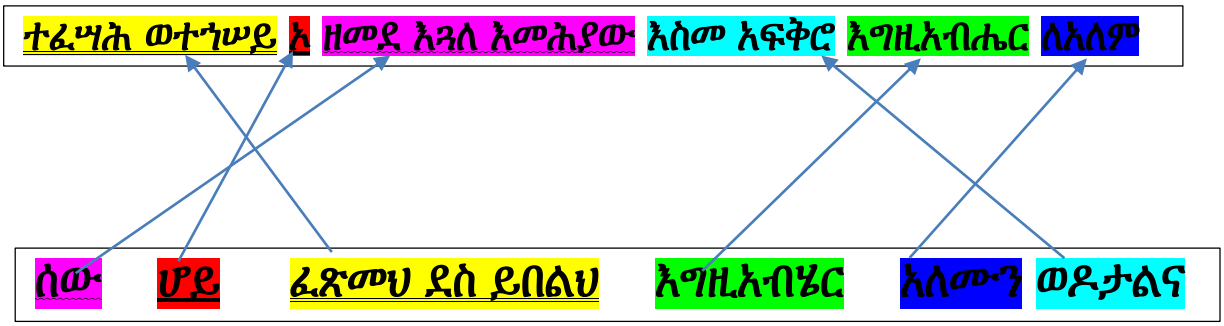


Figure 3-1 Alignments of Amharic and Ge'ez sentence

Such challenges needs to be given attention during applying machine transaltion from source language to target language. Specially, for morphologically rich languages like Amharic and Ge'ez the process of is more serious. To control the large morphological variation morpheme-based machine transaltion is experimented in this study for Amharic and Ge'ez.

Chapter Four

Design and Experimentation

The main objective of this research is, to develop a morpheme-based bidirectional Ge'ez- Amharic machine translation. Therefore, we design an architecture of a bidirectional machine translation for Geez-Amharic. To conduct the experiment we prepare dataset, preprocess, apply morphological segmentation, construct language and translation model.

4.1. Architecture of the prototype

This section is about the prototype of the system starting from input corpus until the translation output and the activities performed at each stage. As describe earlier corpus is collected from online sources, manually prepared as well as adapted ones passes through preprocessing tasks such as tokenization and Normalization. To design bi-directional Ge'ez-Amharic machine transaltion an architecture depicted in figure 4-1is followed.

The architecture works through the following processes. First input corpus goes to preprocessing which includes tokenization and normalization. Then the preprocessed dataset divided into monolingual and bilingual dataset. Monolingual dataset includes either word or morpheme-based which is further processed to build Language Model through language molding. Bilingual dataset is a two files for each transaltion unit in each language. It is used to build the transaltion model via transaltion modeling.

Input Corpus

Input corpus is a corpus that is fundamental for starting the transaltion process. As describe above the unit of transaltion used for this research is word and morpheme. Based on each transaltion unit, we have prepared the input corpus. The word based dataset is a base line for the next transaltion unit which is morpheme. This means input corpus contains two files for each transaltion unit word and morpheme. The morpheme based aligned sentences were prepared using morfessor and rule based. As per this research, two experiment were conducted each time. A total of six datasets were prepared. For word-based translation, two datasets were used for bi-directional Ge'ez and Amharic translation. For morpheme-based MT four datasets were prepared based on unsupervised and rule-based segmentation. Two bilingual files for each techniques were prepared. A total of 13,833 sentence level aligned files were prepared for each transaltion unit.

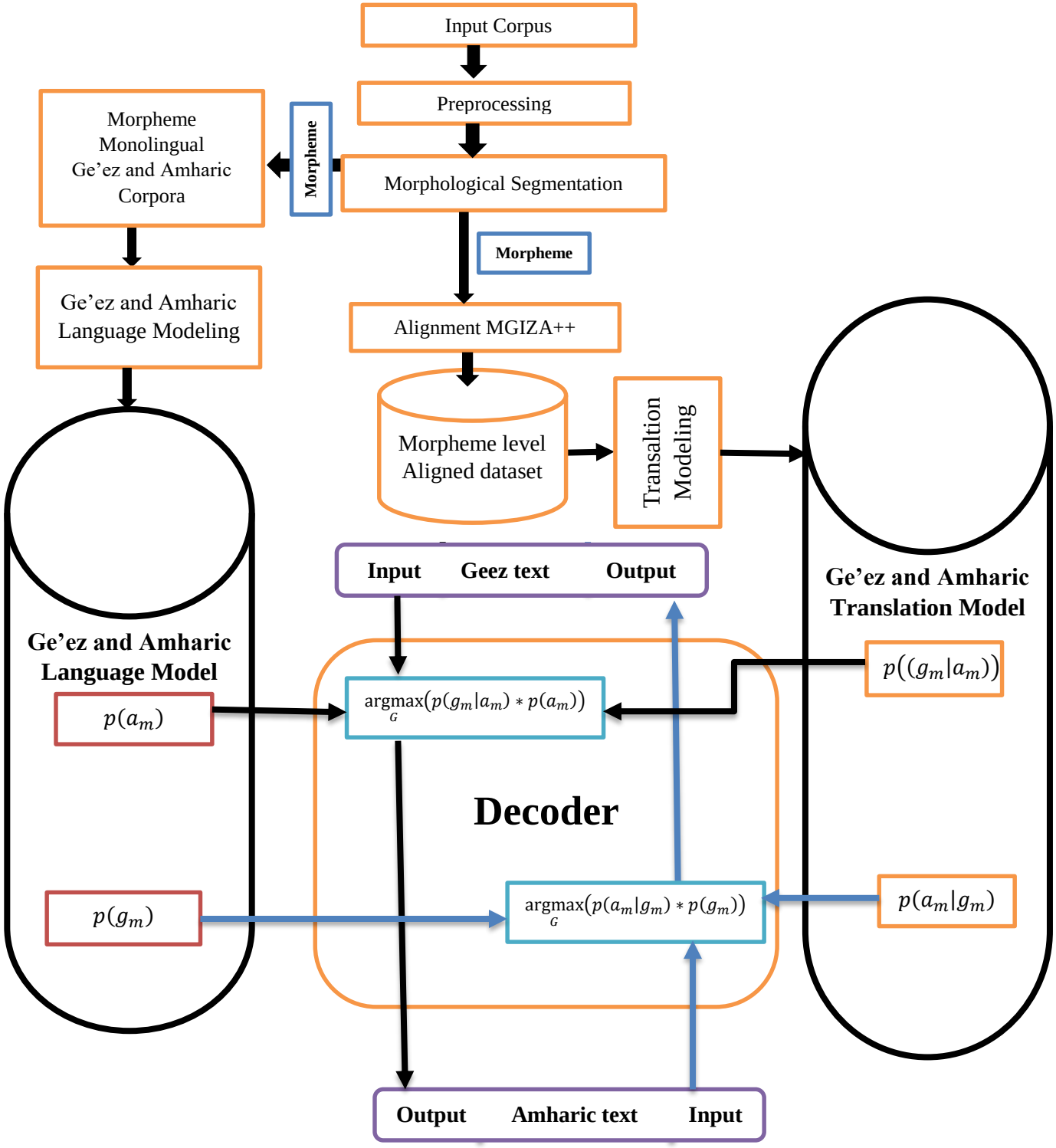


Figure 4-1 Architecture of Bi-Directional Ge'ez-Amharic Transaltion where $\boldsymbol{g_m}$ Ge′ez Morphmes and $\boldsymbol{a_m}$ Amharic Morphmes

Preprocessing

The preprocessing here includes character normalization and space normalization for Amharic and space normalization for Ge'ez language of each sets of training, tuning and test sets. Tokenization and clean also another preprocessing techniques used.

The final output of the preprocessing modules is bilingual corpus for alignment using MGIZA++ to build the transaltion model and monolingual corpus were used for building language model.

In Ethiopic writing there are characters with similar sound and meaning. In Ge'ez the variant characters have different sounds and meaning. For example the use of character "**አ**" and "**ዓ**" in the word **"ሰአሊ"** means **to beg** and **"ሰዓሊ"** mean **to** draw. But, in Amharic words, **"ሰአሊ"** and **"ሰዓሊ"** or **"እግዚአብሔር"** and **"እግዚኣብሄር"** are the same in Amharic meaning draw and God. So, Amharic corpus needs normalization If such words exist, the system consider as one word.

Language Model

For the language model we used monolingual corpora, which is automatically generated by combining the train and tune set. 12, 450 simple and complex sentences are used for both Amharic and Ge'ez language modeling. It is the same amount used for both word-based and morpheme-based MT.

Translation Model

MGIZA++ is used for both word and morpheme level aligned corpus for the translation model. MGIZA++ align the prepared corpus at word and morpheme level by using IBM models (1-5). The result of the output has been used for training and testing the system.

Decoder

A decoder searches for the best sequence of transformations that translates source sentence to the corresponding target sentence. It looks up all translations of every source word or phrase, using word or phrase translation table and recombine the target language phrases that maximizes the translation model probability multiplied by the language model probability. By following the above procedure, the decoder performs the translation process from both directions.

To evaluate the performance of the prototype, BLEU score is computed, which compares the translated document by the system with human translated document (reference translation).

4.2. Dataset Preparation

Two types of dataset prepared for word and morpheme based experimentation.

4.2.1. Dataset Source

For this research, the dataset or corpus is collected from different online sources which includes https://www.ethiopicbible.com, http://ethiopianorthodox.org, and http://eotcmk.org which contains parallel text of Ge'ez and Amharic. All sources of the data were related to Holy Bible of which Old Testament which includes Genesis, Exodus, Leviticus, Numbers, Deuteronomy, Joshua, Ruth, psalms, judge, I and II Samuel, and I Kings. We get these books by crawling from https://www.ethiopicbible.com. To download we were using Beautiful Soup, which is a Python library for pulling data out of HTML and XML files [64]. A total of 66 books were found, 12 of them parallel text of Ge'ez and Amharic. **Appendix II** show the Code Used for crawling the dataset of Ge'ez and Amharic. We also got bitext such as anaphora of Saint John Chrysostom, Saint Epiphanius, and Saint Athanasios from https://www.ethiopianorthodox.org in PPT format. The rest of the bitext which includes seven days Wedase Marya, Anketse Berhan, yewedesewa melahekete, Kidan and Liton were manually prepared bitext.

In preparing the parallel corpus we followed bottom up approach which means align first each book verse level, second merge the aligned books and finally merge all the books to the respective languages. For preprocessing activity, we use python and shell scripting.

For downloading the corpus we used a python script with the BeautifulSoup indicated in Appendix II, since the number of files for both language Amharic is 1187 and that Geez is 463 which is not equal, we write a python script below, to automatically delete/remove files of Amharic which is not found in Ge'ez files.

```
Import os
am = os.listdir ('/home/tadesse/PycharmProjects/ggg/amharic/books/')
retained_file = os.listdir ('/home/tadesse/PycharmProjects/ggg/geez/books/')
for i in am:
    if i not in retained_file:
        os.remove ('/home/tadesse/PycharmProjects/ggg/amharic/books/' + i)
```

4.2.2. Dataset Preprocessing

There are different challenges observed in the collected data. Such challenges are preprocessed by using the dataset for experimentation.

In preparing the dataset for the experimentation the following are main challenges:

1. Misalignment of sentence verse which means the verse exists but it is place in wrong place or placed in another verse in the chapter or in another chapter. For example look መዝሙረ ዳዊት ምዕራፍ ፩ or chapter one of Psalms of verse three in Amharic, when it is translated, they translate with two verse. Due to this the next verse of one of the language encountered miss-transaltion. Therefore; to solve such challenge, we manually check each books verse by verse with the help of professionals. Misplacement of translated verse of the whole chapter or parts of a chapter. Such problems also solved manually.
2. Different ways of writing numerals, for example in Ge'ez ፪፻፸፩ and **ክልኤቱ ምዕት ወአሐዱ.** Dealing with non-standard numerals is also another challenge. For our experiment non-standard numerals are converted to their textual representation manually.
3. Duplication of a single verses in both language Amharic and Ge'ez. We removed redundancies using python script (see Appendix VIII).
4. Mistranslation of the whole chapter in both languages. For example, zelewaweyan 37, 38, 39 and 40. Mistranslation of numerals in both language for example in the ኦሪት ዘኍልቍ ምዕራፍ ፩ verse 28 **"፷፻፸፬፻"** in Ge'ez when it is translated to Amharic it look like this **"ሰባ አራት ሺህ ስድስት መቶ"** **፷፻፸፬፻"** meaning **fifty-four thousand and four hundred** which is different from that of **Amharic.**

As you see from the above listed challenges which makes the dataset preparation difficult we tried to find solution in to two ways. Using manually and automatically solving the difficulty. Due to the above challenges and no pattern exists we forced to check manually each verse of each files of each languages. To solve Challenge 4 we write script.

Most of the Ge'ez dataset have higher verse numbers compared with Amharic dataset. When we merge or rearranging of the verse of Ge'ez dataset manually, it gives the full meaning of the Amharic data set in each books dataset. Due to this by looking on each books chapter we arrange manually.

During dataset preparation especially in Old Testament the following basic activities are performed. As listed above with all the challenges, we write a script aligned and conduct

experiment. We aligned sentences in both languages verse level and those that are manually prepared datasets are aligned with the help of professionals and graduates from Saint Trinity college of EOTC in diploma and degree as well as traditional school teachers in checking the alignment. Misplaced verse, phrases and sentences are corrected.

We expand Ge'ez numbers into expected word manually with the help of professionals due to the way they are written and translated not correct. Challenges indicated in numbers 3 and 7 we encountered most of these problems in Ge'ez dataset. In Ge'ez dataset these problems were happened in to two ways. The first one is simple writing the number even if there is mistranslation indicated in challenge 6. The second one is mixed ways of writing numerals which is simply writing the number and using expanded form of writing numbers style.

Another challenges is misplacement of verse in the chapter of the book and in another books. While making the dataset ready for experimentation we used different python script, for removing verse number (Appendix III) and removing duplicated verses. Manually prepared dataset were aligned automatically.

By conducting this it is possible to prepare better size of corpus for sentence alignment. Word and morphemes are important for the objective. We have used the prepared corpus for word level alignment and MGIZA++ align the corpus using IBM model 1-5. The general steps for base line dataset preparation described in Figure 4-2.

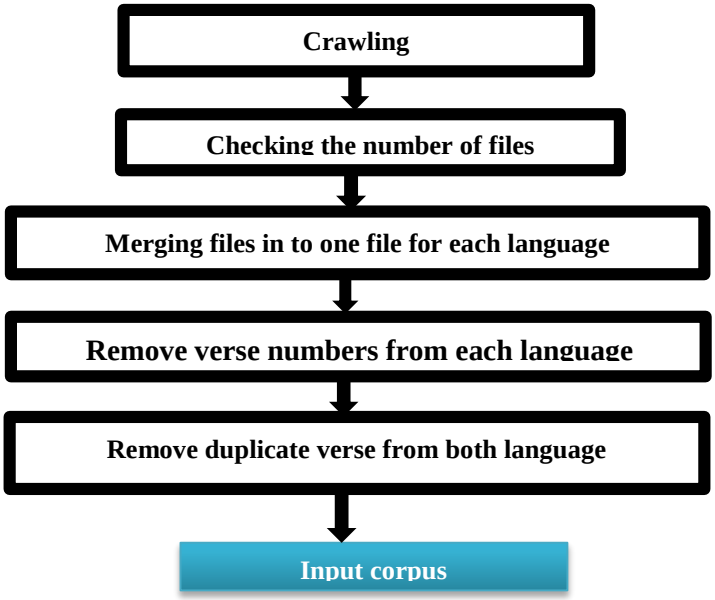


Figure 4-2 Data set Preparation steps for Base line experiment for word based transaltion

Each steps of the process were conducted using python script.

4.2.3. Morpheme-based Dataset preparation

Unsupervised morpheme segmentation

The same corpus has been used for morpheme-based translation. But dataset preparation for morpheme-based translation is different from that of word-based, for this research done using unsupervised segmentation tool called morfessor.

During segmentation morfessor follows the following procedure [49] [31]. The first step is to create a model for both corpus using **morfessor script** using training and test data set. Then model is used to segment an input corpus. Using the created model and morfessor-segment script, text corpus as an input for both language redirect to new file. The third step is to reassemble the segmented new file text using python script. The fourth steps are to provide the morpheme aligned sentences to the MGIZA++.

For the purpose of this research we adapt the standard workflow for Morfessor command line tools which is adapted from **[31]** for segmenting the input corpus as shown below in figure 4-3.

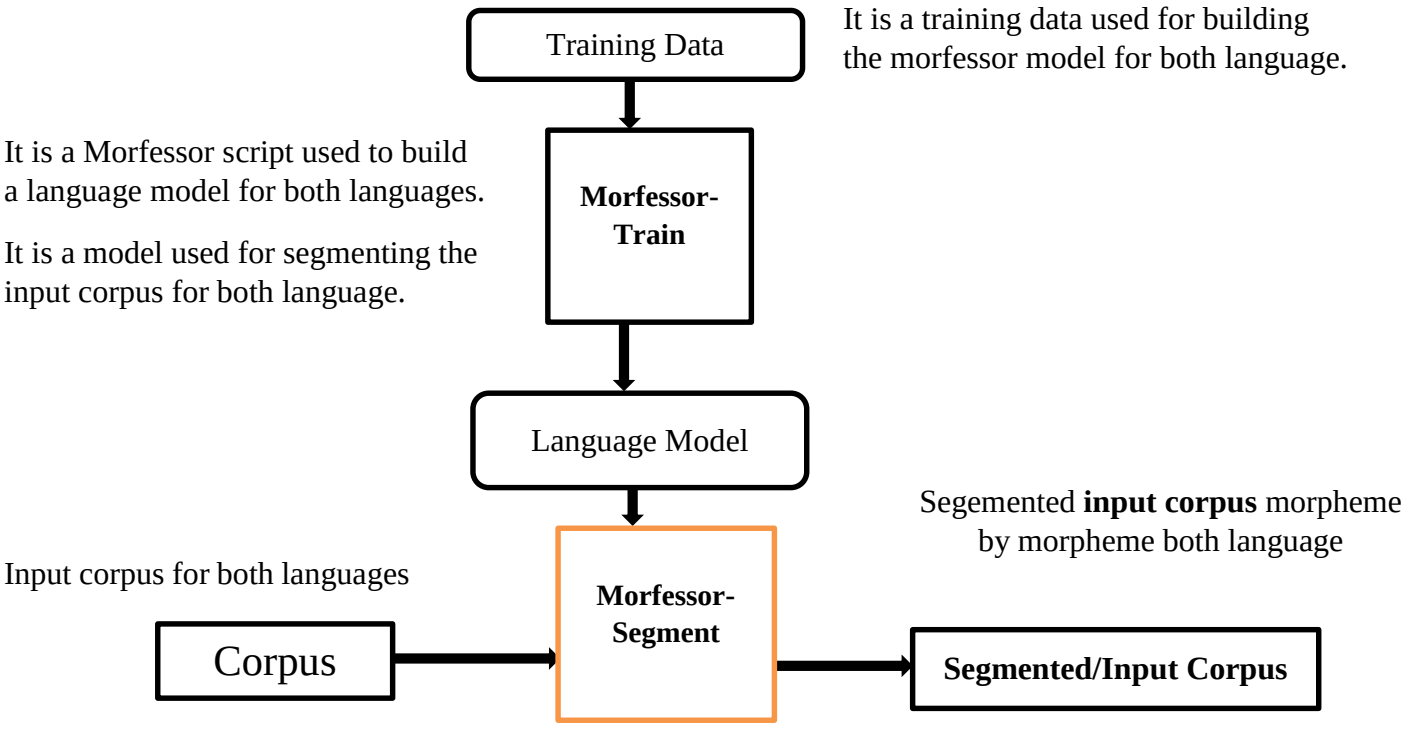


Figure 4-3 Morfessor segmentation processes

As depicted in the above Figure 4-3 the given training data is submitted for morfessor to create the model, then using the model segment the input corpus and redirect to new file. Finally using python script in Appendix IV merge the segmented corpus into sentence level.

For Ge'ez corpus to create the model using training and segment input corpus we use the following syntax

a. **morfessor-train** ginputtext.txt -S geez_model.segm geez_test.txt and
b. **morfessor-segment** -L geez_model.segm geez_corpus.txt > geez.txt-segmented

Accordingly, for Amharic corpus to create the model and to segment input corpus syntax.

c. **morfessor-train** ainputtext.txt -S amharic_model.segm amharic_test.txt
d. **morfessor-segment** -L amharic_model.segm amahric_corpus.txt > amahric.txt-segmented

Ge'ez	Amharic
አኰቴተ	የ ዮሐንስ
ቍርባን	አ ፈ
ዘ ዮሐንስ	ወርቅ
አፈ ወርቅ	የ ቍር ባ ን
ጸ ሎቱ	ምስጋና
ወ በረከቱ	ጸሎቱ ና
የሀሉ	በረከቱ
ምስለ	ከ እኛ
ኵልነ	ጋራ
ሕዝበ	ይ ኑር
ክርስትያን	ለ ዘላለሙ
ለ ዓለመ	አሜን
ዓለም	።
አሜን	ጎሊ ናችሁ
።	ወደ
ላዕለ	ላይ
ይኩን	ይሁን
ጎሊና ክሙ	ልቡና ችሁ
በ ሰማይ	በ ሰማይ
የሀሉ	ይ ኑር
ልብከሙ	የምት ቆሙ በትን
አእምሩ	ዕወቁ
ጎበ	የ ጽድቅ ንም
ዘ ትቀውሙ	ቃል
ወ ስምዐ	ስሙ
ቃለ	በነ
ጽድቅ	ነገር ንም
ወ አጽምዐ	አድምጡ
ዜና	።
ሠናየ	
።	

Table 4-1 sample morpheme generated for Ge'ez and Amharic

As you can see from Table 4-1each words are represented with morpheme including prefixes and suffixes. It is not possible to translate the list of words row wise before concatenating them at sentence level.

To this end we write a python script shown in **Appendix IV** to merge segemented words into sentence level in to two files. By merging the lists of words in table 4-1 the following sentence formed for Ge'ez

አኰቴተ ቍርባን ዘ ዮሐንስ አፈ ወርቅ ጸ ሎቱ ወ በረከቱ የሀሉ ምስለ ኵልነ ሕዝበ ክርስትያን ለ ዓለመ ዓለም አሜን ።
ላዕለ ይኩን ኅሊና ክሙ በ ሰማይ የሀሉ ልብክሙ አእምሩ ኅበ ዘ ትቀውሙ ወ ስምዑ ቃለ ጽድቅ ወ አጽምዑ ዜና ሠናየ ።
In the same way, for Amharic also the following sentence constructed

የ ዮሐንስ አ ፈ ወርቅ የ ቍር ባ ን ምስጋና ጸሎቱ ና በረከቱ ከ እኛ ጋራ ይ ኑር ለ ዘላለሙ አሜን ።
ኅሊ ናችሁ ወደ ላይ ይሁን ልቡና ችሁ በ ሰማይ ይ ኑር የምት ቆሙ በትን ዕወቁ የ ጽድቅ ንም ቃል ስሙ በነ ነገር ንም አድምጡ።

Evaluation of Segementation

	Ge'ez	Amahric
Total number of unique segemented morphemes	28, 826	31,739
10 % for testing for evaltuation of segementation	2,882	3, 173
Evaluation of Morefesor segemtnation	**56.21%**	**52.44%**

Table 4-2 Evaluation of unsupervised morphmes segmentation for Ge'ez and Amharic lamguage using morefessor

As depicted in table 4-2 the total number of unique morphmes for Ge'ez and Amharic is **28,826** and **31, 739** respectivelly. For evaluating of unsupervised morpheme segmenttion **10%** of the total where selected randaomlly. **10%** is also segemnted manually for comparing it with the morefessor output. The evaluation performance shows that **56.21%** and **52.44%** for Ge'ez and Amharic language respectively.

Rule-based Morpheme segmentation

Morphological segmentation is recognized as a potential solution in statistical machine translation (SMT) to deal with data sparsely posed by morphologically complex languages like all Uralic languages [65]. Due the morfessor is perform the segmentation based on corpus size, which is unsupervised, as the corpus size increase the segmentation becomes correct.

For the purpose of this research we used Ge'ez and Amharic both as source and target language. To prepare the dataset for Statistical Machine Translation process we were using python scripting for basic rules for prefixes and suffixes.

The reason for poor performance of the SMT includes **under** and **over segmentation.** Therefore the next two experiments based on rule-based segmentation using basic affixes (prefixes and suffixes) for both language. The prefixes (prefix1, prefix2, prefix3 and prefix4) and the suffixes (suffix, suffix2, suffix3 and suffix4) lists were found from the linguistical relationship between Ge'ez and Amharic language chapter 3.

To conduct these two experiment, we need to segment the words in each of the language based on the prefixes and suffixes in the languages. To do this need to write a python code for both languages.

In addition to unsupervised morpheme segmentation we, also designed rule-based morpheme segmentation. Figure 4-4 shows steps we followed in rule-based prefix and suffix morpheme segmentation. Given Un-segemented input corpus, we perform prefix and suffixes segmentation.

Prefix Segmentation

It is the process which uses lists of multiple prefixes for segmenting prefixes of a single word by iterating through it. This is based on the prefixes location. A single prefix is not occur in fixed position a word. It may appear in any position. It also uses two files that contains root/stem words and files that contains lists of words with no prefixes. In prefix segmentation more than one prefix can be prefixed to a single word. The prefix segmentation runs is from left to right. For example in Ge'ez **ወእምቅድመ** the first prefix is "**ወ**", the second prefix is "**እም**" and the root is "**ቅድመ**".

For example the ወርቅም ፣ ወርቅና ፣ ወርቅህም ፣ and ወርቅህንም. In order to generate the non-prefix containing words from the unsegemented word list of each language we write a script shown in appendix **V**.

Figure 4-4 shows the step we used to segment both prefixes and suffixes form a word.

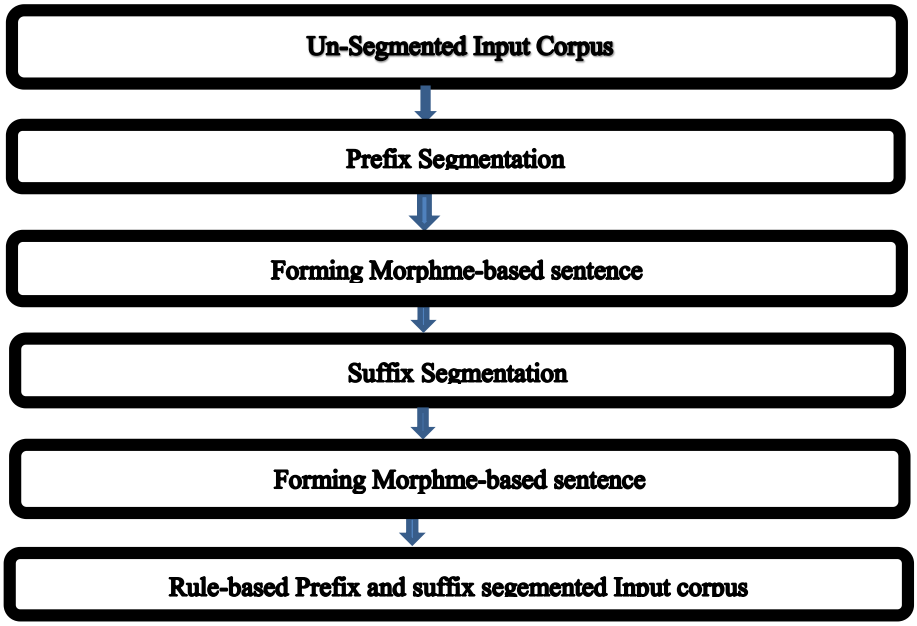


Figure 4-4 Rule Based Prefix and Suffix Segmentation Architecture

Algorithm for Prefix Segmentation refer to appendix VI

```
For pre in N :( N is either (prefix1, prefix2, prefix3 and prefix 4))
        new_snt = " "
        Flag = True
        If l == 1: (the first iteration)
        For pre in prefix1:
                If word.startswith (pre) and len (word [len (pre) :]) >= 2 and Word not in unsegemented
                and word not in nonprefixwords:
                        new_snt = new_snt + pre + " " + word [len (pre): len (word)] + "\n"
                        Flag = False
               If flag == True:
                  new_snt = new_snt + word + "\n"
                 After segmenting based on prefix1 lists write to a file. Then, go for next iteration and
                        prefix which is prefix2 until the end both iteration and prefix.
Finally, we have a file that show segmentation of all prefixes a last, which is all the word/prefixes appear
per line
```

As depicted above, to segmente a prefix type from word/stem, the program first check the word starts with the prefix type, secondly, it checks the length of the remaing subwords, and thirdly, it checks for root word in unsegemented list and finally check for non-prefix word in nonprefix list.

For each iteration the segmentation process were using its own prefix type. Example prefix1 is used for first iteration, prefix2 were used for second iteration and so on. Appendix XIII and XIV shows the prefix and suffix lists of Ge'ez and Amharic language used in prefix and suffix segmentation. Unsegemented list contains root words. Nonprefixwords lists contains non-prefix words.

The reason dividing prefixes into prefix1, prefix2, prefix3, and prefix4 is based on the order of their appearance as a prefix of a word. If two prefix like **"ወደ"** and **"ወ"** exists in the same prefix list the word is segemented two time. For example **"ወደላይ"** if it is segemented using it **ወደ ላይ**, and **ወ ደላይ** which is dual segmentation for a single word which is problem during transaltion.

Prefix segmentation starts from left to right. Prefix1 lists were segemented first from the word by checking the length of the word starting from the prefix length, both root word and non-prefixed words. The result of the segmentation is stored into another new file which is used as an input for the next iteration and prefix type.

Suffix Segmentation

It is segmenting suffixes added at the end parts of the word. It uses the output of prefix segmentation as an input. For segmenting the suffixes we were using suffix1, suffix2, suffix3 and suffix4. It is conducted from right to left. For example consider words in Amharic like, **አባትህ ፣ አባትህን ፣** and **አባትህንም** as you see the root is **አባት.** Suffix1 contains "**ህ**", "**ን**" and "**ም**", suffix2 contains "**ህ**" and "**ን**" and suffix3 contains "**ህ**". Therefore, the suffix **"ህ",** existence in each suffix list is based on its position in each word. One of the challenge is segmenting infixes that we do not work segmentation. Algorithm for suffix Segmentation referring to appendix VII

```
For su in N :( N is a number either suffix1, suffix2, suffix3 and suffix4)
        new_snt
        Flag = True
        If l == 1: (the first iteration)
        For su in suffix1:
                If word.endswith (su) and len (word [len (su):]) >= 2 and word not in
                unsegemented:
                        new_snt = new_snt + su + " " + word [len (su):] + "\n"
                        Flag = False
        If flag == True:
                new_snt = new_snt + word + "\n"
        After segmenting the file based on suffix1, write to a file Then, go for next iteration and
        suffix which is suffix 2 until the end both iteration and suffix.
```

Finally, we have a file that show segmentation of all suffixes and a last, which is all the word/ suffixes appear per line. At this step, we have lists of words having both prefixes and suffixes of a word is segmented. The last file of the suffix segmentation file contains it, which is used for input corpus.

Forming Morpheme -based sentence

The third step in the process of preparing the data for transaltion using rule based is forming the segmented words to their sentence level alignment. For forming segemented morphemes to sentence level we wrote a python script. It is used at the end of prefix and suffix segmentation. After prefix segmentation we merge the result for suffix segmentation. Again the result of suffix segmentation is merged to form morpheme-based sentences for transaltion.

4.3. Experimentation

After designing bi-directional Ge'ez-Amharic translator and preparing dataset, the next step is conducting different experiment using word and morpheme as a translation unit.

4.3.1. Experiment setup

This section describes the toolkit used for building the language and translation model. It also illustrates the hardware and software used in conducting the experiments.

System Environment	
Manufacture	Dell
Model	OptiPlex 3020
Processor	Intel core i3-4250 CPU
Processor speed	3.50 GHZx4
Memory	4GB

(a)

Software Experimental Setup	
OS	Ubuntu 16.04 LTS
Moses-Decoder	For transaltion setup
MGIZA++	for extracting word and morpheme alignments
SRILM	To build the language model of words and morpheme.
Morfessor,	used for segmentation of words
Pycharm	Used for python and shell scripting.

(b)

Table 4-3 Hardware (a) and software (b) experimental Setup

After performing the necessary software installation and preprocessing techniques, we create data folder in desktop having two folders namely **am_ge** and **scripts**. The **am_ge** folder is that contain parallel input corpus for both language Ge'ez and Amharic whereas **scripts** that contains python and shell scripts used in the SMT system.

We conducted six experiments using word and morpheme as a translation unit. While word as a unit of translation two experiments were conducted, four experiments were conducted at morpheme level, (Two experiments using unsupervised morpheme segmentation and the other two using rule-based segmentation). Finally, the one which performs the best is selected as an optimal unit of translation for bi-directional Ge'ez and Amharic MT.

4.3.2. Word-based bi-directional translation

The first two experiments are baseline experiments. We used word aligned corpus for the bi-directional translation process from Ge'ez to Amharic and Amharic to Ge'ez.

Experiment I: Word-based translation from Geez-to-Amharic

The first experiment is conducted to test word-based Geez to Amharic machine translation. The source language is Ge'ez (input text for the translation processes) and target language is Amharic (which is the output of the translation processes).

Experimental result shows that, the system translates the given text to the target language (Amharic) with **8.37%** BLEU score. Figure 4-5 presents sample translation input text in Ge'ez language.

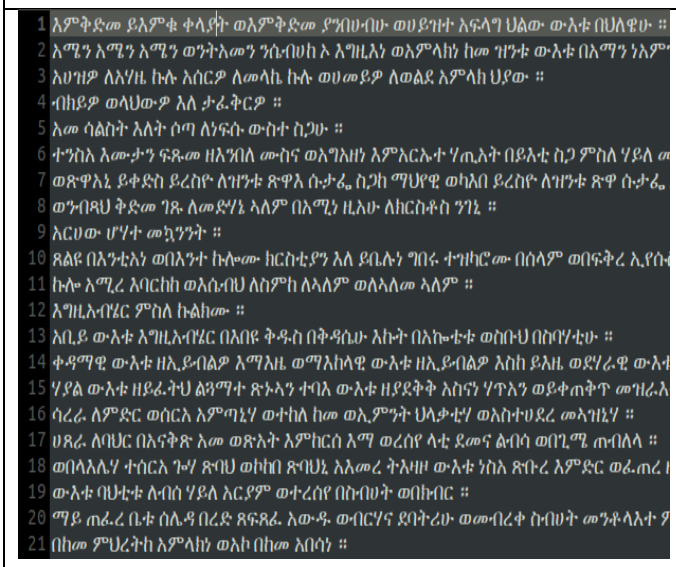
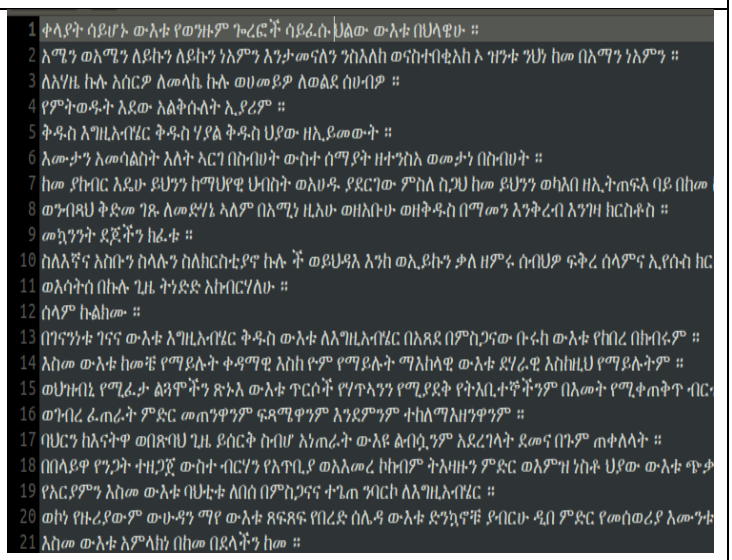
Input text is Ge’ez (a)	Output text is Amahric (b)
	

Figure 4-5 Sample translation input (a) and output (b) for Geez to Amharic translation word level alignment.

However there are many sentences or words are untranslated into Amharic such as **“ይትጋውሁ ብርሓናት”** in the first sentence, **“ዘዓይን ኢርእየ ወእዝን ኢሰምዐ ውስተ ልበ ሰብእ ዘኢተሐለየ ዘአስተዳለወ”** in line 5 and **ብክይዎ ወላህውዎ** last sentence. These occurred because of the out of vocabulary of the training set, morphology, alignment problem and the syntax.

As presented in Amharic one word is **“ሳያሆኑ”** aligned to two words **“እምቅድመ ይእምቁ”** in Ge’ez correctly, which means **“እምቅድመ”** is equivalent to **“ሳያ”** and **“ሆኑ”** is equivalent to **“ይእምቁ””.** The first reason for poor performance of the translation is the **transaltion unit** used. As we all known both of the languages are **morphological richness**.

When a language is morphological rich the number of word produced is based on affixes used in that language. In Morphologically rich language words are ambiguous which a single word expresses a number features based on the prefix, suffixes, circumfix and infixes.

As you can see form, **“ዘዓይን ኢርእየ ወእዝን ኢሰምዐ ውስተ ልበ ሰብእ ዘኢተሐለየ ዘአስተዳለወ”,** a single word is composed of prefixes and suffixes which are specific, manageable and well known. For example in Ge’ez language morphology prefixes like **“ዘ ፣ ኢ ፣ ወ ፣ ኢተ ፣ እም ፣ አስተ** and son on**”** and suffixes like **“ኩ ፣ ነ ፣ ከ ፣ ክሙ ፣ ኪ ፣ ክን ፣ ት** and son on”. These affixes can be used in different words in the same way in each word. In word based transaltion listing all words in a language like Ge’ez and Amharic is difficult during Statically Machine Transaltion.

Therefore, we in this research try to use these units of transition to enhance the performance of SMT system of Ge'ez to Amahric automatic translation. It implies that there is a need to segment both the source and target language corpus.

In Case of Ge'ez and Amharic using these affixes and single word, we can create infinite number of word type, which increases out of vocabulary. To include all this words in the corpus is difficult but being knowing the affixes help us. This is due the number of affixes in a language both Ge'ez and Amharic, which well known, specific, and manageable and consistency.

The second reason for poor performance is **alignment**. We identify all types of alignment **one to one**, **one to many**, **many to one** and **many to many**. For best transaltion one to one transaltion is best alignment for SMT performance enhancement and other alignment type decrease the performance.

For example, **"ቅዱስ"** in Amahric one is aligned with one word of Ge'ez **"ቅዱስ"**, one word **"ሳያሆኑ"** in Amharic aligned with many words in Ge'ez **"እምቅድመ ይእምቁ",** in Amharic many words **"ስድስት መቶ አንድ ሺህ ሰባት መቶ ሰላሳ"** is aligned to one Ge'ez **ስሳ እልፍ አስርቱ ወሰባዕቱ ምዕት ወሰላሳ** word and many words in Amharic **"የሰው ልጅ"** are aligned to many words of Ge'ez **"ደቂቀ እጓለ እመሕያው".** This implies the SMT performance become when many types of alignment exists.

The third reason is **syntactic structure** difference of the languages. Amharic language sentence structure Subject Object Verb (SOV) but sentence structure for Ge'ez Subject Verb Object (**SVO, "እግዚአብሔር ነበቦ ለሙሴ")**, Verb Subject Object (**VSO, "ወነበቦ እግዚአብሔር ለሙሴ")** and Object Verb Subject (**OVS, "ለሙሴ ነበቦ እግዚአብሔር").** In which ever types syntax it is in Amharic it follows SOV, which means **"እግዚአብሔር ሙሴን ተናገር".** This is one of the challenges made the performance to be low.

The position of an adverb and adjective in Amharic is before the verb and noun respectively, while in Ge'ez it is used in both before or after verb and noun. Let think Ge'ez have **three**, word order type and the position of an adjective in Ge'ez have **two**, before and after a noun and that of an adverb is also **two**. If all exists in a single sentence, we have a total of 3*2*2 = 12 types of syntax, which is so challenging for decoder to select the best transaltion.

For this study we can enhance the performance of the transaltion by applying morphological segmentation on surface word on both language Ge'ez and Amahric.

Experiment II: Word-based translation from Amharic to Ge'ez

Because of the system works bi-directional, this experiment checks the performance of the system with the same corpus used in the experiment I. The same text to translate from source language Amharic to target language Ge'ez. We used the following Amharic text as input for translation. Generally, the system translates the given text to the target language (Ge'ez) with **8.42%** BLEU score.

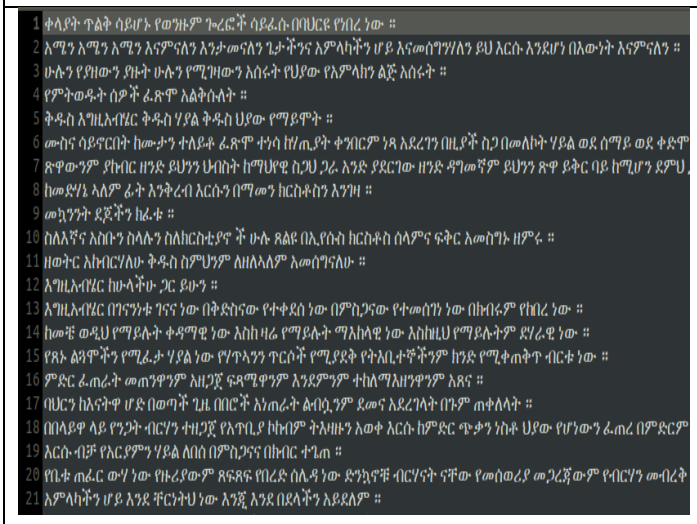
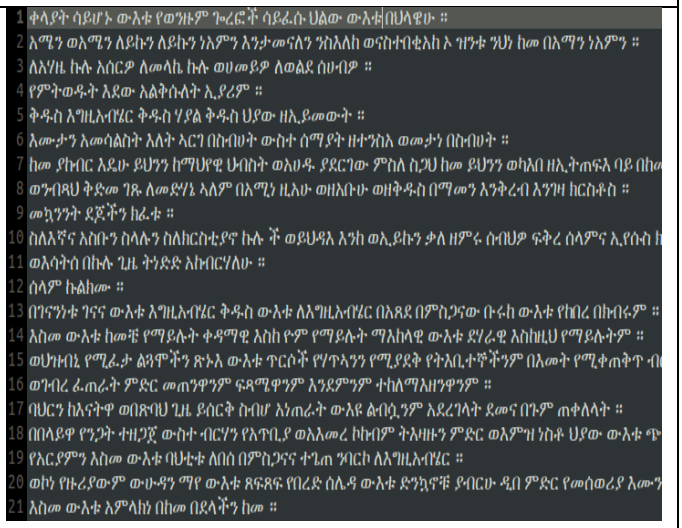
Input text is Amharic (a)	Output text is Ge'ez (b)
	

Figure 4-6 Sample Translation input (a) and output (b) from Amharic to Geez Word as a translation Unit

The first reason for poor performance of the SMT system is the **transaltion unit** used in conducting SMT between Ge'ez and Amharic. As indicated in the first experiment I.

The second reason for poor performance of the transaltion is **Alignment problem**. We identify all types' alignments 1:1, 1: m, m: 1 and m: m as described above.

The third reason **syntactic problem**. As mentioned above in section 4.3.2, Amharic has word order syntax SOV, but Ge'ez have **SVO**, **VSO**, and **OVS** syntax. Conducting SMT, in such situation is challenging. If two language with the same word order the translation performance is better.

From the above two experiments we conclude that the two languages are morphologically equivalent. The performance of the SMT system is poor due to the morphological richness of both language is relevant for this research beside the alignment, and syntactic challenges.

Therefore, to enhance the performance of the SMT system we need to segment words to their sub words, morpheme. In being changing the unit of transaltion we conducted the next experiments based on sub words as morpheme.

4.3.3. Morpheme-based bi-directional transaltion using unsupervised morphological segmentation

The next two experiments are conducted on the basses of morphemes as a translation unit. To conduct the next two experiment the corpus is prepared by applying unsupervised morphological segmentation tool morfessor. For aligning the morphemes, we used MGIZA++.

Experiment III Morpheme-based transaltion from Ge'ez to Amharic

For experiment III we use, Ge'ez text as an input for source language and Amharic is target language. Generally, the system translates the given input text in Ge'ez to the target language Amharic with **14.54%** BLEU score. As compare to the baseline experiment of this research, **6.17%** BLEU performance enhancement were recorded on the transaltion of Ge'ez to Amharic using morpheme generated by using morphological segmented.

This result is achieved due to morphological segmentation of both the source and the target language to their equivalent morphemes. Morphemes in a given language are specific, manageable and consistence. Therefore, managing sub-parts of a word is easy and create consistence.

Form the figure 4-7(a) there is over segmentation and under segmentation; for example, in the first line **"ህሊናከሙ"** should be segmented to **"ህሊና ከሙ"** but as you can see it segmented **"ህሊ ና ከሙ"** which is over segemented. In the figure 4-7 (a) line 3 **"ወረሰየነ"** should be segemented to **"ወ ረሰየ ነ"** but it is segemented to **"ወ ረሰየነ"** which is under segmentation. There are also words that are still un-segmented**,** for example, in the first line **"ልብከሙ"**, in line 2 **"ይትረበባ ፣ አድባራት"** and so on. Finally there is also perfect segmentation like line two **"እምቅድም"** to **"እም ቅድም" ፣ "ዘውእቱ"** to **"ዘ ውእቱ"**.

As you can see from the output of the experiment some sentence or morphemes are not correctly translated in to Amharic such as **"ይትረበባ ፣ አድባራት"** and morphemes that are not exists in the transaltion of Amharic appears. This happened due to the under segmentation such as **ይትረበባ** when it is manually segemented it look like **ይ ት ረበባ** but morfessor segment as **ይትረበባ.**

There is also alignment problem observed with the experimental result. We identify all types of alignment as show in the figure 4-7, namely **1:1, 1: m, m: 1** and **m: m.**

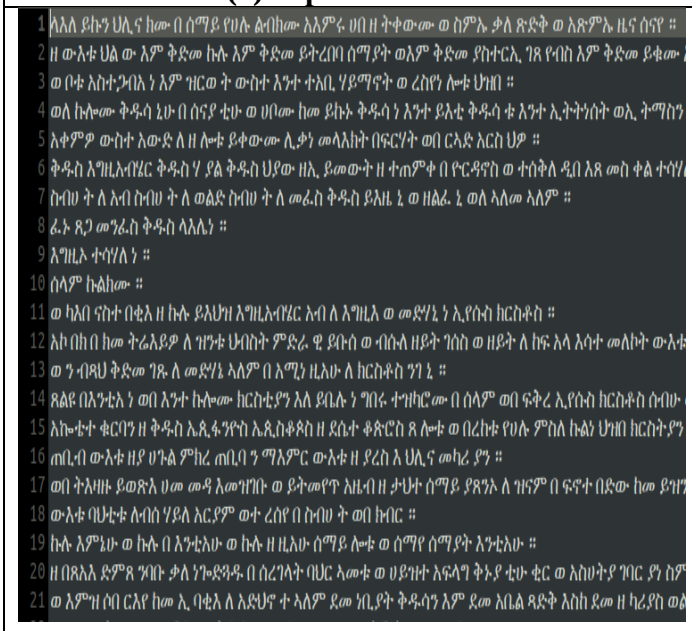
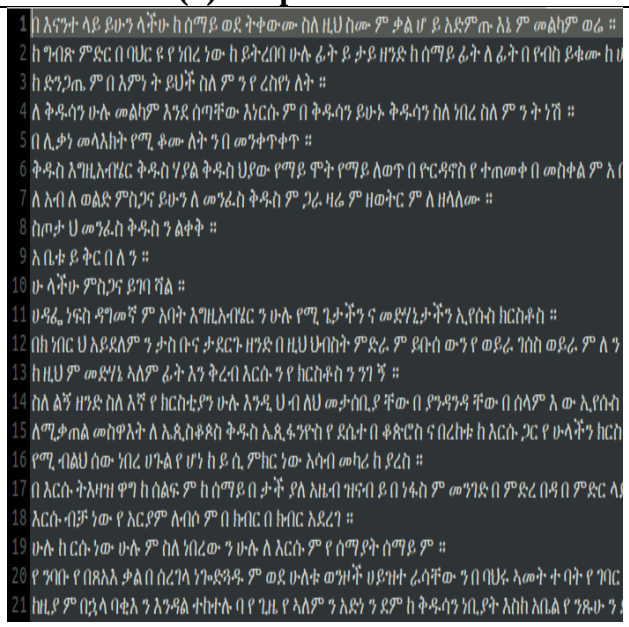
(a) Input text is Ge’ez	(b) Output text is Amharic
	

Figure 4-7 Sample translation input (a) and output (b) for Ge’ez to Amharic translation morpheme level alignment.

Experiment IV: Morpheme-based translation from Amharic to Ge’ez

In this experiment is morpheme-based transaltion is done using Amharic and Ge’ez as the source and target languages respectively. Experimental results shows that the system translates the given Amharic text to the target language (Ge’ez) with **14.88%** BLEU score. As compare to the word based there is enhancement of MT performance of **6.46.**

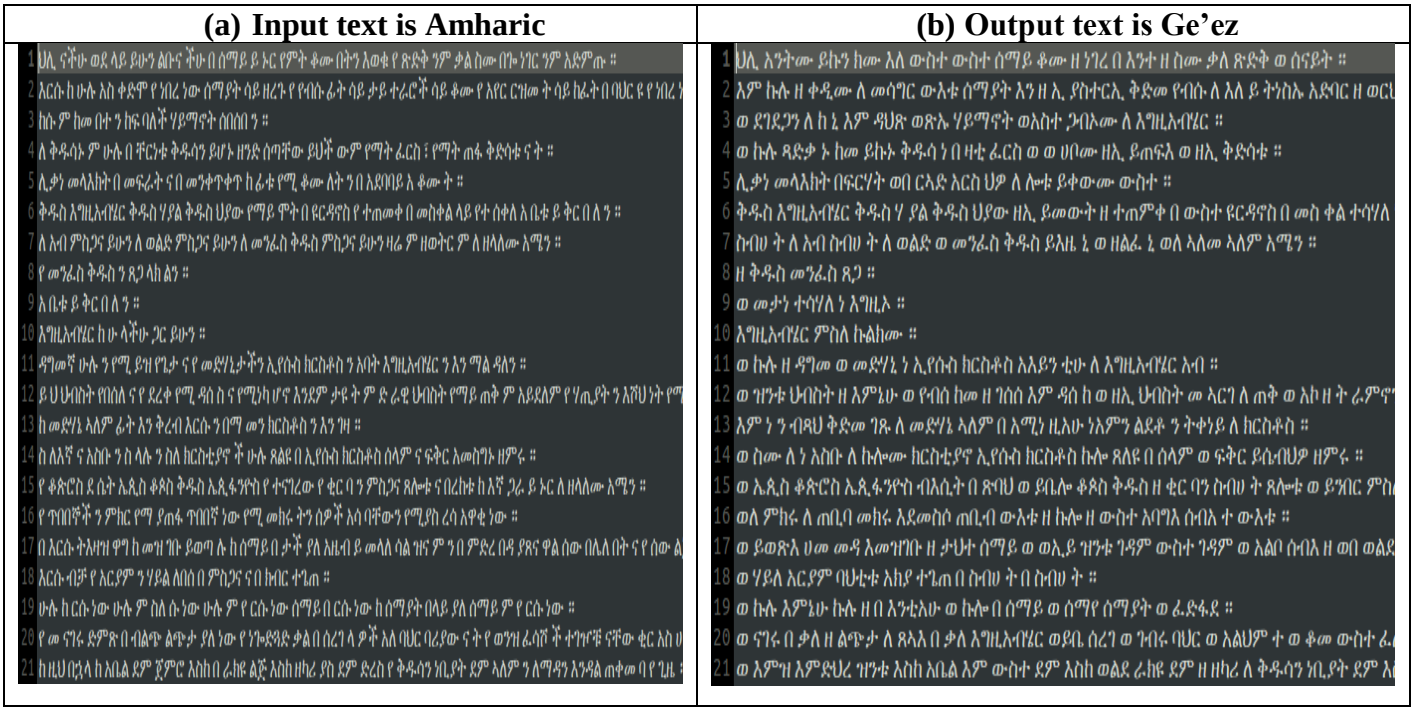


Figure 4-8 Sample translation input (a) and output (b) for Ge’ez to Amharic translation morpheme level alignment.

As you can see from the above figure 4-7 there are under and over segmentation problems for poor performance of the system even if morpheme is enhance the performance of the MT.

In the first line the word **“ህሊናችሁ”** should be segemented to **“ህሊና ችሁ”** but it is segemented to **“ህሊ ናችሁ”,** line 9 **“አቤቱ ይቅር በለን”** should be segemented but it is wrongly segmented to **“አ ቤቱ ይ ቅር በ ለ ን”** in which over segmentation exists,

4.3.4. Morpheme-based bi-directional transaltion using rule based morpheme segmentation

This experiment is conducted with data prepared using rule-based segmentation for both language. The source and target languages are Ge’ez and Amharic respectively.

Experiment V Morpheme-based transaltion from Ge’ez to Amharic

This experiment is conducted with morpheme-based dataset prepared using Amahric as a source language and Ge’ez as a target language. Experimental results shows that the system translates the given text to the target language (Amharic) with **15.14%** BLEU scored.

As compared to the word based and unsupervised morpheme segmentation there is an enhancement of MT performance of **6.73%** and **0.6%** respectively**.** This enhancement is due to the fact that rule-based is the exact experiment how sub-words are formatted to form words.

However, we need extremely to connect all the rules required for morpheme segmentation; such that rules required for word class, how it is inflicted and derived from its stem or root should be well crafted.

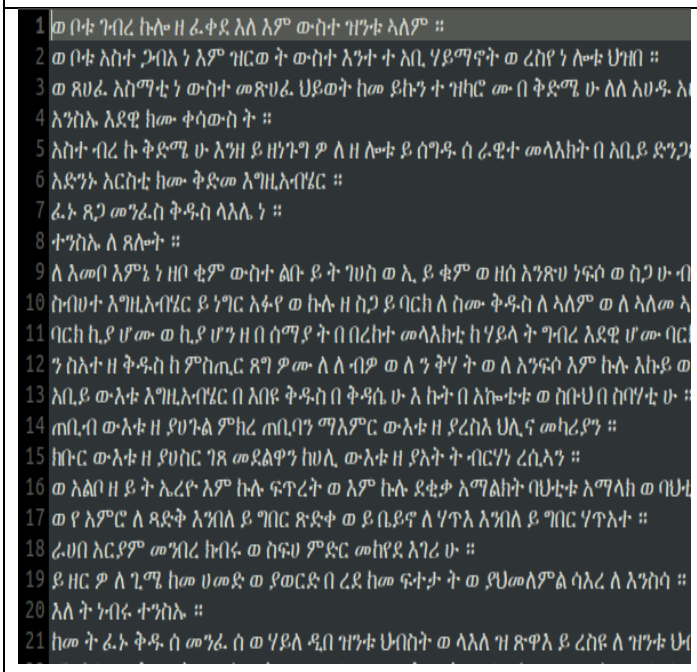
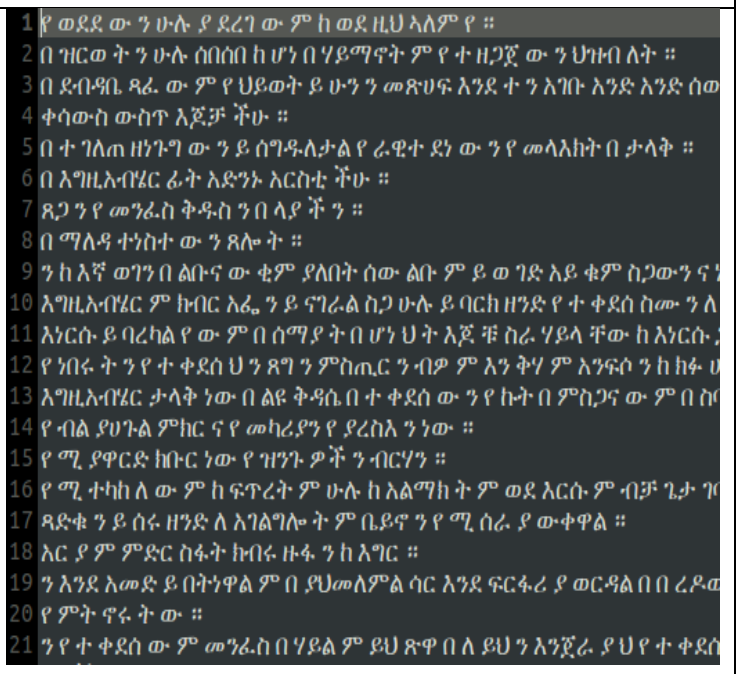
(a) Input text is Ge'ez	**(b) Input text is Amahric**
	

Figure 4-9 Sample Translation input (a) and output (b) for Ge'ez to Amharic, Morpheme as translation Unit using Rule based Approach

The purpose of the experiments were to show that changing the transaltion unit from word to morpheme enhances the performance of the SMT system. While conducting segmentation using Rule-Based Approaches, we need to know each rules of the each word class how it is inflicted and derived from its stem or root.

Experiment VI: Morpheme-based transaltion from Amharic to Ge'ez

This experiment is conducted with data prepared using rule-based segmentation for Amharic as a source and Ge'ez target languages. The system accordingly, translates the given text to the target language (Ge'ez) with **16.33%** BLEU score.

As compare to the word based there is an enhancement of MT performance by **7.73%**, and also **1.27%** enhancement as compared to unsupervised morpheme segmentation.

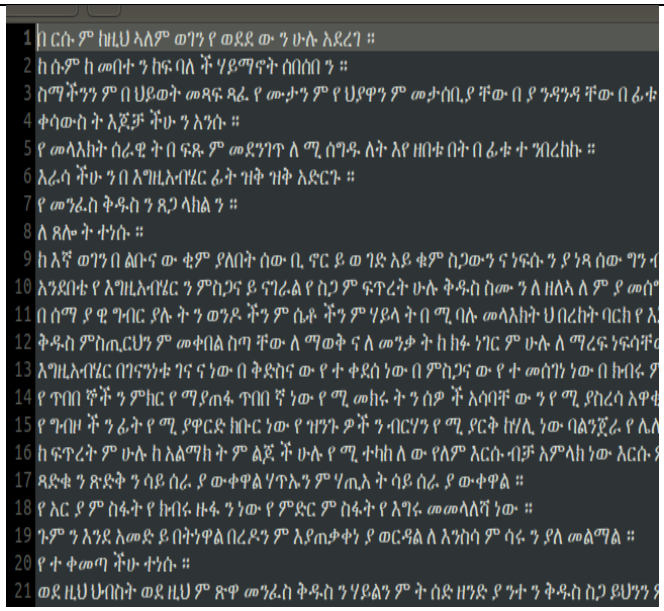
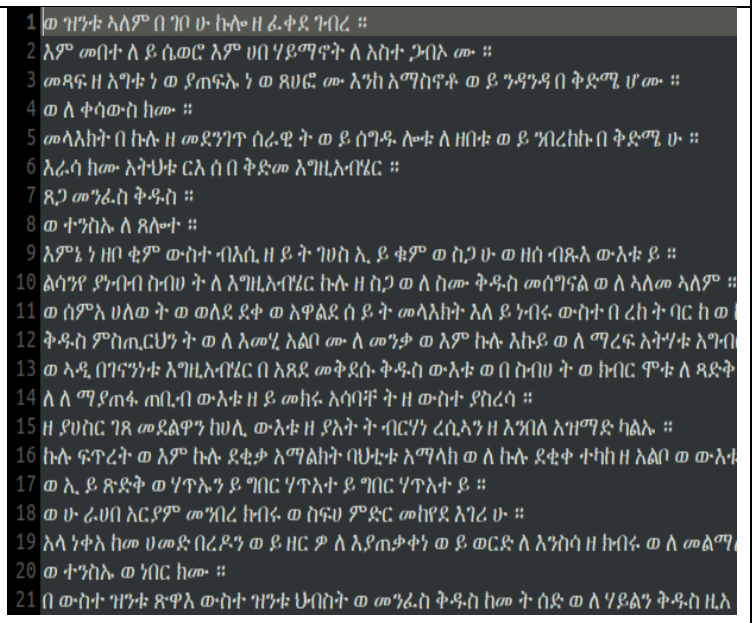
(a) Input text is Ge'ez	(b) Output text is Amharic
	

Figure 4-10 Sample Translation input (a) and output (b) for Amharic to Ge'ez, Morpheme as translation Unit using Rule based Approach

Preparing data set using morfessor is based on the morfessor model which is based on corpus size but that of the rule based is based on the rule that we used in segmentation processes. Morfessor requires corpus knowledge for segmentation which is economical and supported by technology. Rule based segmentation inquires to know detailed linguistic knowledge about the languages that we need to segment which is not economical, time consuming and.

4.4. Discussion of Result

The main purpose of this study is to conduct experiment on morpheme-based bi-directional transaltion of Ge'ez-Amharic for better performance. Six different experiments were conducted from Ge'ez-Amharic and from Amharic-Ge'ez languages. Two and four experiments were conducted using word and morpheme as a transaltion unit respectively. Summary of the experimental result is presented in table 4-3 below.

Types of experiment conducted		**Result of experiment in BLEU from both directions**	
		Ge'ez to Amharic	**Amharic to Ge'ez**
Word-Based Translation		**8.37%**	**8.42%**
Morpheme Based Transaltion	**Using Morfessor**	**14.54%**	**14. 88%**
	Using Rule-based	**15.14%**	**16.15%**

Table 4-4 Summary of experiment result

As depicted in the above table 4-3 morpheme-based transaltion performed better than word-based with performance improvement of greater than **6% BLEU** score. In order to achieve better result the corpus is aligned at morpheme level by using MGIZA++ algorithm. This decreases the number of non-aligned morpheme in the corpus and increase the number of aligned morpheme at phrase translation table. This makes the translation performance better.

Dataset being prepared using unsupervised morpheme segmentation performs **14.54%** and **14. 88%** BLEU score from Geez to Amharic and from Amharic to Geez respectively. And also dataset prepared using rule-based segmentation performs **15.14%** and **16. 15%** BLEU score from Geez to Amharic and from Amharic to Geez respectively. As we compare the result rule-based morpheme segmentation performs better than unsupervised morphological segmentation. This is due to rule-based morpheme segmentation uses rules well crafted by linguist that directs to the morphemes of the language. Rule-based morpheme segmentation requires linguistic knowledge to generate well-crafted rules, time taking, resources insentive and it is long term work plan. On the other hand the unsupervised morpheme segmentation techniques generates the rules from corpus of the language, which is economical and doesn't need linguistic knowledge.

As shown in table 4-3 morpheme-based MT performs better than word-based MT. This is due to, at word-level conducting MT between two morphological rich languages is challenged by many word form of a single word, which is unmanageable, not specific and inconsistent. But, at morpheme-level the MT is not challenged by many forms of a single word since morphemes are specific, manageable and consistent.

Regarding direction of transaltion as depicted in table 4-3, Amharic to Ge'ez MT performs better than Ge'ez to Amharic MT. This is due to the word correspondences from Amharic to Ge'ez is one to many. Based on the dataset we have prepared their exist alignment of one word of Amharic Aligned to many words of Ge'ez. As depicted in figure 4-11, alignment is one of the challenge observed in morpheme-based machine transaltion, especially conducting MT between two morphological rich languages like Ge'ez and Amharic.

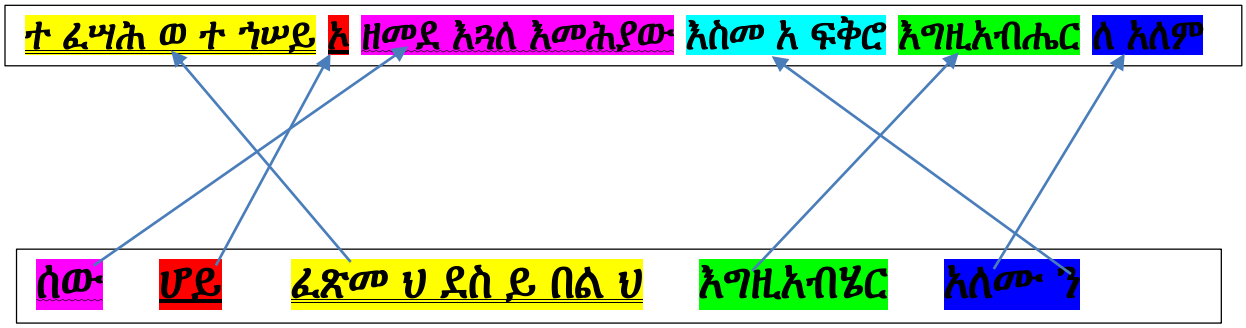


Figure 4-11 Amharic -Ge'ez Alignment Challenges

A comparison is also made with related research done by Dawit [15]. The main focus of the research is applying SMT from Ge'ez to Amharic. The study use word as translation unit. Word level alignment, normalization, and uni-directional (Geez to Amharic only) was done for both languages. Experimental result shows that the system achieves 8.26 % of BLEU score translation performance. As reported by Dawit, the performance of the translator reduced because of morphological richness of the two language. Accordingly, in this study morpheme-based transaltion is experimented, in which we register an improvement in performance.

The result indicate that data set prepared using rules of each language were performing better but we need to have either self-deep linguistic knowledge for both language or professionals that are willing to support the experiment. It also takes time and resources for constructing rules. But, that of the morfessor is unsupervised segmentation which need to increase the corpus size as much as possible.

In general, the translation performance of this study is better than the previous study. However there are translation errors observed. It is better to explore further morpheme-based machine transaltion for Ge’ez-Amharic.

Chapter Five

Conclusion and Recommendation

5.1. Conclusion

Morphologically rich languages like Ge'ez and Amahric pose a challenge for statistical machine translation, as these languages possess a large set of morphological features producing many rich surface forms. Morphologically complex languages are well known to cause problems for contemporary statistical machine translation (SMT) systems. This is because of a single word consists of one or more sub-words called morpheme. Therefore, in this study we aim to explore an optimal translation unit for Ge'ez-Amharic bi-directional transaltion. To achieve this goal, we first studied the morphology and syntax of both Geez and Amharic language. Accordingly, we identify both languages have equivalent morphological richness and Geez is a free grammar language regarding the syntax being SVO, VSO, or VOS. The position of the adverb and adjectives also in Geez is any place before or after a verb and a noun respectively. There is also word correspondence between the two languages one-one, one-many, many-one and many-many.

The design process of bi-directional Geez-Amharic machine translation involves collecting Geez to Amharic parallel corpus. The corpus collected from freely available online sources such as Old Testament holly bible, anaphora or Kidase and manually prepared bitext includes Wedase Marya, Anketse Berhane, yewedesewa melahekete, Kidan and Liton. Corpus preparation involves activities of preprocessing the corpus such as tokenization (for both Geez and Amharic) and character normalization (only for Amharic). Morfessor and morphological rules are used to segment morpheme of Ge'ez and Amharic in unsupervised and rule-based manner respectively. And they were used to find morpheme of Geez and Amharic. MGIZA++ used for word and morpheme level alignment. Moses for used for translation process which integrate all necessary tools for machine translation such as IRSTLM, MGIZA++ and decoder.

To identify an optimal translation unit, we conduct different experiment on each translation unit called word and morpheme. Based on unsupervised morpheme segmentation using morfessor the study creates morpheme-based datasets which achieves **14.54%** and **14.88%** BLEU score from Geez to Amharic and Amharic to Geez respectively.

On the other hand based on rule based segmentations, register **15.14%** and **16.15%** BLEU score from Geez to Amharic and Amharic to Geez respectively. Unsupervised morpheme segmentation is suitable approaches as an IT professional beside the knowledge of linguist is mandatory.

This study achieves a promising result that identifies morpheme as an optimal unit of translation and it enhances the performance of bi-directional Ge'ez-Amharic machine translation. However, being conducting machine transaltion between morphologically rich languages, there are a number challenges observed. One of the challenge is mis-alignment especially when there are many to many correspondence between words/morphemes. The alignment problem becomes also challenging because of multiple syntactic order used in Geez writing. In addition handling morphological richness of the two languages requires standard corpus especially for machine learning algorithms.

5.2. Recommendation

This study explore morpheme-based bi-directional machine translation for Ge'ez-Amharic languages. Based on the finding we would like to recommend the following points for further works:

- One of the challenges in conducting machine translation of Geez-Amharic is the flexibility of syntactic structure of Ge'ez. To simplify the transaltion process this is a need to map all the syntax of Geez to one standard syntax, SVO.
- It is a challenging task to collect and prepare data for local languages. So there is an immediate need to initiate research to prepare standard corpus for local languages that can be used as test bed to evaluate the advancement in machine translation for local languages.
- To exploit the strength of the two major machine learning approaches, further research may be conducted between Ge'ez and Amahric using hybrid of statistical and rule-based machine translation.
- Most of the corpus used for this study is collected from Holly bible and religious documents. To undertake a comprehensive experiments there is a need to prepare a corpus from different disciplines.
- Alignment of Ge'ez-Amharic text is a challenging task because of many-to-many correspondence between words/morphemes of the two languages. Hence, there is a need to identify optimal alignment for Ge'ez-Amharic Machine transaltion.
- In this study we use prefix and suffix for rule-based morphological segmentation. However since both languages are morphological rich, there is a need to apply machine learning algorithms for designing an optimal model for segmentation.
- In this study we focus only on morpheme and word as a translation unit, further research can be done on other unit of translation like phrase, sentence.
- Further research may be conducted Ge'ez to morphologically simple language such English to enhance SMT performance, since there is source to target language asymmetry is another problem for conducting SMT between two morphological rich language.

References

[1] Jurafsky and Martin , An introduction to natural language processing,computational linguistics, and speech recognition, Uited States of America: Prentice-Hall Inc, 2000.

[2] Sloculn, Jonathan, "A survey of machine translation: its history, current status, and future prospects," Jonathan Sloculn, 1985.

[3] Koerner and Asher., Concise history of the language sciences:, Sumerians to the cognitivists, 1995.

[4] Koehn, Philiph, Statistcal Machine Transaltion, United States of America: by Cambridge University Press, New York, 2009.

[5] MA Arba Berdica, " Book of Proceedings, The positive impact of technology in translation," in *International Conference on Linguistics, Literature and Culture*, 2016.

[6] Yitayal, Abate, "Morphological Analysis of Ge'ez Verbs Using Memory Based Learning," A Masters Thesis sumited to Addis Ababa University, Addis Ababa, 2014.

[7] Desta Berihu Weldegiorgis, "Design and Implementation of Automatic Morphological Analyzer for Ge'ez Verbs," A Masters Thesis sumited to Addis Ababa University, Addis Ababa, 2010.

[8] Andrea, DeCapua, Grammar for Teachers, A Guide to American English for Native and Non-Native Speakers, USA, American: Springer Science+Business Media, LLC, 233 Spring Street, New York,, 2008.

[9] መምህር ደሴ ቀለብ, ትነሣኤ ግእዝ, አዲስ አበባ: ማኅበረ ቅዱሳን, 2008 እኢአ.

[10] ባየ ይማም, የአማርኛ ሰዋስው፣የተሻሻለ ሁለተኛ እትም, አዲስ አበባ ፣ ኢትዮጵያ: ካልቸር ኤንድ አርት ሶሳይቲ አፍ ኢትዮጵያ, 2000 አንደ ኢትዮጵያ.አቆጣጠር.

[11] ደሴ በቀለ, , ትንሳኤ ግእዝ, አዲስ አበባ፣ ኢትዮጵያ: በኢትዮጵያ አርቶዶክስ ተዋህዶ ቤተ ክርስቲያን በሰንበት ት/ቤቶች ማደራጀ መምሪያ ማህበረ ቅዱሳን, 2002፣ በአዲስ አበባ ዩንቨርስቲ የስነ ቋንቋ መምህር.

[12] Adam Lopez and Matt Post, "Beyond bitext: Five open problems in machine translation," *Human Language Technology Center of Excellence Johns Hopkins University*, 2013.

[13] M. D. Okpor , "Machine Translation Approaches: Issues and Challenges," *IJCSI International Journal of Computer Science Issues*, vol. 11, no. 5, pp. 159-165, 2014.

[14] አባ ኪዳነ ማርያም ጥዑመ ልሳን, 14ቱ መዝገብ ቅዳሴ, አ.አ ኢትዮጵያ: አኰቴት አሳታሚዎቸ, 2009.

[15] Dawit, Mulugeta, "Geez to Amharic Automatic Machine Translation: A Statistical Approch," Msc Thesis,submitted to Addis Ababa University, Addis Ababa, Ethiopia, 2015.

[16] ዘርአዳዊት, አድሐና, ልሳናተ ሴም (ግእዝ፣ትግራይ፣አማርኛ) ንጽጽራዊ መዝገበ ቃላት, አዲስ አበባ፣ ኢትዮጵያ: ሜጋ አሳታሚና ማከፋፈያ ኃ/የተ/የግ/ማኅበር, 2009 ፣ መምህረ ልሳነ ግዕዝ ወትርጓሜ መጻሕፍት አዲስ ኪዳን ቅድስት ሥላሴ መንፈሳዊ ኮሌጅ.

[17] Rubin, Aaron D., A Brief Introduction to the Semitic Languages, United States of _werica: Library of Congress Cataloging-in-Publication Data, 2010.

[18] Atelach Alemu Argaw and Lars Asker, "An Amharic Stemmer : Reducing Words to their Citation Forms," in *Proceedings of the 5th Workshop on Important Unresolved Matters, pages 104–110*, Czech Republic, June 2007.

[19] Sisay, Adugna, "English-Afaan Oromoo Machine Translation: An expermental using Statistical Approach," A Master Thesis Submitted to Addis Ababa University, Addis Ababa, Ethiopia, 2009.

[20] Ceausu, Alexandru, "Rich morpho-syntactic descriptors for factored machine translation with highly inflected languages as target," Centre for Next Generation Localisation, Dublin City University, 2010.

[21] Eleni, Teshome, "Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus," A Master Thesis , submitted to Addis Ababa University, Addis Ababa, Ethiopia, 2013.

[22] Jabesa, Daba, "Bidirectional English – Afaan Oromo Machine Translation Using Hybrid Approach," A Master Thesis,submitted to Addis Ababa University, Addis Ababa, Ethiopia, 2013.

[23] Yitayew, Solomon, "Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation," A Master Thesis Submitted to Addis Ababa University, Addis Ababa , Ethiopia, 2017.

[24] Tariku, Tsegaye, "English-Tigrina Factored Statistical Machine Translation," A Master Thesis submitted to Addis Ababa University, Addis Ababa, Ethiopian, 2014.

[25] Mulubrahan, Hailegebreal, "Bidirectional Tigrigna – English Statistical Machine Translation," Msc Thesis,Submitted to Addis Ababa University, Addis Ababa, Ethiopia, 2017.

[26] Randil Pushpananda, Ruvan Weerasingh1, and Mahesan Niranjan, "Statistical Machine Translation from and into Morphologically Rich and Low Resourced Languages," Springer International Publishing Switzerland, 2015.

[27] Kumar, Ranjit, Research Methodology a step-by-step guide for beginners, England, London: SAGE Publications Ltd, Thrid Edition, 2011.

[28] Steven M. Ross and Gary R. Morrison, "Experimental Research Methods,," in *In Experimental Research Methods,*, pp. 1021 -1043.

[29] Mahibere Kidusan, "www.eotcmk.org," MK IT. All rights reserved. e-Governance system by Tekle Consulting, 2014. [Online]. [Accessed October 2017].

[30] W. John Hutchins, Harold L. Somers, Introduction to Machine Transaltion, 1992.

[31] Peter Smit, Sami Virpioja, Stig-Arne Gronroos and Mikko Kurimo, "Morfessor 2.0: Toolkit for statistical morphological segmentation," *Association for Computational Linguistics*, pp. 21-24, 2014.

[32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), July 2002, pp. 311-318.*, Philadelphia, 2002.

[33] Mohamed, Amine Chéragui, "Theoretical Overview of Machine translation," in *Proceedings ICWIT*, African University, Adrar, Algeria, 2012.

[34] Och F.J., "Challenges in Machine Translation," *International Symposium on Chinese Spoken Language Processing*, 2006.

[35] M. D. Okpor , "Machine Translation Approaches: Issues and Challenges," *IJCSI International Journal of Computer Science Issues*, vol. 11, no. 5, pp. 159-165, 2014.

[36] Keh-Yih Su and Jing-Shin Chang, "Why Corpus-Based Statistics-Oriented Machine Translation," *Department of Electrical Engineering National Tsing-Hua University Hsinchu, TAIWAN 30043, R.O.C.*, vol. 12, no. 11, pp. 249-262, 1992.

[37] Sonja Nießen and Hermann Ney, "Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information," *Computational Linguistics*, vol. 30, no. 2, pp. 181-205, 2004.

[38] Nagao, Makoto, "Some Rationales and Methodologies for Example-based Approach," [International Workshop on Fundamental Research for the Future Generation of Natural Language Processing, Manchester, 1992.

[39] Abdul-Rauf, Sadaf; Fishel, Mark; Lambert, Patrik; Noubours, Sandra; Sennrich, Rico, "Extrinsic evaluation of sentence alignment systems," *Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*, pp. 6-10, 2012.

[40] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra , and Robert L. Merecer, "The Mathematics of Machine transaltion: Parameter Estimation," *Comptational Linguistics*, vol. 19, no. 2, pp. 263-311, 1993.

[41] Michel Simard and Pierre Plamondon, "Bilingual Sentence Alignment: Balancing Robustness and Accuracy," *Centre for Information Technology Innovation*, pp. 135-144, 1998.

[42] Elif Eyigoz, Daniel Gildea and Kemal Oflazer, "Simultaneous Word-Morpheme Alignment for Statistical Machine Translation," in *Proceedings of NAACL-HLT*, Atlanta, Georgia, 2013.

[43] Sanja Seljan, Angelina Gašpar and Damir Pavuna , "Sentence Alignment as the Basis for Translation Memory Database," *Digital Information and Heritage*, pp. 299-311, 2007.

[44] Fabienne Braune and Alexander Fraser, "Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora," *Institute for Natural Language Processing*, vol. II, no. 12, pp. 81-89, 2010.

[45] Anil Kumar Singh and Samar Husain, "Comparison, Selection and Use of Sentence Alignment Algorithms for New Language Pairs," in *The ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, 2005.

[46] Smith, Jason R., "Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment," in *Annual Conference of the North American Chapter of the ACL*, Los Angeles, California, 2010.

[47] Moore, Robert C., "Fast and Accurate Sentence Alignment of Bilingual Corpora," *Machine Translation: From Research to Real Users*, pp. 1-10, 2002.

[48] al, Peter F. Brown et, "The Mathematics of Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263-311, 1993.

[49] Mathias Creutz and Krista Lagus, "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor," *Neural Networks Research Centre, Helsinki University of Technology*, pp. 1-27, 2015.

[50] Liang Tian, Fai Wong, and Sam Chao, "Word Alignment Using Giza++ And Cygwin On Windows," *International Journal of Engineering Research & Technology (IJERT)*, vol. 2, no. 5, pp. 1762-1765, 2013.

[51] William A. Gale and Kenneth W. Church, "A Program for Aligning Sentences in Bilingual Corpora," *Association for Computational Linguistics* , vol. 19, no. 1, pp. 75-102, 1993.

[52] Adrien Lardilleux, Franc¸ois Yvon and Yves Lepage, "Hierarchical Sub-sentential Alignment with Anymalign," in *Proceedings of the 16th EAMT Conference* , Trento, Italy, 2012.

[53] Ryan Cotterell, Arun Kumar and Hinrich Schutze, "Morphological Segmentation Inside-Out," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,*, p. 2325–2330, 2016.

[54] Linlin Wang and Zhu Cao and Yu Xia and Gerard de Melo, "Morphological Segmentation with Window LSTM Neural Networks," *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* , pp. 2842-2848, 2016.

[55] Lushtak, Sergei A., "Unsupervised Morphological Word Clustering," Computational Linguistics Master of Science, University of Washington, 2012.

[56] Sami Virpioja, Jaakko J. Väyrynen, Mathias Creutz, and Markus Sadeniemi, "Morphology-Aware Stastical Machine Transaltion Based on Morphs Induced in an Unsupervised Manner," in *Published in Proceedings of the Machine Transaltion Summit XI*, Copenhagen, Denmark, 2007.

[57] Fishel, Mark, "Deeper than Words: Morph-based Alignment for Statistical Machine Translation," in *Citeseer*, Tartu, Estonia, 2009.

[58] Mulu Gebreegziabher Teshome, and Laurent Besacier (Prof.), "Preliminary experiments on English to Amharic statistical machine translation," 2012.

[59] Wondwossen Mulugeta, Michael Gasser, and Baye Yimam, "Incremental Learning of Affix Segmentation," *Proceedings of COLING 2012: Technical Papers*, p. 1901–1914, Mumbai, December 2012.

[60] Coulmas, Florian, Writing Systems An Introduction to Their Linguistic Analysis, Deutches Institüt für Japanstudien, Tokyo: Cambridge University Press, 2003.

[61] Karan, Elke, "Writing System Development and Reform," A Msc Thesis in partial fulfillment of the requirements for the degree of Master of Arts, Grand Forks, North Dakota, 2006.

[62] Thomas Lambdin Oden, Introduction to Classical Ethiopic(Ge'ez), USA Amharica: The President and Fellows of Harvard College, 1978.

[63] ዘርአዳዊት አድሐና, መርኖ ሰዋስው ዘልሳነ ግዕዝ, አዲስ አበባ እኢትዮጵያ: ብርሃንና ሰላም , 1996 እ ኢ አ.

[64] Richardson, Leonard, Beautiful Soup Documentation Release 4.4.0, Nov 20, 2017.

[65] Tommi A Pirinen, Antonio Toral and Raphael Rubino, "Rule-Based and Statistical Morph Segments in English-to-Finnish SMT," *Free Publication*, pp. 1-11, 2 February 2016.

[66] "EthiopicBible.com," Powered By Abyssinica Search Engine., 2017. [Online]. [Accessed 21 September 2017].

[67] Peter Smit, Sami Virpioja, Stig-Arne Gronroos, and Mikko Kurimo, "Morfessor 2.0: Toolkit for statistical morphological segmentation," in *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 21–24*, Gothenburg, Sweden, April 26-30 2014..

[68] Reshef Shilon, Nizar Habash, Alon Lavie and Shuly Wintner, "Machine Translation between Hebrew and Arabic:Needs, Challenges and Preliminary Solutions".

[69] Gasser, Michael, Hornmorpho 2.5 User’s Guide, India: Indiana University, School of Informatics and Computing, 2012.

[70] Mulugeta, Seyoum, "The particle ʔɨnde in Amharic," *Studies in Ethiopian Languages,* vol. 3, pp. 83-95, 2014.

[71] Ann Clifton and Anoop Sarkar, "Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction," *Association for Computational Linguistics,* pp. 32-42, 2011.

[72] Ann Clifton and Anoop Sarkar, "Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 32–42, Portland, Oregon, 2011.

[73] W.John Hutchins, ""Machine translation: a brief history," in Concise history of the language sciences:," *Sumerians to the cognitivists,* pp. 445-460, 1995.

[74] Gasser, Michael, "Semitic Morphological Analysis and Genration Using Finite State Transducers with Feature Structures," in *in Proceeding Of the 12th Conference of the European Chapter of the ACL , 309-317,* Athens,Greece, 2009.

[75] Mohamed Amine Chéragui, "Theoretical Overview of Machine translation," in *Proceedings ICWIT,* African University, Adrar, Algeria, 2012.

Appendices

Appendix I: URL for sources of the corpus

1. https://www.ethiopicbible.com/ Ge'ez and Amharic aligned Bible text
2. http://ethiopianorthodox.org/amharic/yezemametsheft/tarik.html/The Anaphora of St Athnasious_Nov2015.pdf
3. http://ethiopianorthodox.org/amharic/yezemametsheft/tarik.html/The Anaphora of Saint Epiphaneous_29Nov2015.pdf
4. http://ethiopianorthodox.org/amharic/yezemametsheft/tarik.html/The anaphora of Saint John Chrysostom_December2015.pdf
5. https://www.stepbible.org/version.jsp?version=Geez

Appendix II: Python Scrip for Downloading the Dataset form Ethiopic Bible Web Sit

```
Import requests
from bs4 import BeautifulSoup
def get_bible_books():
    main_url = "https://www.ethiopicbible.com/amharic-bible-books"
    get_books = requests.get(main_url)
    if get_books.status_code == 200:
        booklists = get_books.content
        soup = BeautifulSoup(booklists, 'html5lib')
        li = soup.select("ol > li > a")
        books_of_bible = []
        for link in li:
            books_of_bible.append(link.get('href'))
        print("We have found " + str(len(books_of_bible)) + " books of bible")
        return books_of_bible
def content_crawl():
    books = get_bible_books()
    for item in books:
        book_iterator = 1
        while book_iterator < 151:
            print('https://www.ethiopicbible.com/' + item + "-" + str(book_iterator))
            get_content = requests.get('https://www.ethiopicbible.com/' + item + "-" + str(book_iterator))
            if get_content.status_code == 200:
                bookcontent = get_content.content
                soup2 = BeautifulSoup(bookcontent, 'html5lib')
                amharic_conetent = soup2.findAll("div", {"class": "amharicBibleChapterContainer"}) or None
                geez_conetent = soup2.findAll("div", {"class": "geezBibleChapterContainer"}) or None
                if amharic_conetent:
                  amharic_book = open("amharic/" + item + "-" + str(book_iterator) + ".txt", "w+")
             amaharictable = amharic_conetent[0].find('table').find_all('tr')
                    for each in amaharictable:
                        amahricverse = each.text
                        amahricverse = amahricverse.replace('\n', ' ')
                        amharic_book.write(amahricverse.strip() + "\n")
                        print(amahricverse)
                if geez_conetent:
                    geez_book = open("geez/" + item + "-" + str(book_iterator) + ".txt", "w+")
                    geeztable = geez_conetent[0].find('table').find_all('tr')
                    for each in geeztable:
                        geezverse = each.text
                        geezverse = geezverse.replace('\n', ' ')
                        geez_book.write(geezverse.strip() + "\n")
                        print(geezverse)
                else:
                    print("false")
                    break
            book_iterator += 1
if __name__ == '__main__':
    print (content_crawl())
```

Appendix III Python scripts used for removing only the first verse number

```
import codecs
import glob

amharic_path = "/home/tadesse/Desktop/Corpus/remove_frist_number/morphemebased/am.txt"
geez_path = "/home/tadesse/Desktop/Corpus/remove_frist_number/morphemebased/ge.txt"

def read_files(path):
    ss = ""
    files = glob.glob(path)
    for name in files:
        with open(name) as f:
            for line in f:
                ss += line + "\n"
    return ss
def write_to_file(fname, cont):

    ft = codecs.open(fname, 'w', 'utf-8')
    ft.write(cont)
    ft.close()
    print('cont written to %s ' % fname)
def remove_num(am_text):
    new_cont = ''
    for line in am_text.splitlines():
        cleaned = ' '.join(line.split()[1:])
        new_cont += cleaned + '\n'
    return new_cont
if __name__ == '__main__':

    cont = read_files(amharic_path)
    am_text = "\n".join([ll.rstrip() for ll in cont.splitlines() if ll.strip()])
    am_text = remove_num(am_text)
    write_to_file("am1.txt", am_text)
```

Appendix IV Python Script for Merging the Segemented Corpus of each Language in different file

```
import codecs

am = codecs.open("am.txt-segemented", "r", "utf-8")
ge = codecs.open("ge.txt-segemented", "r", "utf-8")


def merge_lines(am):
    count = ""
    line = " ".join([line.strip() for line in am])
    for i in line.split("።"):
        count += (i.strip()+"።"+"\n")
    return count


def write_to_file(fname, count):

    ft = codecs.open(fname, 'w', 'utf-8')
    ft.write(count)
    ft.close()
    print('cont written to %s ' % fname)


if __name__ == '__main__':

    count = merge_lines(ge)
    write_to_file('ge.txt', count)
    count = merge_lines(am)
    write_to_file('am.txt', count)
```

Appendix V Python Script for generating non prefix containing from the input corpus of Ge'ez Language

```
import codecs

am = codecs.open("ge1.txt", "r", "utf-8").read().split(" ")
root_words = codecs.open("rootwords_for_geez1.txt", "r", "utf-8").read().split(" ")

def read_from_fiee(am, root_words):
    ff = []
    for root in root_words:
        for word in am:
            if word.startswith(root):
                ff.append(word)
    return ff

def remove_duplcate(cc):
    final_list = ""
    for num in cc:
        if num not in final_list:
            final_list += num + "\n"
    return final_list

def delete_duplacate(xx):
    end = ""
    for word in xx.splitlines():
        if word in root_words:
            del word
        else:
            end = end + word + " " + "\n"
    return end

def write_to_file(fname, cc):
    ft = codecs.open(fname, 'w', 'utf-8')
    ft.write(cc)
    ft.close()
    print('cont written to %s ' % fname)

if __name__ == '__main__':
    cc = read_from_fiee(am, root_words)
    xx = remove_duplcate(cc)
    aa = delete_duplacate(xx)
    write_to_file("nonprefixwords.txt", aa)
```

Appendix VI Python Script for segmenting Prefix containing word lists from the input corpus of Amharic Language

```
import codecs

for l in range(1, 5):

    unsegemented = []
    nonprefixwords = []
    file_used = codecs.open("am" + str(l) + ".txt", "r+", "utf-8").read()
    root_words = codecs.open("rootwords.txt", "r", "utf-8").read()
    for root in root_words.split(" "):
        unsegemented.append(root)

    non_prefix_words = codecs.open("non_prefix_words.txt", "r", "utf-8").read()
    for non_prefix_word in non_prefix_words.split(" "):
        nonprefixwords.append(non_prefix_word)
    def prefix_segemntation(file):
        new_snt = ""
        prefix1 = ["የ", "ለ", "ይ", "አል", "በ", "እየ", "ሳይ", "አት", "አስ", "እንደ", "እስኪ", "ያል", "ባለ",
                   "እንዲ",  "እያስ", "በስተ", "ወደ", "ያስ", "ት" "ል", "ስለ", "እስክ", "ሲ", "እንድ"]
        prefix2 = ["አስ", "ምት", "በስተ", "ወደ", "ያለ", "ማይ", "የ", "ሳት"]
        prefix3 = ["ያስ", "እንዲ", " ት", "ያ", "አላ", "እስከ", "በ", "ተ"]
        prefix4 = ["ት", "ሚ", "እን", "በት", "ክ", "ተ", "ወ", "አይ", "የ"]
        sentence_list = file.split("\n")
        sentence_list.pop()
        for line in sentence_list:
            for word in line.split():
                flag = True
                if l == 1:
                    for pre in prefix1:
                        if word.startswith(pre) and len(word[len(pre):]) >= 2 and
                                word not in unsegemented and  word not in nonprefixwords:
                            new_snt = new_snt + pre + " " + word[len(pre): ] + "\n"
                            flag = False
                    if flag == True:
                        new_snt = new_snt + word + "\n"
                if l == 2:
                    for pre in prefix2:
                        if word.startswith(pre) and len(word[len(pre):]) >= 2 and
                                word not in unsegemented and word not in nonprefixwords:
                            new_snt = new_snt + pre + " " + word[len(pre): ] + "\n"
                            flag = False
                    if flag == True:
                        new_snt = new_snt + word + "\n"
```

```
            if l == 3:
                for pre in prefix3:
                    if word.startswith (pre) and len(word[len(pre):]) >= 2 and
                            word not in unsegemented and  word not in nonprefixwords:
                        new_snt = new_snt + pre + " " + word[len(pre): ] + "\n"
                        flag = False
                if flag == True:
                    new_snt = new_snt + word + "\n"
            if l == 4:
                for pre in prefix4:
                    if word.startswith(pre) and len(word[len(pre):]) >= 2 and
                            word not in unsegemented and  word not in nonprefixwords:
                        new_snt = new_snt + pre + " " + word[len (pre):] + "\n"
                        flag = False
                if flag == True:
                    new_snt = new_snt + word + "\n"
    return new_snt

def write_to_file(fname, count):

    ft = codecs.open(fname, 'w', 'utf-8')
    ft.write(count)
    ft.close()
    print('cont written to %s ' % fname)
if __name__ == '__main__':
    co = prefix_segemntation (file_used)
    j = l + 1
    write_to_file ("am" + str(j) + ".txt", co
```

Appendix VII Python Script for segmenting Suffix containing word lists from the input corpus of Amharic Language

```
import codecs

for l in range(1, 5):
    file_used = codecs.open("am" + str(l) + ".txt", "r+", "utf-8").read()
    unsegemented = []
    root_words = codecs.open("rootwords.txt", "r", "utf-8").read()
    for root in root_words.split(" "):
        unsegemented.append(root)

    def sufix_segemntation(file):
        new_snt = ""
        suffix1 = ["ን", "ና", "ሽ", "ነት", "ቸው", "ህ", "ባት", "ኞች", "ዋ", "ችኋል", "ዎች", "ለህ", "ም", "ለን", "ለት", "ዊ"]
        suffix2 = ["ቹ", "ውያን", "ዎች", "ዋ", "ኝ", "ኞች", "ያ", "ችን", "ቸው"]
        suffix3 = ["ች", "ቸው" "ዊ", "በት", "ችሁ", "ዋ"]
        suffix4 = ["ኛ", "አቸዋል", "ቹ", "ችሁ", "ውያን", "ቻቸው", " ይ", "ቸው", "ህ", "ኞች", "ለ", "ት"]

        sentence_list = file.split("\n")
        for line in sentence_list:
            for word in line.split():
                flag = True
                if l == 1:
                    for su in suffix1:
                        if word.endswith(su) and len(word[0:len(word) - len(su)]) >= 2 and word not in unsegemented:
                            new_snt = new_snt + word[0: len(word) - len(su)] + " " + word[len(word) - len(su):] + "\n"
                            flag = False
                    if flag == True:
                        new_snt = new_snt + word + "\n"
                if l == 2:
                    for su in suffix2:
                        if word.endswith(su) and len(word[0:len(word) - len(su)]) >= 3 and word not in unsegemented:
                            new_snt = new_snt + word[0: len(word) - len(su)] + " " + word[len(word) -
len(su):] + "\n"
                            flag = False
                    if flag == True:
                        new_snt = new_snt + word + "\n"
                if l == 3:
                    for su in suffix3:
                        if word.endswith(su) and len(word[0:len(word) - len(su)]) >= 2 and word not in unsegemented:
                            new_snt = new_snt + word[0: len(word) - len(su)] + " " + word[len(word) - len(su):] + "\n"
                            flag = False
                    if flag == True:
                        new_snt = new_snt + word + "\n"


                if l == 4:
                    for su in suffix4:
```

```
                    if word.endswith(su) and len(word[0:len(word) - len(su)]) >= 2 and word not in unsegemented:
                        new_snt = new_snt + word[0: len(word) - len(su)] + " " + word[len(word) - len(su):] + "\n"
                    flag = False
            if flag == True:
                new_snt = new_snt + word + "\n"

    return new_snt

def write_to_file(fname, count):

    ft = codecs.open(fname, 'w', 'utf-8')
    ft.write(count)
    ft.close()
    print('cont written to %s ' % fname)

if __name__ == '__main__':
    co = sufix_segemntation(file_used)
    j = l + 1
    write_to_file("am" + str(j) + ".txt", co)
```

Appendix VIII: Prefixes and Suffixes used from Ge'ez and Amahric Language

```
import codecs

lang1_file =  'am1.txt'
lang2_file =  'ge1.txt'
lang1 = codecs.open(lang1_file,'r','utf-8').read()
lang2 = codecs.open(lang2_file,'r','utf-8').read()

def to_dic(lang):
    dic = {}
    for count, el in enumerate(lang):
        dic[count] = el
    return dic

def write_to_file(fname,cont):
    fn = codecs.open(fname,'w','utf-8')
    fn.write(cont)
    fn.close()

def remove_repeatet(dic1, dic2):
    repeated_count = 0
    lang1_cont = ''
    lang2_cont = ''
    dic3 = {}; dic4 = {}
    for k, v in dic1.items():
        if v not in dic3.values():
            dic3[k] = v
            dic4[k] = dic2[k]
        else:
            if dic2[k] not in dic4.values():
                dic3[k] = v
                dic4[k] = dic2[k]
            else:
                repeated_count += 1
    for k,v in dic3.items():
        lang1_cont += v + '\n'
        lang2_cont += dic4[k] + '\n'

    print('%d sentence repeated ' % repeated_count)
    write_to_file(lang1_file + '_pr.txt', lang1_cont)
    write_to_file(lang2_file + '_pr.txt', lang2_cont)

if __name__ == '__main__':
    lang1 = lang1.splitlines()
    lang2 = lang2.splitlines()
    dic1 = to_dic(lang1)
    dic2 = to_dic(lang2)
    remove_repeatet(dic1, dic2)
```

Appendix IX: Sample of word level aligned corpus

Geez	Amharic
ብክይዎ ወላህውዎ እለ ታፈቅርዎ ።	የምትወዱት ሰዎች ፈጽሞ አልቅሱለት ።
ዬ ዬ ዬ ክርስቶስ ንጉሥነ ።	ወየው ወየው ወየው ንጉሳችን ክርስቶስ ።
ቅዱስ ሥሉስ እግዚአብሔር ሕያው ተሣሃለነ ።	ልዩ ሦስት ሕያው እግዚአብሔር ሆይ ይቅር በለን ።
ፈኑ ጸጋ መንፈስ ቅዱስ ላዕሌነ ።	የመንፈስ ቅዱስን ጸጋ ላክልን ።
ተንሥኡ ለጸሎት ።	ለጸሎት ተነሡ ።
እግዚኦ ተሣሃለነ ።	አቤቱ ይቅር በለን ።
ሰላም ኵልክሙ ።	እግዚአብሔር ከሁላችሁ ጋር ይሁን ።
ምስለ መንፈስከ ።	ከመንፈስህ ጋራ ።
አንቲ ውእቱ ንጽሕት እምንጹሓን ።	ከንጹሓን ይልቅ ንጽሕት የሆንሽ አንቺ ነሽ ።
ተፈሥሒ አ ገነት ነባቢት ማኅደሩ ለክርስቶስ ዘኮነ ዳግማይ አዳም በእንተ አዳም ቀዳሚ ብእሲ ።	ስለቀደመ ሰው አዳም ሁለተኛ አዳም የሆነ የክርስቶስ ማደሪያው የምትናገሪ ገነት ሆይ ደስ ይበልሽ ።
ወዘእምነገደ ይሁዳ ካሌብ ወልደ ዬፎኔ ።	ከይሁዳ ነገድ የዮፎኒ ልጅ ካሌብ ።
ኢትቅትል ።	አትግደል ።
ኢትዘሙ ።	አታመንዝር ።
ኢትስርቅ ።	አትስረቅ ።
ዐርገ እግዚአብሔር በይባቤ ወእግዚእነ በቃለ ቀርን ።	አምላክ በእልልታ ወደ ሰማይ ወጣ እግዚአብሔር በመለከት ድምፅ ዐረገ ።
ርእዩከ ማያት እግዚኦ ርእዩከ ማያት ወፈርሁ ።	አቤቱ ውኆች አዩህ ውኆችም አይተውህ ፈሩ ።
ዝክረ ጻድቅ ለዓለም ይሄሉ ወኢይፈርህ እምነገር እኩይ ።	የጻድቅ መታሰቢያ ለዘላለም ይኖራል ከክፉ ነገር አይፈራም ።
ስምዐኒ አምላኪየ ስእለትየ ወአፅምአኒ ጸሎትየ ።	አምላክ ሆይ ልመናዬን ስማ ጸሎቴንም አድምጥ ።
ወይቤለኒ እግዚአብሔር ።	እግዚአብሔርም እንዲህ ብሎ ተናገረኝ ።

Appendix X: Sample of morpheme level aligned corpus segmented using morfessor

Geez	Amharic
ብከይዎ ወ ላህውዎ እለ ታፈቅር ዎ።	የምት ወዱት ሰዎች ፈጽሞ አል ቅሱ ለት።
ዬ ዬ ዬ ክርስቶስ ንጉሥ ነ።	ወ የ ው ወ የ ው ወ የ ው ንጉሳ ችን ክርስቶስ።
ቅዱስ ሥሉስ እግዚአብሔር ሕያው ተሣሃለ ነ።	ልዩ ሦስት ሕያው እግዚአብሔር ሆ ይ ይ ቅር በ ለ ን።
ፈኑ ጸጋ መንፈስ ቅዱስ ላዕሌነ።	የ መንፈስ ቅዱስ ን ጸጋ ላክ ልን።
ተ ንሥኡ ለ ጸሎት።	ለ ጸሎት ተነሡ።
እግዚአ ተሣሃለ ነ።	አ ቤቱ ይ ቅር በ ለ ን።
ሰላም ኵልክሙ።	እግዚአብሔር ከ ሁ ላችሁ ጋር ይሁን።
ምስለ መንፈስ ከ።	ከ መንፈስ ህ ጋራ።
እለ ትነብሩ ተ ንሥኡ።	የ ተቀመጣችሁ ተነሡ።
ን ነጽር።	እና ስተውል።
ኢ ትዝክር ለ ነ አበሳ ነ ዘ ትካት ፍጡነ ይርከበነ ሣህልክ እግዚአ።	የ ቀደመ በደላ ችንን አታስብ ብን አ ቤቱ ይቅርታ ህ ፈጥኖ ይደረግ ልን።
ጎቤክ ንጸርሕ ጎቤክ ነዐ ወ ዩ ጎቤክ ንትመሀለል ለ ዓለም ዓለም።	ወደ አንተ እን ጮህለን ወደ አንተ እና ለቅ ሳለን ወደ አንተ እን ማለ ላለን ለዘ ለ አለሙ።
ብርሃን ዘ በአማን ዘ ያበርህ ለ ኩሉ ሰብእ።	በ ዚህ ዓለም ለሚኖሩ ሰዎች ሁሉ የ ም ታበራ ዕውነ ተኛ ብርሃን።
ወ ኮነ ላዕሌሆሙ መንሱተ መዐቱ ለ እግዚአብሔር ወ ሖሩ።	እግዚአብሔር ም ተቈጥቶ ባቸው ሄደ
ወለ ምንት ትቀውሙ ዲበ ትዕይን ቱ ለ እግዚአብሔር።	በ እግዚአብሔር ም ጉባኤ ላይ ለምን ትታ በ ያ ላችሁ ? አሉ።
ወ ሶበ ሰምዐ ሙሴ ወድቀ በ ገጹ።	ሙሴ ም በ ሰማ ጊዜ በ ግምባሩ ወደቀ።
እስመ ሕዝብ ቅዱስ አንተ ለ እግዚአብሔር አምላክ ከ።	ለ አምላክህ ለ እግዚአብሔር አንተ ቅዱስ ሕዝብ ነህ ና።
ወ ኢሰማዕ ክሙ ቃልየ አመ ገበርክሙ ዘንተ።	እናንተ ግን ቃሌ ን አልሰማችሁም።

Appendix XI: Sample of morpheme level aligned corpus segmented using rule based

Geez	Amharic
ብክይ ዎ ወ ላህው ዎ እለ ታ ፈቅር ዎ።	የ ምት ወዱት ሰዎ ች ፈጽሞ አ ልቅሱ ለት ።
ዬ ዬ ዬ ክርስቶስ ንጉሥ ነ።	ወየው ወየው ወየው ንጉሳ ችን ክርስቶስ ።
ቅዱስ ሥሉስ እግዚአብሔር ሕያው ተሣሃለ ነ።	ልዩ ሦስት ሕያው እግዚአብሔር ሆይ ይቅር በለ ን።
ፈኑ ጸጋ መንፈስ ቅዱስ ላዕሌ ነ።	የ መንፈስ ቅዱስ ን ጸጋ ላክ ልን።
ተ ንሥኡ ለ ጸሎት።	ለ ጸሎት ተነሡ።
እግዚኦ ተሣሃለ ነ።	አቤቱ ይቅር በለን።
ሰላም ኵል ክሙ።	እግዚአብሔር ከ ሁላ ችሁ ጋር ይሁን።
ምስለ መንፈስ ከ።	ከ መንፈስ ህ ጋራ።
እለ ትነብሩ ተ ንሥኡ።	የ ተቀመጣችሁ ተነሡ።
ን ነጽር።	እና ስተውል።
ኢ ትዝክር ለ ነ አበሳ ነ ዘ ትካት ፍጡነ ይርከበነ ሣህልከ እግዚኦ።	የ ቀደመ በደላ ችን ን አታ ስብብ ን አቤቱ ይቅርታ ህ ፈጥኖ ይ ደረግል ን።
ኀቤከ ን ጸርሕ ኀቤከ ነዐ ወዩ ኀቤከ ን ት መህለል ለ ዓለም ዓለም።	ወደ አንተ እን ጮህለን ወደ አንተ እና ለቅ ሳለን ወደ አንተ እን ማለ ላለን ለዘ ለ አለሙ።
ብርሃን ዘ በ አማን ዘ ያ በርህ ለ ኩሉ ሰብእ።	በዚህ ዓለም ለ ሚ ኖሩ ሰዎ ች ሁሉ የ ም ታበራ ዕውነተ ኛ ብርሃን።
ወ ኮነ ላዕሌሆሙ መንሱተ መዐቱ ለ እግዚአብሔር ወ ሖሩ።	እግዚአብሔር ም ተቈጥቶ ባቸው ሄደ
ወለ ምንት ትቀውሙ ዲበ ትዕይን ቱ ለ እግዚአብሔር።	በ እግዚአብሔር ም ጉባኤ ላይ ለምን ትታ በ ያ ላችሁ ? አሉ።
ወ ሶበ ሰምዐ ሙሴ ወድቀ በ ገጹ።	ሙሴ ም በ ሰማ ጊዜ በ ግምባሩ ወደቀ።
እስመ ሕዝብ ቅዱስ አንተ ለ እግዚአብሔር አምላክ ከ።	ለ አምላክ ህ ለ እግዚአብሔር አንተ ቅዱስ ሕዝብ ነህ ና።
ወ ኢ ሰማዕ ክሙ ቃል የ አመ ገበር ክሙ ዘንተ።	እናንተ ግን ቃሌ ን አል ሰማ ችሁ ም።

Appendix XII: Lists of University that Teach Ge'ez as Course

Ehtiopian

- ✓ Addis Ababa University
- ✓ Bahir Dar University
- ✓ Dabra Markos University
- ✓ Holy Trinity Theological College in Ethiopia
- ✓ Mekelle University

United States of American

Abilene Christian University

Cambridge University Faculty of Divinity

Catholic University

Florida State University

Frei University Berlin

Göttingen University

Hamburg University

Heidelberg University

Ludwig-Maximilians-Universität München

Oriental University Naples Paris, Institute Catholique, ELCOA

Philipps-Universität Marburg

Pontifical Oriental Institute in Rome

Russian State University of Humanities (Moscow)

Saint Mary Theological College and Ethio-American Cultural Institute, Houston, Texas, online learning,

SOAS, University of London

St Petersburg University

St Tichon University in Moscow

University of Chicago,

University of Texas, Austin,

University of Toronto,

University of Vienna,

University of Washington

Uppsala University

Appendix XIII: Prefixes (a) and Suffixes (b) used for Ge'ez Language

Prefix type	Prefix lists
Prefix1	**["አስተ", "ለ", "ወ", "በ", "ያስተ", "ን", "ኢ", "የ", "እም", "ስ"]**
Prefix2	**["ዘ", "ያስተ", "ን", "ኢ", "የ", "እም", "በ", "ለ", "አስተ"]**
Prefix3	**["ት", "ለ", "ተ", "ለ", "ያስተ", "ን", "ኢ", "የ", "እም", "አ", "ይ", "አስተ"]**
Prefix4	**["ት", "ለ", "ተ", "ን", "ኢ", "የ", "እ", "በ"]**

(a)

Suffix Type	Suffix lists
Suffix1	**["ክሙ", "ኩክ", "ኩኪ", "ክዎ", "ክዋ", "ክን", "ኒ"]**
Suffix2	**["ሁ", "ክ", "ኩ", "ሆሙ", "ክዋ", "ክሙ", "ዎሙ", "ሂ", "ዎ"]**
Suffix3	**["ነ", "ኒ ", "ሙ", "ሆን", "ስ", "ዎን"]**
Suffix4	**["ኒሃ", "ክምዎ", "ዎ", "ኪ", "ቶን", "ሆሙ"]**

(b)

Appendix XIV: Prefixes (a) and Suffixes (b) used for Amahric Language

Prefix type	Prefix lists
Prefix1	["የ", "ለ", "ይ", "አል", "በ", "እየ", "ሳይ", "አት", "አስ", "እንደ", "እስኪ", "ያል", "ባለ", "እንዲ", "እያስ", "በስተ", "ወደ", "ያስ", "ት", "ል", "ስለ", "እስክ", "ሲ", "እንድ"]
Prefix2	["አስ", "ምት", "በስተ", "ወደ", "ያለ", "ማይ", "የ", "ሳት"]
Prefix3	["ያስ", "እንዲ", " ት", "ያ", "አላ", "እስከ", "በ", "ተ"]
Prefix4	["ት", "ሚ", "እን", "በት", "ከ", "ተ", "ወ", "አይ", "የ"]

(a)

Suffix Type	Suffix lists
Suffix1	**["ን", "ና", "ሽ", "ነት", "ቸው", "ህ", "ባት", "ኞች", "ዋ", "ችኋል", "ዎች", "ም", "ለን", "ለት", "ዊ"]**
Suffix2	**["ቹ", "ውያን", "ዎች", "ዋ", "ኛ", "ኞች", "ያ", "ችን", "ቸው"]**
Suffix3	**["ች", "ቸው" "ዊ", "በት", "ችሁ", "ዋ", "ን",, "ህ",]**
Suffix4	**["ኛ", "አቸዋል", "ቹ", "ችሁ", "ውያን", "ቻቸው", " ይ", "ቸው", "ህ", "ኞች", "ለ", "ት"]**

(b)

Declaration

I declare that this research is my original work and has not been presented for a degree in any University, and that all sources of material used for the research have been properly acknowledged.

Declared by:

Name: Tadesse Kassa

Signature: _____

This research has been submitted for Examination with my approval as University advisor.

Name: Million Meshesha (PhD), Advisor

Signature: _____

Date: _____

Addis Ababa, Ethiopia

October, 2018