A Hybrid approach for Machine Translation from Ge'ez to Amharic language

A Thesis Presented by

Samson Tesfaye

to

The faculty of Informatics

of

St. Mary's University

In Partial Fulfillment of the Requirements

for the Degree of Master of Science

in

Computer Science

February 22, 2020

ACCEPTANCE

A Hybrid approach for Machine Translation from

Ge'ez to Amharic language

By

Samson Tesfaye

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

_____

Internal Examiner
{Full Name, Signature and Date}

_____

External Examiner
{Full Name, Signature and Date}

_____

Dean, Faculty of Informatics
{Full Name, Signature and Date}

February 08, 2020

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

_____

Full Name of Student

_____

Signature

Addis Ababa

Ethiopia

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

_____

Full Name of Advisor

_____

Signature

Addis Ababa

Ethiopia

February 08, 2020

Acknowledgments

First, I thank the almighty God for keeping me in my entire life and gave me all what I needed in my life on his good will. I would like to express my gratitude to all the people who supported and accompanied me when I carried out this thesis.

I would like to express my sincere gratitude to my advisor Dr. Michael Melese for the continuous support of my master's research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for this study.

I would also like to thank the Ethiopian Orthodox Tewahedo Church for teaching me Ge'ez language and laid a foundation of this research and experts of the languages who were involved in checking the correctness of the parallel corpus and POS tagging work for this thesis: Mr. Haile Eyesus Abebe & Dawit Medhin.

Last but not the least; I would like to give a special thanks to my family: my mother Azmera Kiros and my brother Ermias Tesfaye for initiating and supporting me from registration for the master's program up to this step.

Table of Contents

List of Acronyms and Abbreviations

A	Amharic
ALPAC	Automatic Language Processing Advisory Committee
BLEU	Bilingual Language Evaluation Understudy
EOTC	Ethiopian Orthodox Tewahedo Church
G	Ge'ez
IBM	International Business Machines
MT	Machine Translation
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NMT	Neural Machine Translation
OCR	Optical Character Resolution
OSV	Object-Subject-Verb
OVS	Object-Verb-Subject
POS	Part of Speech
rG	reordered Ge'ez
RBMT	Rule Based Machine Translation
SOV	Subject-Object-Verb
SVO	Subject-Verb-Object
VSO	Verb-Subject-Object
WSD	Word Sense Disambiguation

List of Tables

List of Figures

List of Algorithms

Abstract

Natural Language Processing can be applied in different areas. From these areas, Machine Translation is the one and its concern is to translate one natural language in the form of text or speech into another language. Human translation has positive sides as far as language translation concerned but it has also its own limitations like slowness when translating than machines, correctness and precision of the texts or speech that are being translated, it has some delays in the Process of translation and it is time and cost consuming. To overcome the problem, many studies have been conducted. Our study, Ge'ez to Amharic machine translation using a hybrid approach, is one of these. Hybrid in this case means using the best features of statistical and rule-based machine translation approaches. Even though Ge'ez and Amharic are the Semitic language family, they have a structural difference in sentence construction. To rectify this issue, in this study we proposed a reordering approach in syntax to make the source language to have a similar sentence structure with the target language. During our research, the source and target languages are Ge'ez and Amharic respectively. There is no prior study conducted on this specific title as the researcher knowledge concerned. We start our study by collecting data from different resources. Unfortunately, our data is only from spiritual books since nowadays Ge'ez language is limited in EOTC literatures. After collecting the data and passing through the preprocessing step, we classified it into two data sets of training and testing. Reordering rules are drafted and applied on the data before classifying. Since our developed machine translation system is unidirectional and the target language is Amharic, we built our language model on it. Translation model which works on probability to generate a target language sentence from a given source language sentence also built and decoder is used to search the best sequence of translation probability. Finally, we conducted four experiments with two different approaches and evaluate the results obtained accordingly. The first and second experiments are performed by the statistical approach by changing the percent of training and testing data then we get a BLEU score of 7.36% and 7.15%. The third and fourth experiments are carried out by hybrid approach in a similar fashion and we get a BLEU score result of 18.62% and 17.38%. Thus, from these we conclude that using a hybrid approach by combining statistical with rule-based machine translation approaches provides a better result for machine translation from Ge'ez to Amharic language.

Keywords: Statistical Machine Translation, Hybrid Machine Translation, Reordering rule

Chapter One

Introduction

1.1. Background

In our day-to-day interactions, we use means of communication [1]. Language is one of the representatives from a different, collective ways of those communications. Human beings have a symbolic mind that is capable of language and not shared by other species. This language knows as human language. As the technology shown a progress, there comes a need to teach computers the human language. The field, Natural Language Processing (NLP) that is the subfield of Artificial Intelligence (Machine Learning) is responsible for this task.

Natural language processing is a theory-motivated range of computation techniques for the automatic analysis and representation of human language [2]. In the natural flow of the NLP, the focus starts from one problem, and heading to another problem. This is because solving the first problem depends on solving the second problem but sometimes the second problem is easy to trace than the first problem, or the second problem got more market interest than the first problem. Since 1950s, there are remarkable progresses in NLP on how to do it and in doing it [3]. NLP applied in different application areas these includes Optical Character Resolution (OCR), Word Sense Disambiguation (WSD), Machine Translation (MT), Text-proofing and Part-of-speech tagging (POS).

One of the applications of NLP, Machine Translation (MT), is a branch of computational linguistics that investigates the use of computers in translating text or speech from one natural language (source language) into another (target language) [4]. At first, a MT system analyzes the input text and creates and internal representation to it. Then, the representation manipulated and transferred to a form suitable for the target language. Finally, the output generated in the target language. There are different approaches available for machine translation namely, statistical, rule-based, example based and hybrid machine translation. In this study, we use hybrid MT approach (statistical and rule-based). Statistical MT uses a probability method to give a best translation while rule-based works by linking the structure of the given input sentence with the structure of demanded output sentence. Hybrid uses the good features of both approaches.

In Ethiopia, there are more than 80 different languages spoken by different ethnic groups. The languages are from different language families. One of the families, Semitic language, contains many Ethiopian and out of Ethiopian languages. From those Ethiopian languages, Ge'ez and Amharic mentioned. Ge'ez was widely spoken in Ethiopia and Eritrea until the 10^{th} to 12^{th} centuries and it uses Ge'ez (Ethiopic) script for its writing system. Nowadays the language has limited number of users around churches and monasteries but there are many literature and spiritual books written on the language and its script (Ethiopic) is widely used in the languages: Tigré (spoken in Eritrea), Amharic and Tigrinya. The script called *'fidel (ፊደል)'*, which is to mean '*alphabet'* [5].

Ge'ez is preferred because as mentioned above there are many useful knowledges that needs to transfer through generations in order to know and exploit well the wisdoms in those books. Amharic language currently is a widely used federal language of Ethiopia and the mother tongue language for many people in the country [6]. Developing a system that translates Ge'ez words/phrases into Amharic used as a bridge for knowledge transfer and conducting a further research. This study uses a hybrid translation approach with some guiding rules to govern the translation from Ge'ez to Amharic language.

1.2. Statement of the Problem

Machine translation is one of the most widely used and developed application of NLP. There have been many studies conducted regarding with machine translation for different foreign languages. As English is the worlds' most spoken language, it is easy to expect that most of the studies to be circled around this language. There are many published papers for the machine translation of other resourced languages in pairing with the English language [7, 8, 9]. Although there are also different studies for machine translation in our country by taking the English language and one from the local spoken languages like [10, 11, 12, 13], it is not as plenty as required since we have over 80 different languages.

Similarly, there is a lack of enough studies on machine translation system between different languages spoken in Ethiopia. However, there were some efforts to fill this gap [14, 15, 16]. As a multi Nations & Nationalities country, we need to have some automated language translation system. The system helps in knowing one nation's culture, thoughts, beliefs, democracy practices (as in *Geda* system), social and cultural heritages and facilitates each other's relationship. The last but not the least thing developing this kind of system gives is in creating the concept of

nationalism, making citizens to stand for their countries sovereignty, since language plays a major role in uniting or separating peoples in a country.

As we speak of Ethiopia, we do not skip mentioning its ancient civilization specially the language that was used (both the spoken and the one used for writing for different purposes). Ge'ez was once the most widely used language in Ethiopia. The people at that time used it as their main language and they wrote many books (religious and others) [5]. Unfortunately, its speakers become limited in number at this time and many of these speakers used it for religious purpose. The good thing is still there is a chance to know about the language because we do have many literatures written on it that are located at Ethiopian Orthodox Churches, heritage preservation authorities in Ethiopia and in other countries. In addition, different local and foreign universities opened a department to teach the Ge'ez language and give a course for the students who enrolled for it. The interesting thing is most of the universities are out of Ethiopia.

Besides this, there is a gap in teaching the language using a computer system since the Ge'ez translation mostly done manually that in turn have a problem of time consumption, lack of accuracy and conciseness, lack of knowledge of the topic and depends on linguistics knowledge of the translator. Developing a machine translation system will rectify the problem in some way since it opens the door to know more about the Ge'ez language. Many literatures written in Ge'ez languages that are helpful to know on what we are standing as history writes the language to be part of us and to move forward to develop our country in various aspects.

Amharic language is selected because as mentioned, it is a widely used federal language of Ethiopia and it is the mother tongue language for many people in the country [6]. The study uses a hybrid machine translation system that is a combination of statistical and rule-based because statistical uses a probability method that concentrates on alignment of words and rule-based uses some guiding rules for efficient translation. Therefore, using these best features of the two approaches, we build an efficient translation system [4]. There are prior studies on Ge'ez to Amharic language, but their developed system efficiency is less that is why we are aiming to develop a system which provide a better result. Due to a difference in sentence structure of Ge'ez and Amharic languages, there is a guiding rule to keep track of the translation. As the researcher knowledge concerned, there is no prior study conducted on this specific title.

So, our study aims to answer the following research questions:

- Does the hybrid machine translation approach give a better result when combining the statistical and rule-based machine translation approaches?
- How much the hybrid machine translation approach improves the system performance as compared with previous studies?

1.3. Objective

1.3.1. General Objective

The general objective of this study is to design and develop Ge'ez to Amharic machine translation system using a hybrid approach.

1.3.2. Specific Objectives

The specific objectives of this research are:

- To review related literature and state of the art in machine translation,
- To identify both languages linguistic behavior,
- To find out the syntactic relationship between the two languages,
- To collect and prepared parallel corpus,
- To design the general architecture of the system,
- To conduct an experiment and examine the result,
- To evaluate the performance of the translation model,
- To report the finding between Ge'ez and Amharic languages translation and
- To show the importance of hybrid machine translation approach over a single machine translation approach

1.4. Methodology

On this study, we used the quantitative experimental research methodology. It's suitable to find impermanent relationships and let the researchers to investigate the possible cause-effect relationship by manipulating independent variables to influence the dependent variables. In this section we discuss the methodologies we used for setting up the experiment in detail.

1.4.1. Literature Review

To conduct this research, published papers, books, articles and other related sources that counted as secondary data used in this study. The aim of using these literature review and related articles

is to have a better knowledge of the problem area and to show the gap and the importance of the study. Furthermore, different machine translation systems for different languages using different approaches and the linguistic behaviors of Ge'ez and Amharic discussed.

1.4.2. Data Collection

Ge'ez-Amharic parallel corpus collected from different sources like the Holy Bible and Wudasie Mariam and Metsehafe Kidase. In this study, we followed the POS tagging mechanisms by means of POS tag sets to re-order the Ge'ez words sentence structure since Ge'ez and Amharic languages have different structure in sentence formation. Since nowadays Ge'ez language is limited in EOTC literatures we face a data scarcity but with minimum data we can get a better result because tagging of each word makes the translation smooth. Two sampling techniques, Convenience and Random sampling, are used. Convenience sampling is one of the types of non-probability sampling where the sample taken from a group of people in which there is no pre-defined rule that govern who to select and it just needs the willingness of the selected people. We applied the convenience sampling in preparing the training and test data sets for experiment. Random sampling on the other hand is a type of probability sampling where all the participants of the sampling have an equally likely opportunity to select with random selection. We used this sampling in collecting the overall data.

1.4.3. Tools and Techniques

The following tools are used for developing this Machine Translation system:

- SRILM toolkit, for language modelling since it consists ready-made set of tools for state-of-the-art for language modelling
- MGIZA, for translation model and word alignment tool. It's powerful tool and best suited for multi-core machines
- Moses, a statistical machine translation system that takes the language and the translation model for translation from one language into other.
- Python programming language
- Ubuntu from version 16 and above
- BLEU (Bilingual Evaluation Understudy) score, to evaluate the MT system. It achieves a high correlation with reference translation
- Notepad, to organize the collected corpus in an easy way
- Microsoft office 2016, for the documentation of the study

1.4.4. Experiment and Evaluation

To make sure that the proposed system meets its design goals, we conduct four different experiments for the two machine translation approaches. Two experiments for statistical machine translation and two for hybrid machine translation approach by making the training and testing data sets 90, 80 and 10, 20 percent respectively. The minimum numbers indicate the percent of the testing data and the maximum ones represent the percent of the training data from the overall collected data. There are two techniques to evaluate a system: Manual (by some person) and Automatic. However, manual system is time consuming. Due to this, the system developed after the accomplishment of this study uses automatic evaluation using BLEU score mechanism.

1.5. Scope and Limitations of the Study

The system, machine translation from Ge'ez to Amharic using hybrid approach, is designed to give a translation from Ge'ez to Amharic language at word, phrase or sentence level based on user's choice and it only accepts inputs which are on a written format i.e. speech translation is not included in this study. The other point is this study is a unidirectional not a bidirectional means that it supports only translations from Ge'ez to Amharic; we do not include translations from Amharic to Ge'ez. The reordering rules are drafted by studying the sentence structure of the language, so these rules include most of the Ge'ez sentence structure. The system uses a hybrid approach and our experiments are based on first the statistical machine translation approach and second on the hybrid approach (statistical and rule-based). The lack of enough bilingual parallel corpora for the study, use of only spiritual books like bible since now Ge'ez language is limited in the literatures of EOTC and absence of publicly available Ge'ez part-of-speech tagger is the main challenge throughout the process this study.

1.6. Significance of the Study

The study helps to translate words, phrases and sentences written on Ge'ez language into its relative of the Semitic language family, Amharic. It also gives a way for those who have an enthusiasm on Ge'ez language and encourages others who do not have any opportunity or interest to know the language well and to read and write whatever they like. It can serve as an alternative learning-teaching tool for universities, research academies, and other related institutions. Anyone who want to read and make a research on ancient documents and ancient Ethiopian Orthodox Church and Ethiopian history will get enough support from this study.

Interested people can use the translation system after post editing it. Finally, it is helpful for other following research on Ge'ez language or related.

1.7. Beneficiary of the Research

Universities, research academies and other related institutions may benefit from this research since it gives a basic knowledge of the language. The study has also great impact for Ethiopian Orthodox Church in a way of addressing followers of the religion since many people does not understand the Ge'ez language very well. Overall, anyone who is keen to know the language gets a better support from the study.

1.8. Organization of the Thesis

This research study has six different chapters. The first chapter gives an overview of Ge'ez and Amharic languages, states the problem, discusses the specific and general objectives, methodology, scope and limitations, significance and beneficiary of the study. Chapter two discusses reviews of literatures which includes Amharic and Ge'ez languages. In addition to these, the details of rule-based, statistical and hybrid approaches of machine translation are discussed. The third chapter mainly discuss about the related works regarding to NLP specifically machine translation. The fourth chapter gives a detail information on the architecture of the proposed system for Ge'ez to Amharic language machine translation system using a hybrid approach. The fifth chapter mainly deals with preparation data, preprocessing and experiments. Finally, chapter six provides conclusion and future works.

Chapter Two

Literature Review

2.1 Overview

In this chapter, we go through the detail information of the two languages from different literatures point of view. Primarily, section 2.2 discuss about the language Ge'ez and Amharic. Then section 2.3 discuss about the linguistics relationships between Ge'ez and Amharic language taking the writing system and syntax. Different machine translation approaches have their portion on this discussion.

2.2 Ge'ez and Amharic languages

Ge'ez is an ancient south Semitic language of Ethiopia and Eritrea in the horn of Africa and becomes Aksum's civilization kingdom official language [17]. Ge'ez is still the liturgical language of EOTC since the early 4th century, Ethiopian Catholic Church and Beta Israel Jewish community of Ethiopia. Despite its' speakers becomes less in number around 13th century, it maintains as the primary written language of Ethiopia up to 20th century. Religious and secular writings are included in the literature list of Ge'ez language [18]. As [17] describes Amharic is the second most spoken Semitic language in the world after Arabic language and the second largest language in Ethiopia after Affan Oromo. Currently Ge'ez has no native speakers, but Amharic language is the official working language of the government of Ethiopia and it has above 30 million native and non-native speakers. Since 14th century, manuscripts for this language prepared and after 19th century, it becomes the general medium of literatures, journalism, education, and communication.

In Ge'ez script, a character represents a consonant and a vowel combination this makes the language alpha syllabary script or "Abugida" [15, 16, 18]. In Abugida a character represents one sound either it's consonant or vowel. Amharic inherits its alphabet scripts from Ge'ez and uses an alpha syllabary writing system in which a single symbol is formed with a combination of consonant and vowel. This makes a person read and write Ge'ez and Amharic easily after knowing the alphabets. There are 26 and 34 basic alphabets ('Fidel' in Amharic) in Ge'ez and Amharic script. Each of the basic alphabets have seven forms created by combining the basic letters with vowels this produces 182 and 238 unique characters respectively and there are other

additional forms that are derived from the basic alphabets like ቈ ቍ ቊ ቋ ቌ which is derived from ቀ, ኰ ኵ ኲ ኳ ኴ from ከ, ጐ ጕ ጒ ጓ ጔ from ጎ and ጘ ጙ ጚ ጛ ጜ from ገ.

Both the languages, Ge'ez and Amharic have complex morphology. If we take the word formation as instance, it has different formations including prefixation, infixation, suffixation, and reduplication. Conjunctions, Prepositions, Article, Pronominal affixes, Negation markers are bound morphemes that attached to the content words that produces complex words consisting of several morphemes [15, 18]. The other characteristic of morphologically complex languages is they show the correspondence between the syntactic part of a sentence like nouns, verbs, person, number, gender, fine and place this impacts the complexity of word generation. In addition, they follow different syntactic structure. Amharic usually uses the Subject-Object-Verb (SOV) form while Ge'ez has a free order of sentence structure but usually falls on SVO, VSO and SOV [16].

2.3 Linguistic Relationships of Ge'ez and Amharic

2.3.1 Writing systems

Writing is a way to represent a specific language in a visual or more understandable form. Symbols used to represent the sounds of speech and punctuations and numerals. Based on studies, there are six different types of writing systems [19, 20]. These alphabets are (English, Russian, and Greek), Abjads (Arabic, Hebrew), Abugidas or alpha syllabaries (Devanagari, Thai, Ge'ez, Amharic), Featural alphabets (Hangul), syllabaries (Japanese, Cherokee), and Logographic (like Chinese). From these, an Abjad or Abugida used for Ge'ez language. Until 330 A.D, The Abjad, which have 26 consonantal letters, were used and vowels were not indicated [21]. Abugida developed by the influence of Christian scripture by adding a must vocalic diacritic to the consonantal letters. The vowels, e, a, i, o, u, diacritics were combined with the consonants in a recognizable and slightly irregular way [22]. Before an Egyptian born and the first patriarch of Ethiopian Orthodox Tewahedo Church Aba Fremnatos change the writing system from left to right, Ge'ez like Arabic was written from right to left [23]. The Amharic language also uses this form for its writing system. Ge'ez has two, the previous and the current, alphabet arrangements. The previous alphabet arrangement uses the አቦጊዳ format and the current alphabet arrangement uses the ሀሁ format [23, 24]. They use almost the same alphabetic arrangement but there is a slight difference on some alphabets or *fidel*.

Table 2. 1: Previous and current alphabet arrangement of Ge’ez language

	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	ጉ	ጊ	ጋ	ጌ	ጎ	ኖ
፲	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
፲፩	የ	ዩ	ዪ	ያ	ዮ	ይ	ዮ
፲፪	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
፲፫	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
፲፬	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
፲፭	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
፲፮	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
፲፯	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
፲፰	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
፲፱	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
፳	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
፳፩	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
፳፪	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
፳፫	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
፳፬	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
፳፭	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
፳፮	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

	ግዕዝ	ካዕብ	ሣልስ	ራብዕ	ሐምስ	ሳድስ	ሳብዕ
፩	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
፪	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
፫	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
፬	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
፭	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
፮	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
፯	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
፰	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
፱	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
፲	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
፲፩	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
፲፪	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
፲፫	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
፲፬	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
፲፭	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
፲፮	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
፲፯	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
፲፰	የ	ዩ	ዪ	ያ	ዮ	ይ	ዮ
፲፱	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
፳	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
፳፩	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
፳፪	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
፳፫	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
፳፬	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
፳፭	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
፳፮	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Table 2. 2: Derived Ge’ez letters

ሏ	ሗ	ሟ	ሧ	ሯ	ቧ	ቇ	ቧ	ኍ	ኗ	ኧ	ኳ	ቈ	ዟ	ዷ	ዻ	ጧ	ጿ	ፗ	ፏ	፟

As described in the above table, Ge’ez language has 182 (7*26) basic letters with the previous and current alphabet arrangement. The other table shows the derived Ge’ez letters from the original ones.

Excluding of some alphabets, Amharic has the same alphabetic arrangement with the Ge’ez language. These excluded alphabets as presented on table are 8 in number which makes a total (7*26) + (7*8) = 182+56 = 238 alphabets on Amharic language alphabets. In addition, Amharic language has a derived alphabets table.

Table 2. 3: Added letters or fidels of Amharic language

	ግዕዝ	ካዕብ	ሣልስ	ራብዕ	ሐምስ	ሳድስ	ሳብዕ
ሇ	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ቈ	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
ኵ	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
ዀ	ኀ	ዂ	ዂ	ዃ	ዄ	ዅ	ዄ
ጒ	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ጕ	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
ፙ	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
ቍ	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ

Table 2. 4: Derived Amharic letters

ኧ	ዠ	ዥ	ዧ	ዣ	ዤ	ጯ	ጧ

To summarize, Amharic language uses the same alphabetic order with Ge'ez language but what makes the difference is the added alphabets and the derived letters from the two languages. Those letters are distinct and may found on each other of the languages.

2.3.2 Syntax

Most of the time Amharic uses the Subject-Object-Verb (SOV) word order. However, in some cases the order may become Object-Subject-Verb (OSV) since Amharic mixes the Semitic and Cushitic languages word structure [22]. On the other hand, the syntax of Ge'ez is a free order and mostly lies on SVO, VSO and OVS.

For instance, we can get the Amharic sentence "እግዚአብሔር ሰማይንና ምድርን ፈጠረ" equivalents in Ge'ez in those three different forms. In SVO, the sentence has this form "እግዚአብሔር ፈጠረ ሰማየ ወምድረ", in VSO "ፈጠረ እግዚአብሔር ሰማየ ወምድረ", and in OVS "ሰማየ ወምድረ ፈጠረ እግዚአብሔር". In the sentence, "እግዚአብሔር" is the subject of Amharic sentence that matches with "እግዚአብሔር" in Ge'ez also; "ሰማይንና ምድርን" is the object in the Amharic sentence that is equivalent with "ሰማየ ወምድረ" in Ge'ez and "ፈጠረ (ላልቶ የሚነበብ)" is the verb of the Amharic sentence which is equivalent with "ፈጠረ (ጠብቆ የሚነበብ)" in Ge'ez [23].

2.3.3 Numerals

Ge'ez has its own numeral system to be used for different linguistics purposes that are involving numbers. Unlike, Arabic numeral system (0-9) that is followed by Amharic, Ge'ez has a bit more representation for some numbers. For instance the numbers 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, and 1000 has their own symbolic representations without combining numbers in the range (0-9) or (፩–፱) these are ፲፣፳፣፴፣፵፣፶፣፷፣፸፣፹፣፺፣፼ respectively. Ge'ez has no symbolic representation for '0' but in writing, it can be expressed as 'አልቦ' to mean ምንም/ዜሮ [21, 23, 24].

Table 2. 5: Ge’ez and Amharic numerals

-	፩	፪	፫	፬	፭	፮	፯	፰	፱	፲
አልቦ	አሐዱ	ክልኤቱ	ሠለስቱ	አርባዕቱ	ሐምስቱ	ስድስቱ	ስብዓቱ	ስመንቱ	ተሰዓቱ	አሠርቱ
0	1	2	3	4	5	6	7	8	9	10
፳	፴	፵	፶	፷	፸	፹	፺	፻	፼	
20	30	40	50	60	70	80	90	100	1000	
አስራ	ሠላሳ	አርብዓ	ሃምሳ	ስድሳ	ሰብዓ	ሰማንያ	ተሰዓ	ምዕት	አልፍ	
፻፻	፻፻፻									
አእላፋት	ትልፊታት									
1000 000	100 000 000									

2.3.4 Similar Letters

Currently, these letters have a similar sound but with different orthographic shape. However, they gave a different meaning when applied each of them on words, phrase and sentences. There are nine Similar Letters (*ተመኩሳይያን ወይም ሞክሼ ፊደላት*) [24].

Table 2. 6: Similar letters (*ተመኩሳይያን*) of Ge’ez

ሠለስቱ	ክልዔቱ	ክልዔቱ	ክልዔቱ
ሀ	*ሠ*	*አ*	*ጸ*
ሐ	*ሰ*	*ዐ*	*ፀ*
ኀ			

Table 2. 7: Similar letters (*ተመኩሳይያን*) of Ge’ez with their definition and reason

Letters	Their Definition	Reasons
ሀ	*ሀሌታው ‘ሀ’*	Beginning of the Ge’ez word *ሃሌሉያ*
ሐ	*ሐመሩ ‘ሐ’*	Beginning of the Ge’ez word *ሐመር*
ኀ	*ብዙኃኑ ‘ኀ’*	Used when *ብዙኃን* is written
ሰ	*እሳቱ ‘ሰ’*	Used when *እሳት* is written
ሠ	*ንጉሡ ‘ሠ’*	Used when *ንጉሡ* is written
አ	*አልፋው ‘አ’*	*አልፋ* is written in it
ዐ	*ዐይኑ ‘ዐ’*	It looks like the shape of an eye and used when *ዐይን* is written
ጸ	*ጸሎቱ ‘ጸ’*	Used when *ጸሎት* is written
ፀ	*ፀሐዩ ‘ፀ’*	It looks like the shape of the sun and used when *ፀሐይ* is written

2.3.5 Word Classes

Words are the center of languages this is to mean that in any language the most recognizable part is its word. On different languages, there are more than tens of thousands of words, but most speakers know and use only a relatively small number among them [25]. Words in general, clauses, phrases and sentences have some guiding rule to follow in order to communicate and get a maximum understanding to the speaker (writer) of a specific language that is called Grammar or in Amharic 'ሰዋሰው'. It also includes phonology, morphology, and syntax. There are seven major parts of speeches in Ge'ez language. Namely, Nouns, Pronouns, Adjectives, Verbs, Adverbs, Prepositions and Conjunctions [22, 26].

Noun

Noun refers anything that represent a thing, feeling, place, animal, person and idea. There are two ways to construct a noun:

- Nouns constructed by nature
- Nouns derived from verbs.

Noun constructed by nature includes for example: ዓራት (አልጋ), ድማህ (አናት), ክሳድ (አንገት), አድግ (አህያ), ስዕርት (ጠጉር). Nouns that are constructed or derived from verbs also used as formal nouns in clauses, phrases, and sentences.

Verb	Noun
ሐለየ (አሰበ)	ሕሊና (አሳብ)
ጥዕየ (ዳነ)	ጥዒና (ደህንነት)
ባረከ (ባረከ)	ቡራኬ (ቡራኬ)

To make plural noun for Ge'ez sentence these alphabets or *fidels* used: አ፣ ት፣ ን፣ ያን፣ ያት፣ ው፣ ል. While 'አ' always added on the begging of the word, all the others come at the end of the word.

Table 2. 8: Making of plural nouns for Ge'ez language by adding fidels

Added *fidels*	Original known	Inflicted to
አ	ደብር (ተራራ)	አድባር (ተራሮች)
አ.....ት	ገበር (አገልጋይ)	አግበርት (አገልጋዮች)
ት	ንጉሥ	ነገሥታት
ን	ባዕድ	ባዕዳን

ያን	ዘማሪ	ዘማሪያን
ያት	ውዳሴ	ውዳሴያት
ው	አብ (አባት)	አበው (አባቶች)
ል	ኪሩብ	ኪሩቤል

Pronouns

Pronouns shortly considered as a substitution to a noun or noun phrase. Without pronouns, it is mandatory to mention nouns and this in turn makes our speech and writing more cumbersome. Ge'ez language has 10 pronouns [27, 22] while Amharic has nine pronouns [26]. Pronouns also used for certain adverbs, adjectives, and other pronouns. Table 7 presents the 10 pronouns of Ge'ez language.

Table 2. 9: Ge'ez pronouns list

	ግእዝ	አማርኛ	English		ግእዝ	አማርኛ	English
፩	አነ	እኔ	I	፪	አንትን	እናንተ	You
፫	ንሕነ	እኛ	We	፭	ውእቱ	እሱ	He
፮	አንተ	አንተ	You	፯	ይእቲ	እሷ	She
፰	አንቲ	አንቺ		፱	ውእቶሙ	እነሱ	They
፲	አንትሙ	እናንተ		፲፩	ውእቶን	እነሱ	

Ge'ez pronouns used as pronouns, as verb to be and as demonstrative pronouns. They further divided into pronouns of gender, pronouns of number, personal pronouns and pronouns based on their task. These can summarize in table 8.

Table 2. 10: Ge'ez pronouns with their respective task

Personal			Gender			Number		Based on their task		
First person	Second person	Third person	Male	Female	Both	Singular	Plural	Near indicator	Far indicator	Common
አነ	አንተ	ውእቱ	አንተ	አንቲ	አነ	አነ	ንሕነ	አንተ	ውእቱ	አነ
ንሕነ	አንቲ	ይእቲ	አንትሙ	አንትን	ንሕነ	አንተ	አንትሙ	አንቲ	ይእቲ	ንሕነ
	አንትሙ	ውእቶሙ	ውእቱ	ይእቲ		አንቲ	አንትን	አንትሙ	ውእቶሙ	
	አንትን	ውእቶን	ውእቶሙ	ውእቶን		ውእቱ	ውእቶሙ	አንትን	ውእቶን	
						ይእቲ	ውእቶን			

As 'verb to be', pronouns also expressed as a past tense

Table 2. 11: Ge'ez pronouns as verb to be

ግእዝ	English	አማርኛ
ውእቱ	Be, is, was	ይሆናል፣ነው፣ነበር
ይእቲ	Be, is, was/will be	ናት፣ነበረች፣ትሆናለች
ውእቶሙ	Are, were	ናቸው፣ነበሩ፣ይኖራሉ
ውእቶን	Are, were/will be	ናቸው፣ይኖራሉ፣ነበሩ
አንተ	Are, were	ነኽ፣ትኖራለህ፣ነበርክ
አንቲ	Are, were/will be	ነሽ፣ነበርሽ፣ትኖሪያለሽ
አንትሙ	Are, were/will be	ናችሁ፣ነበራችሁ፣ትኖራላችሁ
አንትን	Are, were/will be	ናችሁ፣ነበራችሁ
አነ	Am, was/will be	ነኝ፣ነበርኩ፣ትኖራላችሁ
ንሕነ	Are, were/will be	ነን፣ነበርን፣እንኖራለን

Pronouns ***as demonstrative pronouns***

Table 2. 12: Ge'ez pronouns as demonstrative pronouns

ግእዝ	አማርኛ	English	ግእዝ	አማርኛ	English
ዝ፣ ዝንቱ	ይህ፣ ይኸው	This	ዝኩ፣ ዝስኩ፣ ውእቱ	ያ፣ ያው፣ ያውና	That
ዛ፣ ዛቲ	ይች፣ ይችው	This (Feminine)	እታኩቲ፣ አንትኩ፣ ይእቲ	ያች፣ ያችው፣ ያችውና	That (Feminine)
እሉ፣ እሉንቱ	እኒህ፣ እኒሁ	These	እሙንቱ፣ እልክቱ፣ ውእቶሙ	እነዚያ፣ እነዚያው፣ እነዚያውና	Those
እላ፣ እላንቱ፣ እሎን	እኒህ፣ እኒሁ	These (Feminine)	እማንቱ፣ እልክቶን፣ እልኮን	እነዚያ፣ እኒያው	Those (Feminine)

Adjective

A word further describes, define and identify noun or pronoun. While nouns tell us about things nature, adjectives stand to tell us about their behavior or characteristics like type, color, property, shape, size [21]. Adjectives can be constructed by changing the verb into a word which his last alphabet is the third alphabet the arrangement like ፈጠረ → ፈጣሪ or by changing the verb into a word which his last alphabet is the sixth alphabet like ተግሀ (ተጋ) → ትጉህ or by adding a 'መ'

alphabet or fidel on the verb like ዘመረ (አመሰገነ) → መዘምር. Another way of forming an adjective is by adding the fidels ‘ዊ (ይ)’ or ‘ዊት’ like ገሊላ → ገሊላዊ/ይ፣ ኢትዮጲያ → ኢትዮጲያዊት.

In Ge’ez, there are also demonstrative adjectives that used to express near or far things. For example, ዝ/ዝንቱ (ይህ-ለወንድ) ፣ ዛ/ዛቲ (ይች-ለሴት) and ዝክቱ/ዝኩ (ያ-ለወንድ) ፣ እንታክቲ (ያች-ለሴት). There are also adjectives to represent an amount of a thing like, ሕቅ → ጥቂት፣ ንስቲት/ሕዳጥ → ትንሽ፣ ብዙኅ → ብዙ፣ ንሕኑሕ → ብዙ፣ ኵሉ → ሁሉ. Other forms of constructing an adjective are adjectives of numbers like አሐዱ፣ ክልዔቱ, interrogative adjectives like መኑ፣ ምንት፣ አይቴ, and adjectives to plurality for both men and women like ቀተለ (ገደለ) → ቀተልት (ገዳዮች) ፣ ጸሐፈ (ጻፈ) → ጸሐፊዎች.

Verb

Verb is a word to describe an action, state or occurrence and forming the main part of the predicate of a sentence [21]. Amharic verbs derived from roots. They use a combination of prefixes and suffixes to indicate the person, number, active or passive voice, tenses and gender. While Amharic sentence placed at the end of the sentences in most of the times [28], most of Ge’ez sentences have a verb on their middle sentences [21]. Inflection are used to Ge’ez words with respect to person, gender and number. Verbs of the language may be either in a perfect (past form) or imperfect form (present and future forms). The verbs of the Ge’ez language have Semitic non-linear word formation with intercalation of roots with vocalic pattern. Verbs of Ge’ez and Amharic agree with their subject and objects [17]. In Ge’ez language, there are eight root verbs, with different characteristics, that lead the time behavior and using their morphology style [16, 29]. Other similar verbs follow these root verbs.

Table 2. 13: Ge’ez language root verbs

አርእስተ ግስ	ትርጉም
ቀተለ	ገደለ
ቀደሰ	አመሰገነ
ባረከ	ባረከ
ተንበለ	ለመነ
ማህረከ	ማረከ
ሣለመየ	ሾመ
ክህለ	ቻለ
ጦመረ	ጻፈ

Adverb

Adverb is a word that is used to change, modify or qualify several types of words like verb. There are six types of adverbs in Ge'ez language. These are Adverbs of time, frequency, place, manner, reason, and question. Adverbs of time tells the time when it is used on the sentence. For example, ጌሠም (ነገ), ትማልም (ትናንት). Adverbs of frequency describes how many times an event happens, ኩለሄ (ሁል ጊዜ), በበጊዜሁ (በየጊዜው). Adverbs of place gives us an information on a place, ህየ (እዚህ), ጥቃ (አጠገብ). Adverbs of manner describes how one thing takes place, እሙነ (በእርግጥ), ክሡተ (በግልጽ). Adverbs of reason presents the reason on the occurrence a thing, አምጣነ (ያህል), በእንተ (ስለ). Finally, adverbs of question stand to raise a question, እፎ (እንዴት), ምንት (ምን).

Prepositions

Prepositions lied on nouns or pronouns to connect the people, objects, time and locations of a sentence. The following table presents Ge'ez language preposition.

Table 2. 14: Some prepositions of Ge'ez language

ግእዝ	አማርኛ	English
ዲበ፣ ላዕለ፣ መልዕልተ	ላይ፣ በ - ላይ፣ ከ - ላይ	On, above
መትሕተ፣ ታሕተ	ታች፣ በ - ታች፣ ከ - ታች	Under
ውስተ፣ ውሰጤ፣ ማእከለ	ውስጥ፣ በ - ውስጥ (ከ ውስጥ)፣ በመካከል	In/inside/ in the middle
ቅድመ፣ ድኅረ	ፊት - (በፊት)፣ ኋላ/በኋላ	Before/ after
ኀበ፣ መንገለ	ወደ	To
ህየንተ፣ በእንተ፣ በይነ	ስለ	About
እም፣ እምነ	ከ፣ ከ - ይልቅ	From

Conjunctions

Conjunctions are words to connect clauses or sentences or to coordinate words in the same clause. Conjunctions in Ge'ez language are presented on the below table.

Table 2. 15: Some conjunctions of Ge’ez language

ግእዝ	አማርኛ	English
ከመ	እንደ/እንድ	As … As
አምሳለ	እንዳ	
በዘ	እንዲ	
በእንተ	ስለ	About
ህየንተ		
በይነ		
እንበይነ		
አመ	በ-ጊዜ	In…time
ሶበ		
ጊዜ		
አምጣነ	እና/ያህል/ስለ	And/Due to
እስመ		
አኮኑ		

There are positive and negative conjunctions. Positive conjunctions expressed in the positive sentences while negative conjunctions used for negative expressions. ወ & አው are positive conjunctions and ዓዲ, ባሕቱ ዳዕሙ, and አላ are negative conjunctions.

We can conclude what we discussed in the previous sections i.e. the similarities and differences between the two languages, Ge’ez and Amharic, with the following table.

Table 2. 16: Difference and Similarity between Ge’ez and Amharic language

	Writing system	Syntax	Numeral	Similar letters	Word classes
Different		*	*		
Similar	*			*	*

The sign (*) is used to indicate the difference or similarity of the two languages from the listed and previously discussed point of view. They have similarity in writing system, similar letters used and word classes. As discussed, both languages use the same alphabetic order, but they have a difference in added letters in their alphabets. Their difference is on syntax or word order and numbers used. While Amharic follows the SOV word order, Ge’ez uses mostly the SVO, OVS, and VSO word orders. Ge’ez has a bit more numeral representation than Amharic.

2.4 Machine Translation

Machine translation (MT) is an automated translation carried out by a computer to translate from on (source) language to other (target) languages [30]. It involves the use of bilingual data set and other language assets to build language and phrase model for translation. Due to this, it also named as Natural language processing. MT is one of an applied research and it gets its input from linguistics, computer science, artificial intelligence, translation theory and statistics (statistical). The history of machine translation begins in early systems in 1940s and 1960s in the aim of producing high-quality translation. Building a sophisticated method or forcing the input for some restrictions have a great impact on improving the systems translation quality [31]. As indicated by ALPAC (Automatic Language Processing Advisory Committee), the machine translation ability is low as compared with human translator. IBM started to work on statistical machine translation in 1980s and in 1990s, parallel corpus availability has increased. Moses, powerful tool for statistical MT, created in 2006 [32].

Machine translation systems may be bilingual or multilingual. The bilingual refers the involvement of two languages in a translation and it is mostly unidirectional i.e. the source language is the language that the translation begins from while the target language is the language that the source language is translated into. On the other hand, multilingual translation system also consists two languages, but the translation is bidirectional i.e. one language becomes a source language as well as a target language. There are three types of machine designs and all systems may fall in one of the three. One type of the approach is direct translation approach. This system is designed to translate directly from source language to target language and it is bilingual and unidirectional [31].

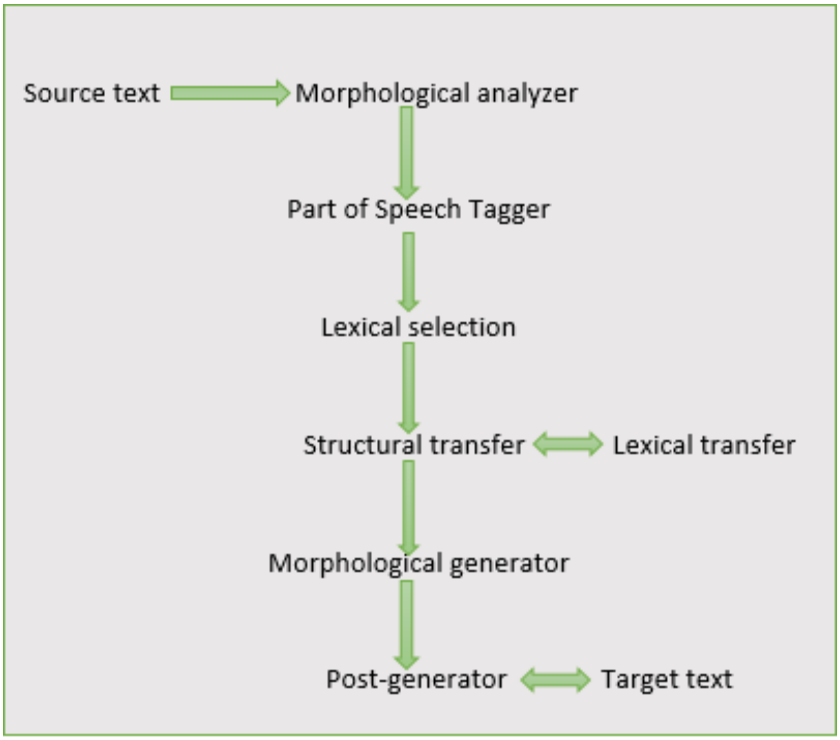
The second design type is an Interlingua approach that based on an assumption that there is a possibility to convert the source language into some internal representation that is common beyond one language. The translation goes from source language to this representation or Interlingua and then to the target language. It will be more economical when it includes more than three languages and the system will be more complex. The final approach is the ambitious transfer approach that designs a three stage for the translation involving abstract representations for source and target languages. The source text first transferred into abstract source language-oriented representation then to the corresponding target language-oriented representation and finally to the target language.

There are many approaches to machine translation. The main ones as described by [4, 33]. Broadly categorized into rule-based, corpus-based and the approach that consists the best feature of the two, hybrid approach. The rule-based approach consists of Direct, Interlingua and Transfer-based machine translation approaches. Statistical and example-based approaches fall under corpus-based machine translation approach. The last category, hybrid approach, takes the merits of the two above-mentioned approaches.

2.4.1 Rule-based MT

RBMT is the first machine translation approach that developed to help the translation and it has a collection of linguistic rules to analyze, transfer and generate [12]. Due to this, the rule-based system needs syntax and semantic analysis and syntax and semantics generation. The overall steps presented for translation categorized in the following figure.

Figure 2. 1: Architecture of RBMT



The morphology of the source text is analyzed then information about the part-of-speech of source word is passed to the next stage. The source word's syntactic information also passed in order to get a full information about it and map it into the structure of target sentence. After the source sentence structure mapped into the structure of the target sentence, the next step is to

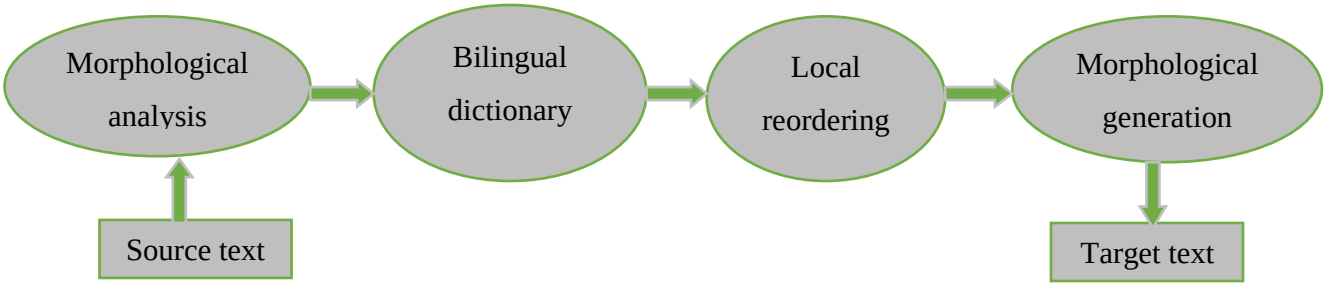
translate the source language words into target language. The final step is to map all entries of the source language sentence into the appropriate forms on the target sentence [4].

RBMT needs great human intervention to write all linguistic resource like part of speech taggers, syntactic parsers, bilingual dictionaries, source to target transliteration, morphological generator, structural transfer and reordering rules [4, 33]. Hence, to give an input to the system regarding with the above-mentioned resources linguistic knowledge is necessary. Under RBMT, there are three sub approaches namely: Direct, Transfer-based and Interlingua MT approaches. The difference between these sub approaches is the in-depth analysis they give to the source language and how far they go to provide a language independent representation of meaning.

Direct Machine Translation

It is a word-by-word translation approach with some simple grammatical reordering. It involves shallow morphological analysis, lexical transfer, based on bilingual dictionary, local reordering and morphological segmentation [30].

Figure 2. 2: Direct Machine Translation



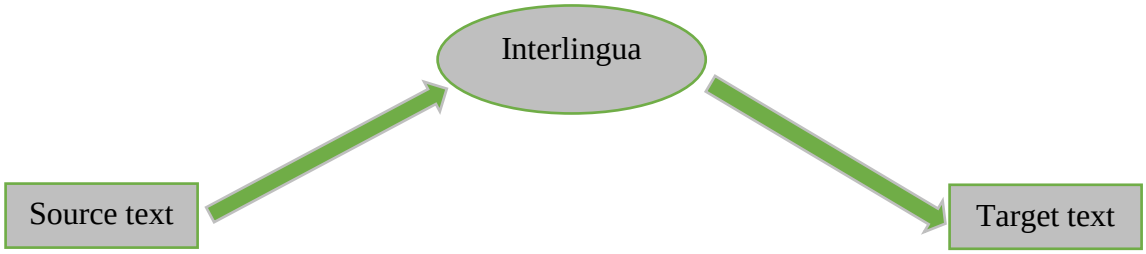
In the morphological analysis phase, there is identification of word endings and reduction of inflected forms to their uninflected basic forms. The result from the morphological analysis phase become the input to large bilingual dictionary program to provide target language word equivalences. Some local reordering rules follow to give more acceptable target language output this may include moving some adjectives or verbs. Then after the morphological generation phase takes its part. On this stage, all the internal representation for a word is converted into its surface form and finally the target text is produced [10, 16].

Although this approach is fast, simple, inexpensive and no translation rules hidden in lexicon, it has some problems: it misses any analysis of the internal structure of the source text and lacks computational sophistication that leads poor translation quality.

Interlingua based Machine Translation

The translation bases on representing the source language text into an intermediary form, Interlingua [15]. The idea is to represent all sentences that tells the same thing in the same way, independent of any language. This approach translates by performing deep semantic analysis on the X input language into the Interlingua representation and providing translation to target language Y from that intermediate representation. It involves analysis and generation: analysis helps to derive an Interlingua representation.

Figure 2. 3: Interlingua Machine translation



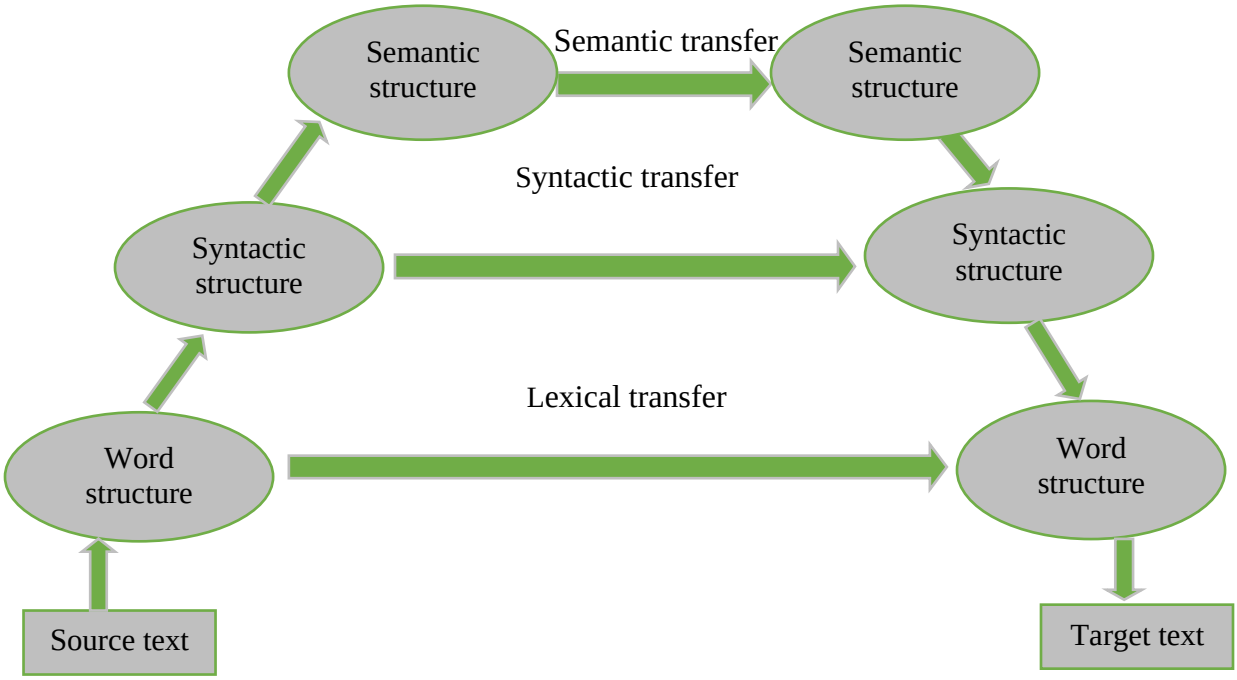
The source language or the sentence to be translated is transformed into an Interlingua that is an abstract language independent representation. Then the target language text is generated from that internal representation. This representation allows analyzers and generators to be written by monolingual system developers and handles very different languages from each other but it's applicable for a specific domain not for a wider domain [30, 33].

Interlingua based MT approach is the most attractive form of the rule-based approaches since it works fine regardless of any language and it permits translation from and into the same language. Its drawbacks: hard to define Interlingua and it fails to take the similarities between languages.

Transfer based Machine Translation

This approach like Interlingua machine translation approach uses an intermediate representation to capture the structure of the source language text to give a correct translation [4]. It involves analysis, transfer and generation to give a syntactic representation of source language sentences using source language parser, to transfer the output of the source language parser into its corresponding target language-oriented representation and to generate the target language text. It requires rules for syntactic, semantic and lexical transfers.

Figure 2. 4: Transfer based machine translation

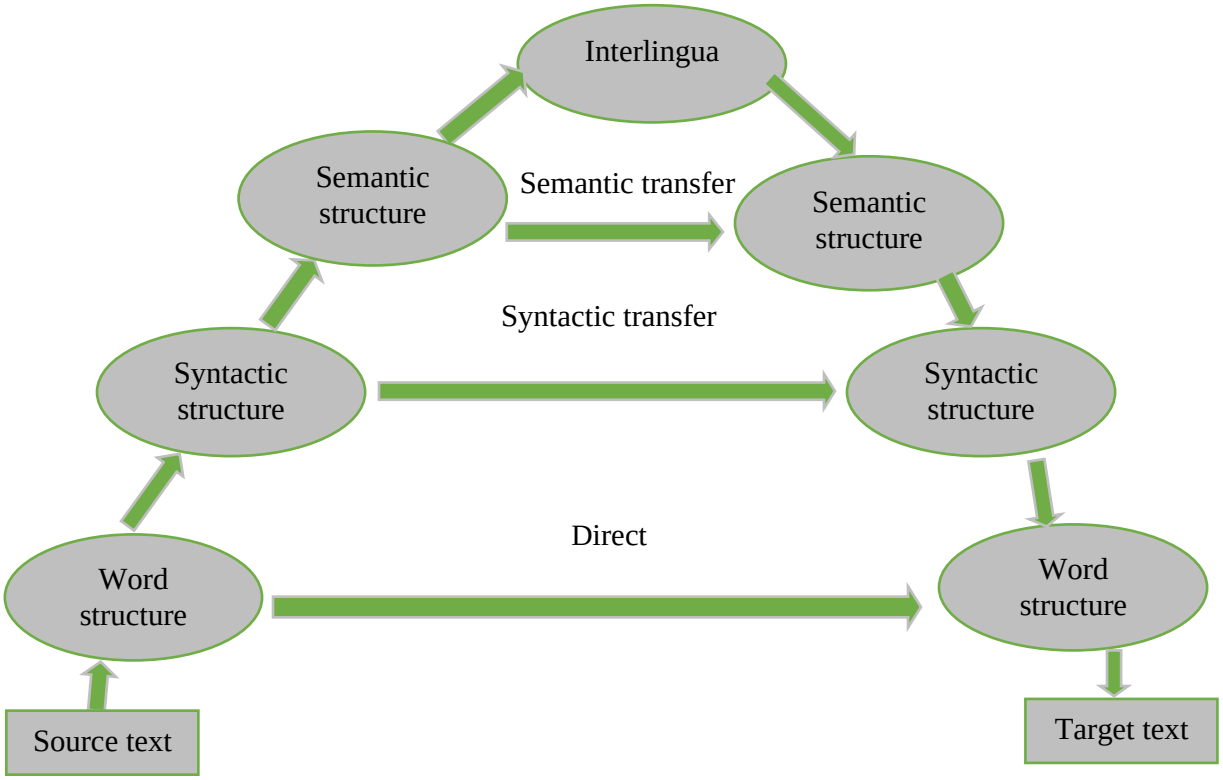


The transfer-based MT approach has the same feature as Interlingua based MT in a way that both use intermediate representation that captures the meaning of the original sentence for correct translation. However, transfer based has a dependence on a language pairs involved. There are three types of transfer from source language intermediate representation to target language intermediate representation. These are lexical, syntactic and semantic transfers. In lexical the word structure of the source text passed to target text through the intermediate representation. Syntactic transfer involves transferring of syntactic structures between the source and the target language. Semantic transfer characterized by creating and transferring semantic or meaning representations that are dependent on the source language. After passing through these different stages, finally the target text is generated [3, 12].

The transfer-based MT approach offers the ability to deal with more complex source language phenomena than the direct approach, high quality translations obtained than direct translation and it has a relative fastness than the Interlingua, and it provides an accuracy of around 90% but it has some difficulties. Some of the disadvantages are rules need to introduce at source language analysis, source-to-target transfer and target language generation, in reusable modules it is difficult to do as much work as possible and the transfer modules cannot simply keep [4].

Finally, the overall system involving Direct, Interlingua based and Transfer based approaches represented by a Vauquois triangle. The triangle shows comparative depths of intermediary representation, Interlingua machine translations at the peak, followed by transfer-based, then the direct translation. The below figure shows this overall process [4].

Figure 2. 5: The Vauquois Triangle



The above figure shows the increasing depth of analysis required as we move from direct to Interlingua and the decreasing amount of transfer knowledge needed as we move up the triangle.

2.4.2 Corpus-based Machine Translation Approach

The aim of corpus-based machine translation approach is to rectify the knowledge acquisition problem of the rule-based approach. It takes a large amount of raw data in the form bilingual parallel corpora to gather a knowledge on the coming translation [30]. Due to this, it can alternatively name as Data driven MT. The raw data consists parallel source and target language texts in which translation can smoothly conducted with the help of suitable methods. The approach classified into other two sub approaches: Statistical and Example based approaches.

Statistical Machine Translation Approach

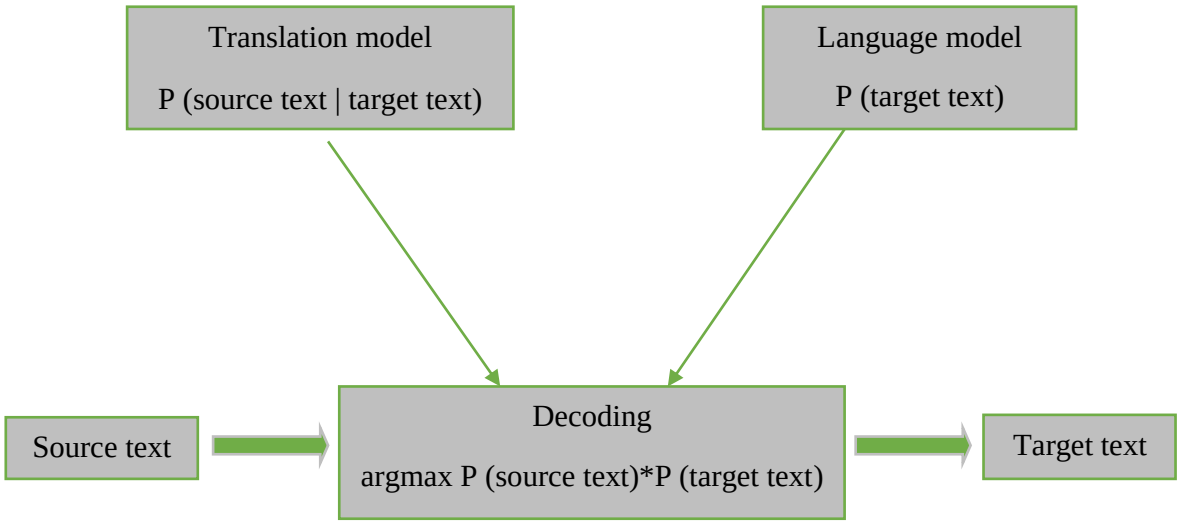
The approach bases on statistical models that finds the most probable target text give a source text. The noisy channel model is applicable in this approach. The model assumes the source language sentence is a corrupted version of the target language sentence then the task is to discover the hidden or target language sentence that generated the observation or source language sentence. The best target sentence $T=w_1, w_2 \ldots w_n$ is the one whose probability $P(T|S)$ is the highest [34]. The model uses a Bayes' theorem [15], that assumes each sentence in a given target language is a possible translation to the sentence in the source language and the best translation is the one whose target sentence has the highest probability. The theorem assigns a probability P(T/S), the probability that a translator will produce T in the target language when presented with S in the source language, for every pair of source and target sentence (S, T).

Given the target language sentence T, then the aim is to find the source language sentence S from which the translator produced T. Mathematically written as:

$$P(S|T) = P(S)\, P(T|S)/P(T)$$

The noisy channel model of statistical machine translation requires three components: a language model, a translation model and a decoder.

Figure 2. 6: Architecture of SMT



The decoder in statistical machine translation is responsible to produce the best translation according to the product of the translation and the language model by taking the source language. Translation model takes the source and target texts, language model takes only the target text.

Language model

Language modelling is the process of determining the probability of a sequence of words. It is applicable in speech recognition, optical character recognition (OCR), handwriting recognition, machine translation, information retrieval, and spelling correction. It is a probabilistic way to take regularities of a language in the form of word-order constraint. Language modeling component provides a language model for the target language by taking a monolingual corpus. Sequences of words that are convincing to the provided input text given high probabilities while sentences with less related sense to the given sentence get a low probability. Language models used n-gram models, which based on sequence of n words. Given a word string '*a'* with n words a=$w_1w_2 \dots w_n$ mathematically written as:

$$\Pr(\mathrm{a}) = \Pr(Wi \,|W1, W2, W3, \dots, W\mathrm{n}) = \prod_{i=1}^{n} \Pr(Wi \,|W1, W2, W3, \dots, W\mathrm{n}-1)$$

W_i is the i^{th} word and n is the word length

The N-gram Model

Jelinek and Mercer proposed the n-gram model and it is a dominant one among statistical language models [35]. It assumes the probability of the n^{th} word depends only on the n_1 preceding words based on the Markov assumptions which says only the prior local context consisting of last few words affect the next word. Accordingly, the n-gram has (n-1) th order of the Markov model [36]. While a high n provides a detail information concerning the context of a given sequence, the low n provides more cases that will be seen in the training data and this implies a more reliable estimate. When the size of the corpus gets large, the reliable count of the n-grams will be higher.

Translation model

Translation models tells the bilingual relationship between the source and the target languages text from the parallel corpus. The training corpus for this model is a sentence level aligned

corpus for the languages to be involved in the translation phase. Due to the data sparsity problem, it is difficult to conduct sentence level translation. Hence, decomposing the sentences into smaller chunks is preferable. Based on this, most of the time word based, phrase based, and syntax based statistical translation models are used widely [37].

Statistical word-based translation model assumes that every target sentence is a possible translation of every source sentence. The assumption arises from relying on the fact that there is a more suitable word choice to get a reliable output on the process of translation [11].

As [38] described the translation model is the reverse process because of the Bayesian inversion. The following mathematical expression puts the above assumption in short:

$$P\ (s|t) = \sum_{a} P(s, a|t)$$

Where P (s|t) = probability of the source sentence given the target sentence

P (s, a|t) = probability of word alignment of source sentence to the target sentence

Statistical phrase-based translation model the transition model has the responsibility to find the probability that E generates F where E is a source language let us say English sentence and F is for target (foreign) language sentence [39]. This translation model bases on using phrases or sequences of words instead of a single word as a unit of translation. It has three steps throughout its translation. The primary step is grouping the English source words into phrases e_1, e2… e_i then followed by translating each English phrase e_i into a foreign phrase f_j. The final step is reordering of each foreign phrase. Its probability model depends on the translation probability and a distortion probability. ϕ ($f_j|e_i$) is the translation probability of generating foreign phrase f_j from English phrase e_i.

Distortion probability or *d* applied to order foreign phrases. It refers to a word having a different (distorted) position in the foreign sentence than it had in the English sentence. The distortion probability when applied to phrased based machine translation means that the probability of two consecutive English phrases separated in foreign by foreign words of a length. The distortion parameterized by $d\ (a_i\text{-}b_{i-1})$, where a_i is the start position of the foreign phrase generated by i_{th} English phrase e_i, and b_{i-1} is the end position of the foreign phrase generated by the '*i-1th'* English phrase e_{i-1}. The translation model in short will be:

$$P(F|E) = \prod_{i=1}^{I} \phi(fi, ei) d(ai - bi - 1)$$

Where ϕ (f_i,e_i) = the translation probability of generating foreign phrase *fi* from English phrase *ei*

$d\ (a_i - b_{i-1})$ = distortion probability

Finally, the phrase-based model needs two additional things to use: model for decoding and model for training. Model for decoding needed to go from a foreign string to the hidden English string while model for training is to learn parameters [39].

Statistical syntax-based model phrase-based model has a drawback that it works whenever there is a phrase to be used in other word it does not have any room for syntax [40]. One of the syntactic behaviors is changing a word order whenever it is necessary to cop up with situation occurred on the move. The statistical syntax-based model uses syntactic rules to follow for machine translation because it is mandatory to have some guiding rule about the syntax and the sentence structure of a given language. Word re-ordering is one of the rules for this type of model. The sentence structure in one language may differ from the other so rules play a key role in understanding and matching with this scenario. For example, word order of Ge'ez language has a VSO, SVO, or OVS structure while Amharic mostly follows the SOV sentence structure. The syntactic rules applied on input, output or both languages.

The syntactic rules applied using tree-to-string, string-to-tree and tree-to-tree models. String-to-tree syntax-based translation model views the input language string as provided by parse tree of output language and passed through a noisy channel [41]. A syntactic parser assigns a parse tree to the English string followed by insertion of words at each node and translation of leaf words. The gains from the alignment tool of parallel text sentences reported and the decoder presented. Tree transducers also developed to be able to compute transformations of trees [42]. Other works on complex rules extracted from parallel text to build models of string-to-tree alignment discussed on [43, 44, and 45]. On the other hand, syntax-based translation systems using tree-to-tree resented on [46, 47]. Syntax based translation is based on translating syntactic units rather than a single word or strings of words.

Decoder

The decoder is responsible to produce the best translation according to the product of the translation and the language model by taking the source language [11]. The problem on the search is to find a sentence that maximizes the translation and the language model probabilities. To solve the problem the decoder uses the best-first search algorithm that informed by knowledge from the problem domain. It selects node n in the search space to explore based on evaluation function f (n). They are the variants of A* search algorithms which is a specific kind of best-first search.

The A* search main aim is to keep the priority queue which is traditionally referred to as a stack with the entire partial translation hypothesis, together with their scores. The search space can be limited by only considering the possible translation for foreign sentence F hence the entire unnecessary search space of source sentences ignored. The decoder is responsible to find the highest scoring sentence in the target language based on the translation model in relation with the given source sentence. The decoder is also able to provide a ranked list of the translation candidates and to supply various types of information about how the decision made [11, 48].

Example Based Machine Translation

Example based machine translation also called machine translation by example-guided inference, machine translation by analogy principle. The main concern is translation by people don't always involve deep linguistic analysis of a sentence and the translation is conducted by decomposing sentences into fragments, translating each of them and composing them into one long sentence properly [49]. On this machine translation approach set of phrases in the source language and their corresponding translations in the target language are given then the system uses these examples to translate new similar source phrases into the target language. The main idea behind this is if a previously translated phrase occurs again, then the same translation is likely to be correct again [4, 33].

Example Based Machine Translation approach passes through three different steps. The first step is to match the source language input against the example database. Then, selecting the corresponding fragments in the target language is proceeds. Finally, recombining the target language fragments to form a correct text takes place. The approach makes advantageous since fragments of human translation which result higher quality, but it may have limited coverage depending on the size of the example database, needs the production of dependency trees from

analysis and generation modules and computational efficiency, for large databases, despite of using parallel computation techniques [4].

2.4.3 Hybrid Machine Translation

The approach makes use the strong side of the statistical a rule-based translation approach [11]. It has a better efficiency from all the machine translation approaches and used in different ways. One way is to perform the translation at the first stage using a rule-based approach followed by adjusting the output using statistical information. Moreover, in some cases rules used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system. The later way is more suitable since it has more power, flexibility, and control in translation [4, 33]. Example of the hybrid machine translation system is Oepen that integrate the statistical method with the rule-based method.

2.4.4 Neural Machine Translation

Neural machine translation is a newly introduced machine translation system proposed by Kalchbrenner and Blunsom (2013) [50]. The main goal of neural machine translation system is constructing a single neural network that can jointly tuned to maximize the translation performance. The model belongs to a family of encoder-decoders, every language has its own encoder and decoder, or it involve a language specific encoder that is applicable to each sentence whose outputs then combined [50].

NMT by means of encoder-decoder approach encodes a completely input sentence into a fixed length vector then the translation is decoded but this mechanism has a problem to translate long sentences. Letting a model to search a set of input words or their annotations computed by an encoder when generating each target word will rectify the issue [51].

Overall, Machine Translation may or may not need human intervention in the process of translating from one source language to another target language. From our discussion we can mainly categorize Machine translation approaches into four namely Rule-based, Corpus-based, Hybrid and Neural. Direct, Interlingua and Transfer based approaches are of the Rule-based type and under Corpus-based there are Statistical, and Example based MT approaches. Hybrid is the use of best feature of two approaches while Neural MT concerned with using single neural network to increase the translation performance.

2.4.5 Evaluation of Machine Translation

Evaluation is important to check continuously whether the algorithm we used provides the expected result or giving us the unwanted/unexpected result. Translations evaluated along fidelity and fluency combination of the two and evaluated by using human raters or automatically [52, 53]. Using human raters provide the most accurate evaluation when evaluated along fluency and fidelity. Fluency is evaluated by means of how the translation is intelligible, how clear, how readable, or how natural does the output is. To do this the human raters will be given a scale and ask them to rate each sentence of the machine translation output and other mechanism will rely less on the conscious decision of the raters.

Fidelity measures the adequacy and informativeness. Adequacy judged by whether it contains the information that existed in the original sentence and informativeness bases on whether the information in the machine translation output is enough to perform some task [52]. However, human evaluation mechanism is time consuming and expensive in terms of finance and the inter evaluator agreement (different evaluators may give different results) and intra-evaluator consistency (the same evaluator may produce different result at different times). To overcome these problems automatic evaluation by means of BLEU (Bilingual Evaluation Understudy) score, NIST, TER and other mechanisms are used [52].

In summary, there are a lot of studies conducted in Machine Translation between different languages in Ethiopia concerning with Statistical, rule-based and hybrid machine translations. In most cases Statistical MT was used but there are also researches which are involving rule-based approach by combining with statistical approach. Here the rule-based approach is used to syntactical reorder the source sentence in order to have the same sentence structure with the target sentence to improve the efficiency of the translation system. The statistical approach comes with the use of translation and language models. Statistical MT approach is widely used in machine translations so in this study we are aiming to show the improved performance on the translation system by adding rules to syntactically reorder the Ge'ez sentences to look alike the sentence structure of the Amharic sentence.

Chapter Three

Related Works

3.1 Overview

This chapter mainly focuses on previously conducted works related with machine translation system involving languages from European and Asian and of course from different languages spoken in Ethiopia. Studies that involve English with Ethiopian languages and machine translation system between languages in Ethiopia are well covered.

3.2 Machine Translation systems involving European languages

From European languages, a research work to translate French and German languages to English language using a statistical approach mentioned. Stat-XFER framework developed to translate MT systems on different data conditions and it's a search based, and syntax led framework. In this study, statistical methods, which permit extraction of syntax-based transfer rules from parallel corpora with word alignments in hand, and constituency parses are used. Bilingual translation lexicon and a transfer grammar, which manually developed by language experts, are built. Parallel sentences parsed with Stanford, Xerox XIP and English and German version of Stanford parser for English, France and Germany languages respectively. The 2007 WMT shared task used to evaluate the performance and based on the result the Stat-XFER systems' get a low score on the evaluation [54].

English-Spanish machine translation, developed by Preslav Nakov, mainly focused on domain adaptation, sentence paraphrasing, tokenization and recasting using statistical approach. Experiments also conducted to these elements. Domain adaptation uses small in-domain news bi-text and a large out-of-domain from Europarl corpus. Two translation models and two separated language models built in this study. Experimental results on tokenization and recasting on WMT'07 news test data provides 35.09% Bleu score which shows an improvement from the previous result on this dataset. On the other hand, 21.92% Bleu score achieved using WMT'08. Building separate translation and language models has a remarkable effect on the efficiency improvement of the translation system [8].

3.3 Machine Translation systems involving Asian languages

English Thai (Thailand) machine translation study comes up with reordering rules based on phrase. The rules applied before training and testing steps English sentences in a preprocessing step. The study aims to improve a phrased based statistical MT in language pairs with different word orders using reordering rule with statistical MT mechanism. The source language sentences parsed using a parser, a Stanford parser, and then followed by reordering rule to make the sentence structure more like the target language. Reordering rules constructed from the classified parse trees of the training set. Training, testing and translating stages will proceed once the reordering rule applied on the prepared corpus. The study gets a BLEU score of 57.45% that shows a remarkable advancement from the previous experimental result [55].

English to Chinese machine translation system also uses a syntactic reordering approach under its study. The system reorders the English sentences to look alike Chinese word order in the sentence using a Penn Chinese Treebank guideline to get a convenient way of reordering rules. The reordering has three categories namely Verb phrases, Noun phrases and Localizer phrases (to map to prepositional phrases in English) and as the researchers' identified other phrase types does not require a reordering rule. The study uses 637K pairs of parallel sentences from various resources. NIST MT evaluation data for Chinese from 2002 to 2006 that have four human generated English reference translation for each Chinese input used for tuning and testing. 2347 sentences for tuning to optimize various parameters using minimum error training and 2320 sentences for different analysis experiments used. Before applying the reordering rules, segmentation, part-of-speech tagging, and parsing are applied. They used the perception-learning algorithm to train the Chinese Treebank-style tokenizer and part-of-speech tagger. After this, reordering rules on the parse tree of each input used. The output of this step is an input to re-tokenization to make sure it is consistent with the baseline system. The evaluation result shows 30.86% BLEU score which improved as compared with the previous results [56].

Mossa Ghurab et al studied on Bidirectional Arabic-Chinese machine translation systems using phrase-based statistical approach. As the previous two research works involving Asian languages, which discussed above, this research also uses a phrase-based statistical approach. A corpus from the United-Nations website and different news engine websites are used. To evaluate the efficiency of the system the study used BLUE and NIST evaluation metrics. The system gets a BLEU and NIST score of 0.4916 and 7.9905 respectively from Arabic to Chinese

while on the other hand a 0.4678 and 7.0643 evaluation score from Chinese to Arabic language translation. The study succeeds on integrating models into statistical machine translation architectures to make a smooth interaction. However, the failed to reason out why BLEU score evaluation metrics gets a low score as compared with the NIST [57].

3.4 Machine translation systems involving Ethiopian Languages

Michael Gasser conducted a study on English-Amharic translation system by using a rule-based approach. He states the implementation of rule based bidirectional Amharic-English machine translation system in L3 framework and using an extensible dependency on grammar that relies on constraint satisfaction on parsing and generation. In addition, Michael Gasser focuses on features and advantages that L3 framework offers for handling structural divergences between the two languages and the capacity to accommodate shallow and deep translation within a single system. The proposed system only shows translation using simple sentence means that it does not have a room for complex Amharic sentences [12].

Machine translation system conducted by Jabesa Daba and Yaregal Asabie uses a hybrid approach, statistical and rule based, for bidirectional English-Oromiffa translation. Reordering rules implemented on this study since the two languages have different sentence structure. The ordering applied on simple, interrogative and complex sentences of the two languages to make similar sentence structure with their respective target language. Two experiments performed using statistical and hybrid approaches that yields a BLEU score of 41.50% and 32.39% when translating from English-Oromiffa and Oromiffa-English respectively using statistical approach. On the hand, the hybrid approach provides a BLEU evaluation score of 37.41% and 52.02% from English-Oromiffa and Oromiffa-English respectively. As the result shows, the hybrid approach is a way better than a pure statistical approach [11].

Mulu Gebreegziabher and Laurent Besacier also studied on English-Amharic machine translation using a statistical approach. They used 632 parallel corpora from which 115 is for experimentation purpose. Pre-processing like text conversion, trimming (performed before and after aligning at document level), sentence splitting (performed before start aligning at sentence level), sentence aligning and tokenization (done after aligning at the sentence level) were performed on the parallel documents to retain and convert the overall content to a valid, suitable format for the system. The researchers used Hunalign aligner to align at sentence level and they

found a BLEU score of 35.32% and 0.32%. The performance can increase by applying morphological analyzer and generator for Amharic language [13].

Another study conducted involving Amharic and English languages in a bidirectional form using a statistical approach is by Eleni Teshome. She collected 1020 simple and 1951 complex, separate, parallel sentences for the two languages each from various resources. Then she performed two different experiments on these simple and complex sentences of the Amharic and English languages. The BLEU score for simple sentences is 82.22% and 73.38% from English-Amharic and Amharic-English respectively. For complex sentences, the BLEU score evaluation shows a result of 73.38% from English-Amharic and 84.12% from Amharic-English translations. From the result of the experiment, we can conclude that the translation system better performs when translating from Amharic-English language. However, she failed in using the testing data again in training data this in turn raises a reliability question on her system performance [11].

A team of researchers from different universities in Ethiopia studied on bidirectional English-Ethiopian languages statistical machine translation [58, 59]. They selected five languages from Semitic (Amharic, Tigrigna and Ge'ez), Cushitic (Afan-Oromo), and Omotic (Wolaytta) language families. Corpuses are collected from different sources of religious, historical and legal domains. As they stated the performance of the statistical machine translation greatly affected by the morphological richness of the languages and the linguistic features of the target languages. These features include the writing system, word ordering and morphological complexity. The collected corpuses then passed through a series of preprocessing stages like Character normalization, Sentence tokenization and Alignment. In this study, the SMT system from Ethiopian language-English languages have a higher BLEU score result than that of English-Ethiopian languages. The one-to-many alignment when English is used as a target language favors on the better performance of the system. The other reason for the better performance is the suitability of the language model for English language since it's not morphologically complex as Ethiopian languages.

A study on Ge'ez to Amharic automatic machine translation system using statistical machine translation performed by Dawit in the aim of providing a means of knowing Ge'ez language. The data that is collected from various resources only includes spiritual books i.e. biblical data since most of the time literatures on Ge'ez language are limited to some religion especially Ethiopian orthodox church. To overcome the alignment problem of some sentences from the corpus, he

manually aligned at verse and sentence level. As the researcher stated, the result of the translation gets a higher score when the testing data is large and gets a low score while the data selection for testing becomes small. To rectify the problem, he split each book of the bible into training and testing set in order to check the performance of the system. The evaluation of the designed system got a BLEU score of 8.26% after conducting an experiment on the collected data using a statistical approach. Dawit suggests more works to conduct on morpheme level since these languages are morphological rich. The translation performance will increase after applying morphological segmentation and synthesizing mechanisms [15].

Tadesse Kassa added a study on Ge'ez and Amharic languages. He designed a morpheme-based bi-directional machine translation system from Ge'ez to Amharic and vice versa. The study emphasizes on morphemes, smallest grammatical units, as both languages are morphologically rich in their nature. As the research states, at word level there is a data scarcity, difficulty to manage many forms of a single word, not specific and lacks consistency but the morpheme level overcomes these limitations. Parallel corpuses consisting of 13,833 sentences for each language gathered from various resources of religious books. These data passed through preprocessing tasks such as tokenization, cleaning and normalization to facilitate the efficiency of the translation system. After the experimentation of the system, a BLEU score evaluation of 15.14% and 16.15% found from Ge'ez-Amharic and Amharic-Ge'ez respectively [16].

Akubazgi Gebremariam who conducted a research on Amharic-Tigrigna machine translation system is among the mentioned researches on machine translation system on Ethiopian languages. Despite Amharic and Tigrigna have the same language family and used similar sentence structure; there is also a big difference in building different types of phrases. Hence, they used a hybrid approach that involves a statistical and rule-based to fill the void that arises from some phrases in the languages. The research follows a POS tagging mechanism on source language (Amharic) along with preparing 19 different tag sets for each word in the sentence. After tagging with a tag sets, local reordering rule applied on source language in order to make its sentence structure more like the target language's sentence structure. The developed system under this study provides a BLEU score evaluation of 7.02% and 17.47% after conducting a separate experiment on statistical and hybrid approaches respectively [14].

Chapter Four

Design of Ge'ez to Amharic Machine Translation

4.1 Overview

To meet the goal of this study that is building a translation system from Ge'ez to Amharic language using hybrid approach, a fully functioned system architecture developed. Data used for the experiment properly collected, preprocessed and language and translation model built. Different rules to guide the translation process also drafted. In this chapter, we will see how the proposed system works in detail.

4.2 Architecture of the system

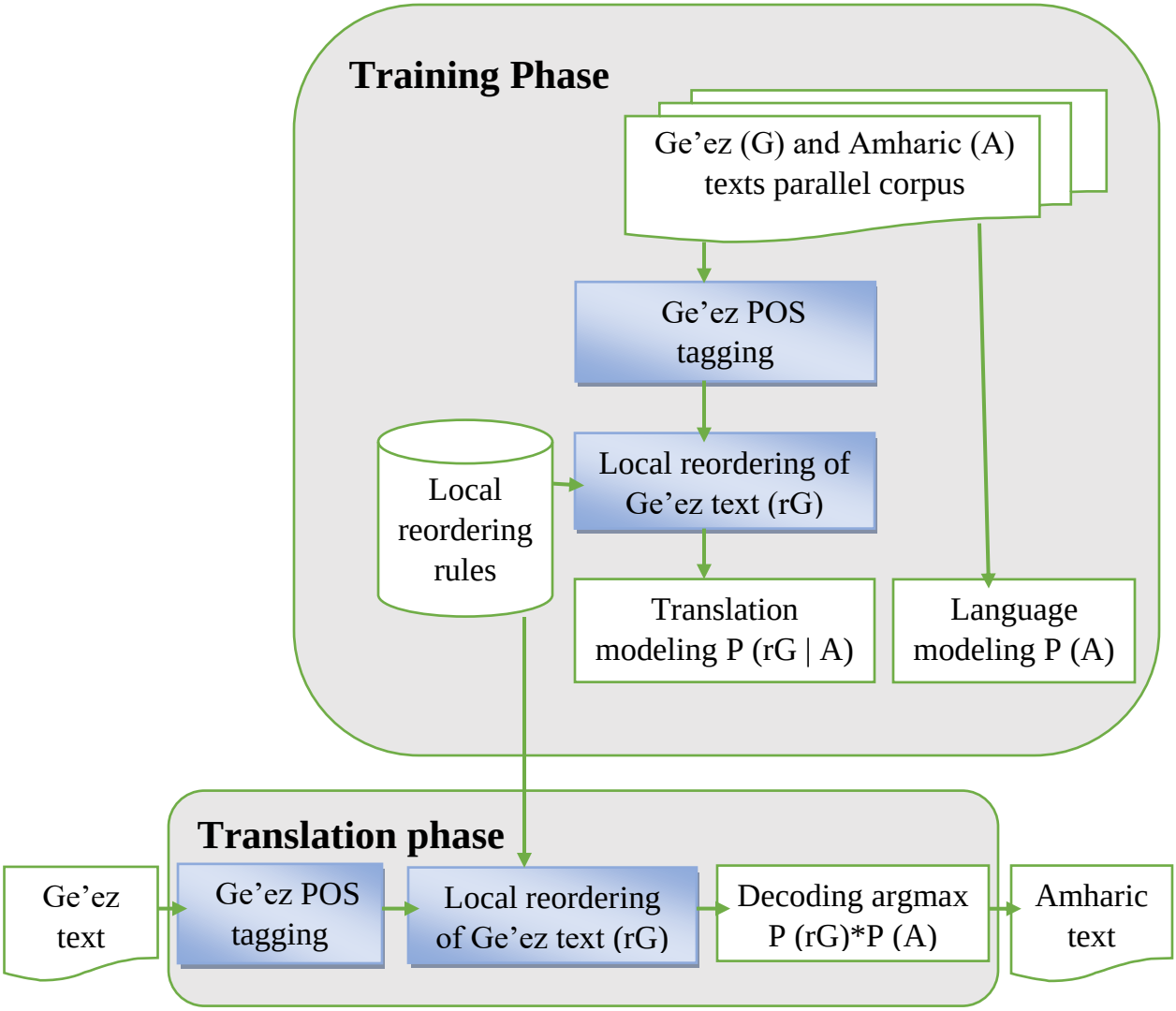
As we described in chapter 2 Figure 6 of the general architecture of statistical machine translation or statistical MT which is based on statistical model takes some form of source text which passes through different preprocessing to find the most probable target sentence. Language modelling, which bases on the target language, is responsible for describing how words arranged. Translation modeling takes the source and target language to compute a probability of source text/sentence given the translated sentence/text. The role of the decoder on the other hand is to take the source language and provide a best translation based on the product of translation and language model.

The architecture of our system designed to take bilingual and monolingual corpuses as an input and the data from the corpuses preprocessed with different preprocessing tools. We put POS tags on each words of our sentences then locally reordered them to look alike the sentence structure of the target language. The translation model takes the bilingual corpus and segment it into several sequences of consecutive words. The languages model takes the target language or in this case the Amharic language to determine the word order in the sentence formation. When conducting a decoding operation, the decoder searches the best translation from the given possible translations based on the probability. Tuning is responsible to find the optimal weights that maximize the translation performance on a small set of parallel sentences.

Hence, we followed the general architecture of the statistical machine translation. The only difference is applying a process of POS tagging on each words and local reordering since we are

aiming to build our system based on statistical and rule-based approach. The shaded rectangles in the following figure shows the additives in the general architecture of the statistical machine translation. So, the overall architecture discussed can be summarized in the below figure.

Figure 4. 1: Architecture of the proposed system



4.2.1 Training Phase

Parallel Corpus

A parallel corpus is a corpus that contains a collection of original texts in language L_1 and their translation into a set of languages $L_2 \ldots L_n$ but in most cases, the parallel corpus refers a set of two languages. For the purpose of this study, parallel corpuses that comprised of simple and complex sentences of the two languages, Ge'ez and Amharic, prepared independently. These text files collected from different resources like Bible (old and new testament), Wudasie Mariam, and

Metsehafe Kidase. In order to protect an ambiguity from happening, we carefully adjust the total number of sentences in each corpus to be the same in number.

Translation model

Translation model shows the bilingual relationship between the source and the target languages text from the parallel corpus. For this model, the training corpus is a sentence level aligned corpus for the languages to be involved in the translation phase. Our study uses a locally reordered Ge'ez sentences (rG) with POS tags with their corresponding translations in Amharic language (A). Since it has a problem in accuracy of translation, sentence level translation not recommended instead word, phrase and syntax-based translations are preferable. The translation model takes locally reordered source language (rG) and target language (A) with the probability denoted by P (rG | A).

Language model

Language modelling aims at characterizing, capturing and exploiting the restrictions imposed on the way that words can combine to form sentences and describing how words arranged in a natural language. The language model is always constructed using a target language, so we built our language model on our target language i.e. Amharic language P (A). Language model is applied in different areas like in Automatic Speech Recognition, Character and handwrite recognition, SMT, POS tagging. Statistical language modelling or SLM is one of the type of approaches in language modelling. The approach bases on corpus based probabilistic approach. It predicts the probability of the next word based on a sequence of given words. It applies a chain rule in calculating P (W) that is as a product of conditional probabilities.

$$P(W) = \prod_{i=1}^{n} p(wi/w1, \dots, wi-1)$$

However, SLM has a drawback in calculating conditional probability for all words and all sequence length. To overcome this problem an N-gram model based on Markov's assumption used. The assumption works by predicting the probabilities of a word based on few previous words. N refers the number of words in a sequence and the probability of a word *w* calculated based on N-1 previous words. The N-gram model is useful in this study in a way on the target language to compute the probability of each word. The probability may calculate as a unigram,

bigram and trigram depending on the *n* that we assigned. In general, the N-gram probability uses the following equation.

$$P(W_n/W_1W_2 \ldots W_{n-1})$$

We can elaborate how the N-gram probability model works using the following Ge'ez language sentences.

እግዚአብሔር ታላቅ ነው/God is great

እግዚአብሔር ዓለምን መረጠ/God choses this world

እግዚአብሔር ዓለምን ባረከ/God blessed the world

The unigram probability calculated as follows:

$P(a_1)$ = count (a_1)/total words

P (እግዚአብሔር) = count (እግዚአብሔር)/count (total words)

= 3/9

= <u>0.33</u>

where *a* refers the selected Amharic word to calculate the probability of, *'3'* and *'9'* represent the occurrence of the word '*እግዚአብሔር'* in the sentences and the total number of words respectively.

Similarly, the bigram probability looks like this when computed:

$P(a_2/a_1)$ = count (a_1a_2)/count (a_1)

P (ዓለምን/እግዚአብሔር) = count (እግዚአብሔር ዓለምን)/count (እግዚአብሔር)

= 2/3

= <u>0.67</u>

Here also, a_1, a_2 refers the selected Amharic words. The numbers *'2'* and *'3'* represent the total occurrence of *'እግዚአብሔር ዓለምን'* and '*እግዚአብሔር'* in the sentences respectively.

It is also possible to calculate the trigram probability as follows:

$P(a_3/a_1\ a_2)$ = count $(a_1\ a_2\ a_3)$/count $(a_1\ a_2)$

P (ባረከ/እግዚአብሔር ዓለምን) = count (እግዚአብሔር ዓለምን ባረከ)/count (እግዚአብሔር ዓለምን)

= 1/2

= <u>0.5</u>

Again, *'1'* and *'2'* refer the frequency of appearance of the words *'እግዚአብሔር' 'ዓለምን'* and *'ባረከ'* and the number of times the words *'እግዚአብሔር'* and *'ዓለምን'* appear together respectively.

Ge'ez POS tagging

In order to reorder the source language, Ge'ez, words and phrases of the languages must be tagged with POS tag sets to look alike the sentence structure of the target language (Amharic). As a result, these tagged sentences pass to the next level that is local reordering. As the researcher's knowledge concerned, there is no publicly available POS tagger tool for Ge'ez language. Since Amharic and Ge'ez are of the same language family, Semitic, the tag sets that we are going to use in Ge'ez language are similar with the Amharic language.

The most common tag sets are PN (personal noun), CN (compound word that changes order of words in the target language), N2 (compound words that never change order of words in the target language), N (noun), VN (verbal noun), PRON (pronoun). V (verb), AXU (auxiliary verb), VREL (relative verbs). ADJ (adjective), NUM (number), NUMCR (cardinal number), NUMOR (ordinal number). PRP (prepositions that have similar positional order like the target language), ADV (adverb), PUN (punctuation), CC (conjunctions and subordinate conjunction) and UNC (unclear).

However, for the purpose of this study we use N, NP, NC, NPC, VN, CN, N2 (noun (N), N attached with preposition, conjunction and with both at a time, compound words that change and never change order of words in the target language respectively). ADJ (adjective), PRON, PRONP, PRONC, PRONPC (pronoun (PRON), PRON attached with preposition, conjunction and with both at a time respectively), PRP (preposition), CC (conjunction). ADV (adverb), VN (verbal noun), V, VP, VC, VPC (verb (V), V attached with preposition, conjunction and with both at a time respectively) and VREL (relative verb) [60].

Ge'ez reordering rules

Ge'ez and Amharic belong to the same language family that is Semitic. They have also used a common '*alphabet'.* Despite these similarities, they have a great difference in their sentence structure. Amharic most of the time uses Subject-Object-Verb (SOV) structure while Ge'ez falls into three structures, VSO (Verb-Subject-Object), SVO (Subject-Verb-Object) and OVS (Object-Verb-Subject). The main aim of reordering the words or phrases in a sentence is to overcome the gap of missing a common sentence structure. Reordering rule was applied for machine

translations involving different languages with different sentence structure. In this study, we construct rules on Ge'ez sentences to make reordering on their words/phrases to make them look alike the sentence structure of the Amharic language. Hence, it facilitates the translation process to meet its goal.

In general, on our study we have a total of ten rules on which reordering of words/phrases could happen. General rules that govern our translation system to work smoothly mentioned below.

- ✓ Rule 1: reordering rule involving Verb (V, VP, VC, VPC, and VREL) and Noun (N, NP, NC, and NPC) and Compound Word (CN, N2)
- ✓ Rule 2: reordering rule involving Verb (V, VP, VC, VPC, and VREL) and Adjective (ADJ)
- ✓ Rule 3: reordering rule involving Verb (V, VP, VC, VPC, and VREL) and Pronoun (PRON, PRONP, PRONC)
- ✓ Rule 4: reordering rule involving Verb (V, VP, VC, VPC, and VREL) and Adverb (ADV)
- ✓ Rule 5: reordering rule involving Adverb (ADV) and Adjective (ADJ)
- ✓ Rule 6: reordering rule involving Adverb (ADV) and Pronoun (PRON, PRONP, PRONC)
- ✓ Rule 7: reordering rule involving Adverb and Noun (N, NP, NC, and NPC) and Compound Word (CN, N2)
- ✓ Rule 8: reordering rule involving Pronoun (PRON, PRONP, PRONC) and Noun (N, NP, NC, and NPC) and Compound Word (CN, N2)
- ✓ Rule 9: reordering rule involving Pronoun (PRON, PRONP, PRONC) and Adjective (ADJ)
- ✓ Rule 10: reordering rule involving Noun (N, NP, NC, and NPC) and Compound Word (CN, N2) and Adjective (ADJ)
- ✓ Rule 11: reordering rule involving Compound words with changing word order (CN, CNP)

The above listed rules are highlights regarding with the discovery of this research. We will go through on each point in detail with examples in the coming pages.

Rule 1: reordering rule involving Verb (V, VP, VC, VPC, and VREL) and Noun (N, NP, NC, and NPC) and Compound Word (CN, N2)

A Verb (V) tells us what the subject of the sentence is doing in other word it refers the part of the sentence that holds the action words. They are the main parts of a sentence/phrase along with nouns to describe what is taking place. There is a need to add some -fixes on verbs in order to connect with the sentences' objects. These -fixes might be prepositions, conjunctions or the combination of the prepositions and conjunctions. When Preposition or Conjunction lie on a verb, the resulting word termed as prepositional or conjunctional verb respectively or VP/VC. In addition, verbs can appear by holding a preposition and conjunction at the same time. In this case, the verb thought as Verb with Preposition and Conjunction or VPC. VREL or Verb Relative on the other hand refers the variation of the verbs 'to be' or 'to have' to describe the relationship between two things.

Nouns (N) in a sentence are most of the time referred as 'subjects' that take the ownership of an action (verb). They identify any of class of people, places, or things collectively or particularly in a sentence. Similar to verbs, prepositional, conjunctional or a combination of prepositional and conjunctional –fixes might be applied on nouns to form Noun Preposition (NP), Noun Conjunctions (NC) and Noun with Preposition and Conjunction (NPC). There are also Compound Words that made up on two or more words. They made up with nouns that modified by adjectives or other nouns. When translating compound words, they may or may not have different order in the target language than the source language. In our study if they have different word order in the target language, they tagged us CN and if they do not their tag is N2.

If verbs precede, any of the nouns in the source language (Ge'ez) they need to change their word order when translating from Ge'ez to Amharic language in order to have the same sentence structure with the source language (Amharic). We can see these with examples:

G: ንበትክ/V መኣሥሪሆሙ/N

A: ማሰርያቸውን/N እንበጥስ/V

G: ወተወከሉ/V ለእግዚአብሔር/NP

A: በእግዚአብሔርም/NP ታመኑ/V

G: ወዘኢቆመ/VP ፍኖተ/CNP ታጥአን/CN

A: በታጠኣተኞችም/CNP መንገድ/CN ያልቆመ/VP

As it's seen from the above examples, the word order of Ge'ez sentences does not match with the word orders of Amharic language, so we need to change the word order in the source language to make the same word order with the target language. Based on this after reordering, the Ge'ez sentences with their translation on the Amharic sentences looks like this:

rG: መኣሥሪሆሙ ንበትክ

A: ማሰርያቸውን እንበጥስ

rG: ለእግዚአብሔር ወተወከሉ

A: በእግዚአብሔርም ታመኑ

rG: ኃጥኣን ፍኖተ ወዘኢቆመ

A: በኃጢአተኞችም መንገድ ያልቆመ

The reordering algorithm looks like this:

Algorithm 4. 1: Reordering rule for noun, compound noun and verb

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 2: reordering rule involving Verb (V, VP, VC, VPC, and VREL) and Adjective (ADJ)

Adjectives are words, which describe or modify other words. They tell how much or how many of something that mentioned, which thing someone passed to him, or which kind of something

that someone want. They placed in a sentence before a noun or pronoun since it identifies or quantify individual people and unique things. However, in Ge'ez language it may be found after a verb because of different word structure usage of the language. Here is an example to show the position of the verb and the adjective that needs swapping.

G: ወትሬዕዮ·ሙ/V በበትረ/N ጎጺን/ADJ G: ወዘኢነበረ/VP መንበረ/N መስተሳልፋን/ADJ

A: በብረት/ADJ በትር/N ትጠቢቆቸዋለህ/V A: በዋዘኞችም/ADJ ወንበር/N ያልተቀመጠ/V

Here the positions of the verb and adjective on the two languages differs so there is a need to apply the rule on the Ge'ez sentences to swap this POS tagged words and create a similar word structure with its translation in Amharic sentence. Based on this, the sentences look like this after applying the reordering rule:

rG: ጎጺን በበትረ ወትሬዕዮ·ሙ rG: መስተሳልፋን መንበረ ወዘኢነበረ

A: በብረት በትር ትጠቢቆቸዋለህ A: በዋዘኞችም ወንበር ያልተቀመጠ

The algorithm, which implemented on this rule, is:

Algorithm 4. 2: Reordering rule for verb and adjective

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 3: reordering rule-involving Verb (V, VP, VC, VPC, and VREL) and Pronoun (PRON, PRONP, and PRONC)

A pronoun is a word or phrase that replaces the role of the noun or used as a substitution for a noun or noun phrase. When they used in a sentence, they can do everything a noun do and they are building blocks of sentences. As in noun and verb, prepositional and conjunctional -fixes also used in pronouns. The resulting word called PRON, PRONP or PRONC (pronouns with adjectives, conjunctions and with prepositions and conjunctions respectively). Like adjectives, pronouns may find after verbs so in this case we need to swap these words. Some examples:

G: ወትገድፎሙ/V ይነቡ/PRON ሐሰተ/NP

A: ሐሰትን/NP የሚናገሩትን/ PRON ታጠፋቸዋለህ/V

G: ወንገድፍ/V እምላዕሌነ/PRONP አርዐቶሙ/NP

A: ገመዳቸውንም/NP ከእኛ/PRONP እንጣል/V

One can understand from the examples that the position of the pronouns and verbs are not in the right in Ge'ez sentences as compared with the Amharic sentences. In this case, we need to swap the position in order to have a similar word structure and to have a good translation. Therefore, applying the reordering rule makes our sentences like this:

rG: ሐሰተ ይነቡ ወትገድፎሙ

A: ሐሰትን የሚናገሩትን ታጠፋቸዋለህ

rG: አርዐቶሙ እምላዕሌነ ወንገድፍ

A: ገመዳቸውንም ከእኛ እንጣል

The reordering algorithm is:

Algorithm 4. 3: Reordering rule for verb and pronoun

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 4: reordering rule involving Verb (V, VP, VC, VPC, and VREL) and Adverb (ADV)

An Adverb is part of a sentence that change, modify or qualify different types of words like an adjective, a verb, a clause, another adverb, or any other type of word or phrase. They provide a detail information on how, where, when, in what manner and to what extent something happens. They always modify verbs in a way by giving a more specific information on the action that takes place when they used to do so. If they found after the verb in Ge'ez sentences, they need to swap their position to meet the translation efficiency. Here are some sentences as examples in the two languages with adverbs involving:

G: ወሕዘብኒ/NP ነበቡ/V ከንዶ/ADV

A: ሕዝቡም/NP ከንቱን/ADV ይናገራሉ/V

G: ተቀነዩ/V ለእግዚአብሔር/NP በፍርሀት/ADV

A: ለእግዚአብሔር/NP በፍርሃት/ADV ተገዙ/V

Adverbs as always must come before verbs in Amharic sentences but in Ge'ez sentences this rule does not work. Since we are translating into Amharic, the sentences on Ge'ez must follow its word structure. Therefore, we must swap the positions of verbs and adverbs when the former appears before the later to make the translation smooth.

rG: ወሕዘብኒ ከንዮ ነበቡ

A: ሕዝቡም ከንቱን ይናገራሉ

rG: ለእግዚአብሔር በፍርሀት ተቀነዩ

A: ለእግዚአብሔር በፍርሃት ተገዙ

The algorithm for this rule is:

Algorithm 4. 3: Reordering rule for verb and adverb

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 5: reordering rule involving Adverb (ADV) and Adjective (ADJ)

As it described in the previous section adverbs used to change, modify or qualify different types of words mainly verbs while adjectives are words that are important in describing or modifying other words usually nouns. Since adverbs placed mostly before verbs and adjectives before nouns, if adjectives preceded by adverbs in Ge'ez sentences swapping rules must takes place in order to achieve the word order of Amharic sentence. The below sentence is an example for this case.

G: ቀስቶሙ/ADV ናሁ/ADJ ኃጥኣን/N ወሰቁ/V G: በጊዜሁ/ADV ኩሉ/ADJ ሰብእ/N የኀልፍ/V

A: እነሆ/ADJ ኃጢአተኞች/N ቀስታቸውን/ADV ገትረዋልና/V A: ሁሉም/ADJ ሰው/N በጊዜው/ADV ያልፋል/V

As the POS tag indicates, the underlined words are of the class of adjectives and adverbs. The Ge'ez sentence must have the same word order with Amharic sentence so after applying the reordering rule the above sentence looks like this:

rG: ናሁ ኃጥኣን ቀስቶሙ ወሰቁ rG: ኩሉ ሰብእ በጊዜሁ የኀልፍ

A: እነሆ ኃጢአተኞች ቀስታቸውን ገትረዋልና A: ሁሉም ሰው በጊዜው ያልፋል

The algorithm for this reordering rule is:

Algorithm 4. 4: Reordering rule for adverb and adjective

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 6: reordering rule involving Adverb (ADV) and Pronoun (PRON, PRONP, and PRONC)

Pronouns as we discussed are words or phrases that replace the role of the noun or used as a substitution for a noun or noun phrase. Due to different word order, structure of the Ge'ez language pronouns may find after adverbs, which is not the case in the word structure of Amharic language. Therefore, there is a need to have a rule to govern this mismatch when translating to Amharic language. Here is example to show the necessity of this rule:

G: ኵሉሙ/ADV ይትፈሥሑ/V ብከ/PRONP

A: ባንተ/PRONP ሁሉም/ADV ደስ ይላቸዋል/V

G: ሠናይቶ/ADV መኑ/PRON ያርእየነ/V

A: ማን/PRON መልካሙን/ADV ያሳየናል/V

Pronouns in Amharic sentence placed always before the position of the adverbs. Hence, in our translation from Ge'ez to Amharic we need to consider swapping sentences with this kind of arrangement in Ge'ez. Based on this the above example sentence will look like this after applying the reordering rule:

rG: ኵሉሙ ይትፈሥሑ ብከ

A: ባንተ ሁሉም ደስ ይላቸዋል

rG: መኑ ሠናይቶ ያርእየነ

A: ማን መልካሙን ያሳየናል

The reordering algorithm is:

Algorithm 4. 5: Reordering rule for adverb and pronoun

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 7: reordering rule-involving Adverb (ADV) and Noun (N, NP, NC, and NPC) and Compound Word (CN, N2)

Nouns in the order of Amharic sentence must come first when compared with adverbs since most of the time nouns are the subjects of the sentence and adverbs used to modify verbs. When nouns and compound words appear by succeeding adverbs this reordering rule swaps these words that belongs to the class of adverbs and nouns and compound words. The below example clearly illustrates this case.

G: የጎልቅ/V እከዮሙ/ADV ለታጥኣን/N

A: የታጥኣን/N ክፋት/ADV ይጥፋ/V

G: ሰላመ/ADV ይሁብ/V ቤተ/CNP እግዚአብሔር/CN

A: የእግዚአብሔር/CNP ቤት/CN ሰላምን/ADV ይሰጣል/V

As one can see from the examples, the positions of nouns, compound nouns and adverbs is not in the right position when compared with the word order of the Amharic sentences so in this case we need to apply our rule to produce a sentence structure resembling to the Amharic sentence.

rG: ለታጥኣን እከየሙ የኅልቅ

A: የታጥኣን ክፋት ይጥፋ

rG: ቤተ እግዚአብሔር ሰላመ ይሁብ

A: የእግዚአብሔር ቤት ሰላምን ይሰጣል

The algorithm for this rule is:

Algorithm 4. 6: Reordering rule for verb and adjective

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 8: reordering rule-involving Pronoun (PRON, PRONP, PRONC, and PRONC) and Noun (N, NP, NC, and NPC) and Compound Word (CN, N2)

The main function of pronouns in a sentence is to replace the function of a noun. When it appears within a sentence, it can function and consider as a noun. Nouns as mentioned most of the time they are the subject of the sentence. Every sentence has a subject that takes the responsibility of the action that occurred at some point of time. Compound words are a combination of two words to produce a new meaning. When pronouns appear before nouns, swapping must exist. Here is an example to elaborate this kind of situations.

G: አንተ/PRON እግዚኦ/N ዕቀበነ/V

A: አቤቱ/N አንተ/PRON ጠብቀን/V

G: ተፈነወ/V ለነ/PRON መንፈስ/N2 ቅዱስ/N2

A: መንፈስ/N2 ቅዱስ/N2 ለኛ/PRON ተላከ/V

Reordering the sentence with this rule yields:

rG: እግዚኦ አንተ ዕቀበነ

A: አቤቱ አንተ ጠብቀን

rG: መንፈስ ቅዱስ ለነ ተፈነወ

A: መንፈስ ቅዱስ ለኛ ተላከ

The algorithm for this rule is:

Algorithm 4. 7: Reordering rule for pronoun and noun, compound noun

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 9: reordering rule involving Pronoun (PRON, PRONP, and PRONC) and Adjective (ADJ)

Adjectives can modify subjects or nouns and pronouns. They have also the ability to act as a compliment to linking verbs or the verb to be. The interesting point to add on this is sometimes a word that used as a noun may found to be an adjective based on its placement on the sentence. In some Ge'ez language sentences, pronouns may appear before adjectives that needs to swap its position when translating it to Amharic language. Let us have a look this example:

G: ወኪየከ/PRON እሴፎ/V ኵሎ/ADJ አማረ/N

A: ሁሉን/ADJ ቀን/N አንተን/PRON ተስፋ አድርጌአለሁ/V

After swapping the two underlined words from the Ge'ez sentence, the sentence will have the same order with that the Amharic sentence word order.

rG: ኩሎ አሚረ ወኪየከ እሴፎ

A: ሁሉን ቀን አንተን ተስፋ አድርጌአለሁ

Its reordering algorithm will be:

Algorithm 4. 8: Reordering rule for pronoun and adjective

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 10: reordering rule involving Noun (N, NP, NC, and NPC) and Compound Word (CN, N2) and Adjective (ADJ)

Nouns may find in the sentences with different forms. Prepositions and conjunctions or a combination of prepositions and conjunctions could lie and make nouns to change their forms but never change their word class. Nouns and compound words in Ge'ez language sentences may precede adjectives in their word order structure. However, when translating the Ge'ez sentence to Amharic adjectives and nouns must swap their positions. The below example shows the necessity of this reordering rule.

G: ወማኅበረ/N አሕዛብኒ/ADJ የፀውዶከ/V

A: የአሕዛብም/ADJ ጉባኤ/N ይከብብሃል/V

G: ወዘኢነበረ/VP መንበረ/N መስተሳልቃን/ADJ

A: በዋዘኞችም/ADJ ወንበር/N ያልተቀመጠ/V

The underlined words with their POS tags must swap their position from the Ge'ez sentence.

rG: አሕዛብኒ ወማኅበረ የዐውደከ

A: የአሕዛብም ጉባኤ ይከብብሃል

rG: መስተሳልቃን መንበረ ወዘኢነበረ

A: በዋዘኞችም ወንበር ያልተቀመጠ

The algorithm used for this rule is:

Algorithm 4. 9: Reordering rule for noun, compound noun and adjective

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Rule 11: reordering rule involving Compound words with changing word order (CN, CNP)

Compound words as it stated in the above sections is a combination of two words to form a new, meaningful word. When these words translated into Amharic language from Ge'ez language, there may be a need of swapping their word order. Let us see some example for this case:

G: ፍኖተ/CNP ጎጥአን/CN

A: የጎጢአተኞች/CNP መንገድ/CN

G: ነገሥተ/CNP ምድር/CN

A: የምድር/CNP ነገሥታት/CN

After swapping the word orders for Ge’ez sentences, the example looks like this:

rG: ኃጥኣን ፍኖተ rG: ምድር ነገሠተ

A: የኃጢአተኞች መንገድ A: የምድር ነገሥታት

The algorithm will be:

Algorithm 4. 10: Reordering rule for compound noun

```
1: function Load_Ge’ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge’ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

In general, rules that are discovered on this study to meet its goal play an important role in translating Ge'ez words or phrases into Amharic language by arranging Ge'ez words position in a sentence or phrase whenever it is needed using the POS tags attached to them. To give priority to high order word classes in a sentence we use the dictionary method. The method takes the order by number from largest to smallest and act accordingly whenever swapping of words necessitates. As it shown in the above algorithms, they look like the same but what makes the difference is their priority in the dictionary that is used. The overall algorithm looks like this:

Algorithm 4. 11: The overall algorithm for the reordering rule

```
1: function Load_Ge'ez_Sentence (S)
2:     data_reader ← read_data from the Load_Ge'ez_Sentence
3:     words ← split ( data_reader, new line)
4:     dict ← dictionary_value (words, Wj)
5:     for i=0 to size (words) do
6:         for j=0 to size (words) do
7:             pos1 ← split (words [i] [j], tab_space) [1]
8:             pos2 ← split (words [i] [j], tab_space) [1]
9:             if dict [pos1] > dict [pos2] and not j==0
                   Temp ← word [i] [j]
                   word [i] [j]← word [i] [j-1]
                   word [i] [j-1]← Temp
10:            end if
11:        end for
12:    end for
13: end function
```

Repeat the above process of swapping all words until they placed on the right position

Since there are multiple lines in our corpus and multiple words in each line, the 'for loop' never stops until all the words that need swapping get the right position. The above-mentioned steps are all parts of the training phase. The next crucial phase of this process is the translation phase.

4.2.2 Translation Phase

Ge'ez Input text

For the translation to be conducted first the input language set (in this case the Ge'ez language sentences) must be provided since the translation is from one source language to the other target language. Based on this we prepared Ge'ez language corpus for this translation as a source language then tagged with different POS tag sets and pass through appropriate reordering rule to resemble the training model so that the reordered text has the form rG. Then the translation system or the decoder accepts this reordered text to translate it into a better output on the target language.

Decoder

The job of the decoder is to take a source language and translates it into its corresponding target language according to the product of translation model that consists both language sets and language model, which have only the target language. The main problem in translation is to find word/phrase that maximizes the translation and language model probabilities. To do so the decoder uses a best first search approach. The decoder looks all the possible translations of source word/phrase from word or phrase translation table and recombine the target language word or phrase that maximizes the translation model probability with the language model probability.

For our research, the translation model takes locally reordered Ge'ez sentence with target Amharic sentence so that the decoder takes the reordered Ge'ez sentence to translate it into its corresponding Amharic language sentence. To put it mathematically,

$$\mathrm{A} = \underset{\boldsymbol{a}}{\mathbf{argmax}}\ P\ (r\mathrm{G}|\mathrm{A}) * P\ (\mathrm{A})$$

$P\ (r\mathrm{G}|\mathrm{A})$ Ge'ez-to-Amharic translation model
$P\ (\mathrm{A})$ Amharic language model

Amharic Output Text

For our translation system, the target language is Amharic language that gives the corresponding word/phrase for the source language i.e. Ge'ez.

Chapter Five

Experiment

5.1 Overview

The designed architecture that is mentioned in the previous chapter must be evaluated with a set of data in order to check its performance. Thus, in this chapter we are going to conduct two independent experiments and discuss on the results found after the experiments. The first experiment is based statistical approach only while the second is a combination of rule-based and statistical machine translation approach.

5.2 Data collection

Large amount of data, monolingual and bilingual, needed to conduct statistical machine translation. Monolingual corpus is required to estimate the right word order to guide the target languages to resemble the source language while a sentence aligned bilingual corpus helps to build the translation model training and decoding to determine the word alignment between two aligned sentences [15]. For the purpose of this research, we collect our corpus (a total of 2009 parallel sentences for both languages) from different online sources including https://www.ethiopicbible.com, http://ethiopianorthodox.org, and http://eotcmk.org. These sources contain parallel data of Ge'ez and Amharic language from the Holy Bible with PDF format that makes suitable for statistical and rule-base machine translations. In addition, texts are collected from other spiritual books like Wudasie Mariam and Metsehafe Kidase similarly with having parallel data of the two languages.

5.2.1 Data Preprocessing and Preparation

Throughout the preparation of the parallel data, there were many challenges that can limit the performance of the system. To overcome these, we follow different mechanisms.

- Breaking the document in sentence level to make separate sentences appear on separate lines and corresponding Ge'ez and Amharic documents on different files.
- Misaligned sentence verses which means they exist in the data but in a wrong place. This cause the entire forthcoming sentences aligned with different, unrelated sentence of the target language or vice versa. This problem rectified manually with the help of experts.

- Duplication of a verse in the two languages solved by manually searching the case and removing since duplication alters the reliability of the system.

After passing through the data collection and preprocessing steps, we used notepad tool to organize the texts we got from the above-mentioned sources in different files for each language. To make the prepared corpus ready for training and testing of the proposed translation system the following two processes conducted.

Tokenization: refers inserting of spaces between words

Cleaning: includes removing of empty, misaligned and long sentences. The occurrence of these can cause a potential problem during training. It cuts long sentences into small, suitable ones and removes unusual spaces between words and sentences.

When preparing the data first we align all the chapters in verse level then merge these verse level aligned chapters in the two languages differently. In total, we have 2,009 aligned parallel sentences. From these, to conduct four independent experiments we allocated slightly more than 90% and 80% for training since training the system is the crucial thing in getting a better translation from the system and approximately 10% and 20% for testing the system [15]. Putting in number, the training set used for training the translation model consists 1800 and 1600 sentences and 209 and 409 sentences respectively allocated to test the system for the collected parallel data set of each language.

One of the aims of this research paper is to show the importance of using hybrid approach over statistical approach alone in machine translation. Thus, we conducted two experiments one for showing the result obtained from using statistical machine translation approach alone and the other is for hybrid machine translation of rule-based and statistical approaches. We used similar tools for training and testing in both experiments. The next sections discuss the overall steps and the results obtained from the experiments.

5.3 Experiment 1: Statistical approach

Our first experiment is conducted based on statistical approach by taking 90% for training and 10% for testing from the total size of our corpus. The approach uses a probability method to give a best translation and it bases on statistical models that finds the most probable target text given a source text.

5.3.1 Training the translation System

As mentioned in the previous section we have used in total 2009 sentences to conduct the training and testing for each language. From the total number of aligned corpus of the languages, 1800 sentences used for training and the rest are for testing the proposed system. During the training process of Ge'ez to Amharic machine translation system, Moses, a freely available tool is used. There following steps are performed under training the system.

Language Model Training

Language model is used to determine the probability of a sequence of words of the target language by taking a monolingual corpus to ensure fluency of the output. In our case, the target language is Amharic, so we built our model on it. In this study, SRILM, language-modelling toolkit, is used.

Training the system

Up to now, lexicalized reordering tables and Moses configuration file are created with the use of word-alignment, phrase extraction and scoring. On this step the file 'moses.ini' that is responsible for decoding is created. The phrase table, mainly contains the probabilities of a word following words to the given word, was also created. MGIZA toolkit used for system training.

Tuning

After the creation of the file 'moses.ini' for decoding, it's possible to immediately start the querying process but weight optimization is needed since the weights used by Moses to weight the different models against each other are not optimized. To rectify the problem and get a better weight the translation system must be tuned. This step also creates another '.ini' file for decoding. The above-mentioned steps are all conducted for training the system. After all this, the testing process takes place to evaluate the performance of the translation system.

5.3.2 Result of Experiment 1

For testing the performance of the translation system, we have used 209 Ge'ez and Amharic parallel sentences. The performance is measured in terms of translation accuracy to translate a single Ge'ez sentence to Amharic sentence. For evaluation purpose, BLEU score methodology that is discussed in chapter 2 was used. After passing through this evaluation process, our developed translation system got a BLEU score of 7.36% using statistical method i.e. from the

overall corpus size the mentioned amount correctly translated from Ge'ez to Amharic texts. The below figure shows the result obtained on this experiment.

Figure 5. 1: Experimental result of statistical approach I

```
sami@sami-Satellite-L755:~$ /home/sami/SMT/moses-on-nov-07-2019/moses-script/gener
ic/multi-bleu.perl /home/sami/smt-translation/data/test.104.np.tok.lc.am < /home/s
ami/smt-translation/translated/translated.ge-am
BLEU = 7.36, 22.8/7.6/4.8/3.5 (BP=0.997, ratio=0.997, hyp_len=368, ref_len=369)
```

The BLEU score is low because we used a minimum of data.

5.4 Experiment 2: Statistical approach

Our first experiment is conducted based on statistical approach by taking 80% for training and 20% for testing from the total size of our corpus. The approach uses a probability method to give a best translation and it bases on statistical models that finds the most probable target text given a source text.

5.4.1 Training the translation System

As mentioned in the previous section we have used in total 2009 sentences to conduct the training and testing for each language. From the total number of aligned corpus of the languages, 1600 sentences used for training and the rest are for testing the proposed system. During the training process of Ge'ez to Amharic machine translation system, Moses, a freely available tool is used. There following steps are performed under training the system.

Language Model Training

Language model is used to determine the probability of a sequence of words of the target language by taking a monolingual corpus to ensure fluency of the output. In our case, the target language is Amharic, so we built our model on it. In this study, SRILM, language-modelling toolkit, is used.

Training the system

Up to now, lexicalized reordering tables and Moses configuration file are created with the use of word-alignment, phrase extraction and scoring. On this step the file 'moses.ini' that is responsible for decoding is created. The phrase table, mainly contains the probabilities of a word following words to the given word, was also created. MGIZA toolkit used for system training.

Tuning

After the creation of the file ‘moses.ini’ for decoding, it’s possible to immediately start the querying process but weight optimization is needed since the weights used by Moses to weight the different models against each other are not optimized. To rectify the problem and get a better weight the translation system must be tuned. This step also creates another ‘.ini’ file for decoding. The above-mentioned steps are all conducted for training the system. After all this, the testing process takes place to evaluate the performance of the translation system.

5.4.2 Result of Experiment 2

For testing the performance of the translation system, we have used 409 Ge’ez and Amharic parallel sentences. The performance is measured in terms of translation accuracy to translate a single Ge’ez sentence to Amharic sentence. For evaluation purpose, BLEU score methodology that is discussed in chapter 2 was used. After passing through this evaluation process, our developed translation system got a BLEU score of 7.15% using statistical method i.e. from the overall corpus size the mentioned amount correctly translated from Ge’ez to Amharic texts. The below figure shows the result obtained on this experiment.

Figure 5. 2: Experimental result of statistical approach II

```
sami@sami-Satellite-L755:~/SMT-II/ge-to-am$ /home/sami/SMT/moses-on-nov-07-2019/
moses-script/generic/multi-bleu.perl /home/sami/SMT-II/data/test.209.np.tok.lc.a
m < /home/sami/SMT-II/translated/translated.ge-am
BLEU = 7.15, 9.7/6.9/6.8/5.9 (BP=0.991, ratio=0.991, hyp_len=688, ref_len=694)
sami@sami-Satellite-L755:~/SMT-II/ge-to-am$
```

The BLEU score is low because we used a minimum of data.

5.5 Experiment 3: Hybrid approach

This is the second experiment conducted on Ge’ez to Amharic machine translation system. We applied the reordering rules mentioned on chapter 4 on training and testing data sets so both data sets are ready for training and testing of the proposed translation system. There is no difference in training and testing steps of hybrid machine translation approach with that of statistical approach. The reason behind is the rules are applied before the training and testing steps in both approaches. During translation, the reordering rules are applied on Ge’ez tagged sentences to have a similar sentence structure with the Amharic text. All the POS tagging labels removed once the reordering of words takes place successfully. After applying the reordering rules and

Ge'ez sentences get the same sentence structure with the Amharic sentences, we finally apply the statistical approach on the well-prepared and reordered dataset.

5.5.1 Training the translation System

As mentioned in the previous sections we have used in total 2009 sentences to conduct the training and testing for each language. From the total number of aligned corpus of the languages, 1800 sentences used for training and the rest are for testing the proposed system. During the training process of Ge'ez to Amharic machine translation system, Moses, a freely available tool is used. There following steps are performed under training the system.

Language Model Training

Language model is used to determine the probability of a sequence of words of the target language by taking a monolingual corpus to ensure fluency of the output. In our case, the target language is Amharic, so we built our model on it. In this study, SRILM, language-modelling toolkit, is used.

Training the system

Up to now, lexicalized reordering tables and Moses configuration file are created with the use of word-alignment, phrase extraction and scoring. On this step the file 'moses.ini' that is responsible for decoding is created. The phrase table, mainly contains the probabilities of a word following words to the given word, was also created. MGIZA toolkit used for system training.

Tuning

After the creation of the file 'moses.ini' for decoding, it's possible to immediately start the querying process but weight optimization is needed since the weights used by Moses to weight the different models against each other are not optimized. To rectify the problem and get a better weight the translation system must be tuned. This step also creates another '.ini' file for decoding. The above-mentioned steps are all conducted for training the system. After all this, the testing process takes place to evaluate the performance of the translation system.

5.5.2 Result of Experiment 3

For testing the performance of the translation system, we have used 209 Ge'ez and Amharic parallel sentences. The performance is measured in terms of translation accuracy to translate a single Ge'ez sentence to Amharic sentence. For evaluation purpose, BLEU score methodology

that is discussed in chapter 2 was used. After passing through this evaluation process, our developed translation system got a BLEU score of 18.62% using statistical method i.e. from the overall corpus size the mentioned amount correctly translated from Ge'ez to Amharic texts. The below figure shows the result obtained on this experiment.

Figure 5. 3: Experimental result of hybrid approach I

```
sami@sami-Satellite-L755:~$ /home/sami/SMT/moses-on-nov-07-2019/moses-script/generic/mult
i-bleu.perl /home/sami/Desktop/hmt/data/test.104.np.tok.lc.am < /home/sami/Desktop/hmt/tr
anslated/translated.ge-am
BLEU = 18.62, 35.0/20.5/16.4/11.0 (BP=0.984, ratio=0.984, hyp_len=363, ref_len=369)
```

5.6 Experiment 4: Hybrid approach

This is the second experiment conducted on Ge'ez to Amharic machine translation system. We applied the reordering rules mentioned on chapter 4 on training and testing data sets so both data sets are ready for training and testing of the proposed translation system. There is no difference in training and testing steps of hybrid machine translation approach with that of statistical approach. The reason behind is the rules are applied before the training and testing steps in both approaches. During translation, the reordering rules are applied on Ge'ez tagged sentences to have a similar sentence structure with the Amharic text. All the POS tagging labels removed once the reordering of words takes place successfully. After applying the reordering rules and Ge'ez sentences get the same sentence structure with the Amharic sentences, we finally apply the statistical approach on the well-prepared and reordered dataset.

5.6.1 Training the translation System

As mentioned in the previous sections we have used in total 2009 sentences to conduct the training and testing for each language. From the total number of aligned corpus of the languages, 1600 sentences used for training and the rest are for testing the proposed system. During the training process of Ge'ez to Amharic machine translation system, Moses, a freely available tool is used. There following steps are performed under training the system.

Language Model Training

Language model is used to determine the probability of a sequence of words of the target language by taking a monolingual corpus to ensure fluency of the output. In our case, the target

language is Amharic, so we built our model on it. In this study, SRILM, language-modelling toolkit, is used.

Training the system

Up to now, lexicalized reordering tables and Moses configuration file are created with the use of word-alignment, phrase extraction and scoring. On this step the file 'moses.ini' that is responsible for decoding is created. The phrase table, mainly contains the probabilities of a word following words to the given word, was also created. MGIZA toolkit used for system training.

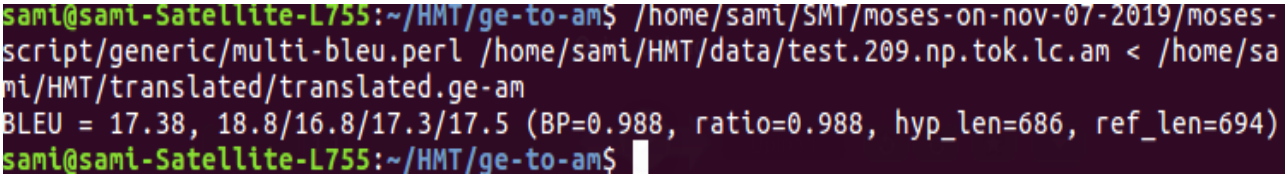
Tuning

After the creation of the file 'moses.ini' for decoding, it's possible to immediately start the querying process but weight optimization is needed since the weights used by Moses to weight the different models against each other are not optimized. To rectify the problem and get a better weight the translation system must be tuned. This step also creates another '.ini' file for decoding. The above-mentioned steps are all conducted for training the system. After all this, the testing process takes place to evaluate the performance of the translation system.

5.6.2 Result of Experiment 4

For testing the performance of the translation system, we have used 409 Ge'ez and Amharic parallel sentences. The performance is measured in terms of translation accuracy to translate a single Ge'ez sentence to Amharic sentence. For evaluation purpose, BLEU score methodology that is discussed in chapter 2 was used. After passing through this evaluation process, our developed translation system got a BLEU score of 18.62% using statistical method i.e. from the overall corpus size the mentioned amount correctly translated from Ge'ez to Amharic texts. The below figure shows the result obtained on this experiment.

Figure 5. 4: Experimental result of hybrid approach II



5.7 Discussion

As described on the previous sections the main aim of this study is to show the machine translation performance using hybrid approach by developing a system. During this study, we conducted two independent experiments to show how the performance of the proposed system varies when using hybrid machine translation approach rather statistical machine translation approach alone. As it can be seen from the BLEU score of the two experiments, the hybrid approach has performed well and provide a better result. This is because of the reordering rules we applied on Ge'ez sentences to have a same sentence structure with that of the Amharic sentence pair. However, the size of the corpus has an impact on the performance of the proposed system since statistical machine translation approach takes bilingual corpus. When the size of the corpus increases the accuracy also increases and so does the BLEU score.

This study shows an improvement from previous studies on Ge'ez-Amharic language pair. Dawit [15] and Tadesse [16] got a BLEU score of 8.26% and 15.14% approximately as machine translation from Ge'ez to Amharic machine translation concerned with different approaches. However, our system gets a BLEU score of 18.62% with minimum amount of parallel data that is better for Ge'ez to Amharic language translation as compared with the studies conducted in the mentioned research papers.

Chapter Six

Conclusion and Recommendations

6.1 Conclusion

This study focuses on hybrid machine translation approach i.e. a combination of rule-based and statistical machine translation approaches. Rules are pointed out to govern the translation process from Ge'ez to Amharic language. We have discussed the historical background of Ge'ez and Amharic languages. In addition, we have discussed the linguistic relationships between both languages including writing system, syntax, numbering system, and word classes. In general, similarities and differences between the two languages are discussed.

For this research, parallel corpuses for both languages are collected from different sources and all the sources are spirituals books since Ge'ez language is currently widely used and limited in the Ethiopian Orthodox Church literatures. The corpus is prepared and organized into two different files for each language and divided into two sets of training and testing. POS tagging, applying reordering rules on Ge'ez sentences with the help of python programming language, language modelling using SRILM, translation modelling with MGIZA and training the translation system using Moses are the tools and mechanisms used during this research work.

In preparing the reordering rules, we consider the differences in syntactic structure between the two languages. The reordering rule is applied by means of POS tagging. Since there is no publicly available POS tag tool for Ge'ez language, we used a manual mechanism to tag all the words in the sentences. The main purpose of setting out reordering rules on Ge'ez sentences is to have the same sentences structure with the Amharic sentences since the translation is unidirectional that is from Ge'ez to Amharic.

After all things and preliminary conditions set, the last step is testing the proposed, developed system. In this study, four experiments were conducted in order to check the accuracy of our translation system. We got a BLEU score of 7.36% and 7.15% from two experiments in statistical approach by changing the training and testing data set sizes and 18.62% and 17.38% from hybrid machine translation approach. From this, we conclude that using hybrid approach for machine translation gives a best result as compared with statistical machine translation.

6.2 Recommendations

For a translation system to be considered as more accurate and efficient, there are many things to be fulfilled. These things are by themselves have a capability to become problem areas and create a room for further research to be conducted around them.

The below mentioned points are possible areas of research as a future work:

- ✓ Increasing the size of the corpus has a direct relation with the accuracy of the translation and performance of the developed system. Therefore, Ge'ez to Amharic translation using POS tags may perform better when there is more data collected.
- ✓ It is also possible to work on Speech to text and text to speech translation since the translation is from Ge'ez to Amharic it could help a lot for proceeding research.
- ✓ In this research, we only applied reordering rules to resemble the source language into the target language since there exist a structural difference. However, it is possible to add rules like morphological rules.
- ✓ Based on our study, rule-based approach using POS tagging makes a research to have a better result when machine translation concerned for other Ethiopian language pairs. There have been good improvements on Ethiopian languages but still a lot to do.
- ✓ The main challenge of this research paper is to find a standard, pre-collected corpus and well-prepared POS tag sets. These problem areas have a potential to be explored more and work on as a research idea.
- ✓ The POS tagging mechanism we used by reordering words in the sentence can also be applied for the bidirectional machine translation from Ge'ez to Amharic and Amharic to Ge'ez language translation.

References

[1] R. Ferrer, Cancho and R. V. Sole´, "The small world of human language," Proc. Roy. Soc. London Ser. B, 268 (2001), pp. 2261–2265.

[2] E. Cambria and B. White, "Jumping NLP curves: A review of natural language processing research." *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 48–57, 2014.

[3] S. Jones. "Natural language processing: a historical review." *Linguistica Computazionale,* vol. 9, pp. 3–16, 1994.

[4] Okpor MD. "Machine translation approaches: issues and challenges." *International Journal of Computer Science Issues (IJCSI)* 11, no. 5 (2014): 159.

[5] SIL international. "Ethnologue language of the world." Internet: www.ethnologue.com/country/ET, last visited August 17, 2018.

[6] EPRDF. "Constitution of the Federal Democratic Republic of Ethiopia", Internet:

www.wipo.int/edocs/laws/en/et/et007en.pdf, December 8, 1994.

[7] Takahashi S., Wada H., Tadenuma R., and Watanabe S. "English Japanese Machine Translation." *Information Processing*, Butterworths Scientific Publications, London, 1960.

[8] Preslav Nakov. "Improving English-Spanish Statistical Machine Translation: Experiments Domain Adaptation, Sentence Paraphrasing, Tokenization, and Recasting", *in proceedings of the third workshop on statistical machine translation*, pages 147-150, Ohio, 2008.

[9] Holger Schwenk, Jean-Babtiste Fouet and Jean Senellart, "First Steps towards a general-purpose French/English Statistical Machine Translation System", *In: proceedings of the third workshop on statistical machine translation*, pages 119-122, Ohio, 2008.

[10] Eleni Teshome. "Bidirectional English – Amharic Machine Translation: An Experiment using constrained corpus." MSc thesis, Addis Ababa University, Ethiopia, 2013.

[11] Jabesa D. and Yaregal A. “A Hybrid Approach to the Development of Bidirectional English-Oromiffa Machine Translation”, *In: Proceedings of the 9th International Conference on Natural Language Processing* (PolTAL2014), Springer Lecture Notes in Artificial Intelligence (LNAI), Vol. 8686, pp. 228-235, Warsaw, Poland, 2014.

[12] Michael G. “Toward a Rule-Based System for English-Amharic Translation.” In *LREC-2012: SALTMIL-AfLaT Workshop on Language technology for normalization of less-resourced Languages,* 2012.

[13] Mulu Gebreegziabher Teshome and Laurent Besacier. "Preliminary experiments on English-Amharic statistical machine translation." in *Proceedings of the 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU),* pp. 36-41, 2012.

[14] Akubazgi Gebremariam. “Amharic-to-Tigrigna Machine Translation Using Hybrid Approach.” MSc thesis, Addis Ababa University, Ethiopia, 2017.

[15] Dawit Mulugeta. “Ge’ez to Amharic Automatic Machine Translation: A Statistical Approach.” MSc thesis, Addis Ababa University, Ethiopia, 2015.

[16] Tadesse Kassa. “Morpheme-Based Bi-directional Ge’ez to Amharic Machine Translation.” MSc thesis, Addis Ababa University, Ethiopia, 2018.

[17] Rubin A. D. *A Brief Introduction to the Semitic Languages*. Piscataway, NJ, USA: Gorgias Press LLC, 2010.

[18] Dillmann, August and Carl Bezold. *Ethiopic grammar*. Amsterdam: Wipf and Stock, 2003.

[19] Coulmas F. *Writing Systems an Introduction to Their Linguistic Analysis*. Cambridge, United Kingdom: Cambridge University Press, 2003.

[20] Karan, Elke. "Writing System Development and Reform." MSc Thesis in partial fulfillment of the requirements for the degree of MA, Grand Forks, North Dakota, 2006.

[21] መምህር ደሴ ቀለብ። *ትንሣኤ ማእዝ*። አዲስ አበባ፣ኢትዮጵያ፡ ማኅበረ ቅዱሳን፣ 2008 እኢአ።

[22] አባ ኪዳነ ማርያም ጥዑመ ልሳን። *ፈቃደ መዝገበ ቅዳሴ*።አአ፣ ኢትዮጵያ፡ አኰቴት አሳታሚዎች፣ 2009።

[23] 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 ቅ/ሥላሴ መንፈሳዊ ኮሌጅ።

[24] *ኃይለ ኢየሱስ መንግሥት። የልሳነ ግእዝ መማርያ፣ሁለተኛ እትም።* በኢ/ኦ/ተ/ቤ/ክ በሰንበት ት/ቤቶች ማደራጃ መምሪያ ማህበረ ቅዱሳን ልማት ተቋማት አስተዳደር የአቡነ ጎርጎርዮስ ሥልጠና ማእከል፣ ፳፻፲፩ ዓ.ም።

[25] Andrea Decapua. *Grammar for Teachers: A Guide to American English for Native and Non-Native Speakers*. New York, USA: Springer Science Business Media LLC, 2008.

[26] ባየ ይማም። *የአማርኛ ሰዋስው፣የተሻሻለ ሁለተኛ እትም።* አዲስ አበባ ፣ ኢትዮጵያ፡ ካልቸር ኤንድ አርት ሶሳይቲ አፍ ኢትዮጵያ፣ 2000 አንደ ኢትዮጵያ አቆጣጠር።

[27] ደሴ በቀለ። *ትንሳኤ ግእዝ።* አዲስ አበባ፣ ኢትዮጵያ፡ በኢትዮጵያ ኦርቶዶክስ ተዋህዶ ቤተ ክርስቲያን በሰንበት ት/ቤቶች ማደራጃ መምሪያ ማህበረ ቅዱሳን፣ 2002። በአዲስ አበባ ዩንቨርስቲ የስነ ቋንቋ መምህር።

[28] Sisay F. “Part of speech tagging for Amharic using conditional random fields.” In *Proceedings of ACL-2005 Workshop on Computational Approaches to Semitic Languages*, 2005.

[29] Desta Berihu Weldegiorgis. "Design and Implementation of Automatic Morphological Analyzer for Ge’ez Verbs." A Master’s Thesis submitted to Addis Ababa University, Addis Ababa, 2010.

[30] SYSTRAN. “What-Is-machine-translation.” Internet: www.systransoft.com/systran/corporate-profile/translation technology/, last visited March 7, 2017.

[31] Hutchins J. “Example Based Machine Translation – a review and commentary.” In *Recent advances in example-based machine translation*, 2003.

[32] Hieu Hoang and Philip Koehn. “Design of the Moses Decoder for Statistical Machine Translation.” *in Proceedings of ACL Workshop on Software engineering, testing, and quality assurance for NLP*, 2008, pages 58-65.

[33] Antony P. “Machine Translation Approaches and Survey for Indian Languages.” Computational Linguistics and Chinese Language Processing, Vol.18, No.1, pp.47-78, 2013.

[34] Hal Daume III and Daniel M. “A noisy- channel model for document compression.” In *Proceedings of the 40th Annual Meeting on ACL*, 2002, pages 449–456.

[35] L. Bahl, Jelinek, and R. L. Mercer, “A maximum likelihood approach to continuous speech recognition.” *IEEE Trans. Pattern Anal. Machine Intel/*, vol. PAMI-5, pp.179-190, 1983.

[36] Bushra J. ‘’Statistical machine translation between languages with significant word order difference.’’ In Master Thesis, Univerzita Karlova v Praze & University of Malta, August 2010.

[37] Gao, J. “A Quirk Review of Translation Models.”, 2011.

[38] Peter F., Stephen A. Della P., Vincent J., and Robert L. "The Mathematics of Machine translation: Parameter Estimation." Computational Linguistics, vol.19, no. 2, pp. 263-311, 1993.

[39] Daniel J., James H. Martin, *Speech and Language Processing: an introduction to NLP, Computational Linguistics, and Speech Recognition*, Prentice-Hall, 2006.

[40] James Brunning. “Alignment models and algorithms for statistical machine translation.” Ph.D. Thesis, Cambridge University Engineering Department, 2010.

[41] Yamada, Kenji, and Kevin Knight. "A syntax-based statistical translation model." *In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 2011, pp. 523-530.

[42] Kevin Knight and Graehl Jonathan. "An overview of probabilistic tree transducers for natural language processing." In International Conference on Intelligent Text Processing and Computational Linguistics, Springer Berlin Heidelberg, 2005.

[43] Chiang D., Kevin K., and Wei W. "11,001 new features for statistical machine translation." *In Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, 2009, pp. 218-226.

[44] Michel G., Mark H., Kevin K., and Daniel M. “What is in a translation rule?” *In Proceedings of HLT-NAACL*, 2004, pages 273–280.

[45] Li, Zhifei, Chris C., Chris D., Juri G., Sanjeev K., Lane S., Wren N., Jonathan W., and Omar F. "Joshua: An open source toolkit for parsing-based machine translation." *In Proceedings of the Fourth Workshop on Statistical Machine Translation*, 2009, pp. 135-139.

[46] Alshawi, Hiyan, Srinivas, and Shona D. "Learning dependency translation models as collections of finite-state head transducers." *Computational Linguistics*, 26(1): 45-60, 2000.

[47] Melamed I.D. "Statistical machine translation by parsing," In *proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 2004, pp. 653-660.

[48] Michael Collins. “Statistical machine translation: IBM models 1 and 2.” *Columbia: Columbia University*, 2011.

[49] Robert L. Mercer & Paul S. Roossin. “A statistical approach to machine translation.” *Computational Linguistics*, 16(2), 79-85, 1990.

[50] Kalchbrenner N. and Blunsom P. "Recurrent continuous translation models," In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013, pages 1700–1709.

[51] D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate." Technical report, arXiv preprint arXiv: 1409.0473, 2014.

[52] James B. “Alignment models and algorithms for statistical machine translation”. Ph.D. Thesis, University of Cambridge Engineering Department, 2010.

[53] John S. White, Theresa A. O'Connell, and Lynn M. Carlson. "Evaluation of machine translation," In *Proceedings of the workshop on Human Language Technology*, 1993, pp. 206-210.

[54] Hanneman G., Edmund H., Abhaya A., Vamshi A., Alok P., Erik P. and Alon L. "Statistical transfer systems for French--English and German--English machine translation," In *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 163-166.

[55] Nawaphol L., Rapeeporn C., Richard B., Annupan R. “English Syntactic Reordering for English-Thai Phrase-Based Statistical Machine Translation.” *in 6th JCSSE,* 2009.

[56] Chao W., Michael C. and Philipp K., “Chinese Syntactic Reordering for Statistical Machine Translation.” In *proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, Pages 737–745.

[57] Mossa G., Yueting Z., J. Wu and Maan Y. A. “Arabic-Chinese and Chinese-Arabic Phrase based statistical machine translation systems”. *Inform. Technol. J*, vol. 9, pp. 666-72, 2010.

[58] Solomon T., Michael M., Martha Y. Million M., Solomon A., Wondwosen M., Yaregal A., Hafte A., Biniyam E., Tewodros A., Wondmagegn T., Amanuel L., Tsegaye A., Seifedin S., “Parallel

Corpora for bi-lingual English-Ethiopian Languages Statistical Machine Translation" In *proceedings of the 27th International Conference on Computational Linguistics*, 2018, Pages 3102–3111.

[59] Solomon T., Michael M., Martha Y. Million M., Solomon A., Wondwosen M., Yaregal A., Hafte A., Biniyam E., Tewodros A., Wondmagegn T., Amanuel L., Tsegaye A., Seifedin S., "Parallel Corpora for bi-lingual English-Ethiopian Languages Statistical Machine Translation" In *proceedings of the 27th International Conference on Computational Linguistics*, 2018, Pages 3102–3111.

[60] Björn Gambäck. "Tagging and Verifying an Amharic News Corpus." Language Technology for Normalization of Less-Resourced Languages (2012):79.

Annex I: Sample Parallel Corpus for Training

Ge'ez Sentences	Amharic Sentences
ብፁዕ ብእሲ ዘኢሖረ በምክረ ረሲዓን	ምስጉን ነው በክፉዎች ምክር ያልሄደ
ወዘኢቆመ ፍኖተ ኃጥኣን	በኃጢአተኞችም መንገድ ያልቆመ
ወዘኢነበረ መንበረ መስተሳልቃን	በዋዘኞችም ወንበር ያልተቀመጠ
ዘዳእሙ ሕገ እግዚአብሔር ሥምረቱ	ነገር ግን በእግዚአብሔር ሕግ ደስ ይለዋል
ወዘሕጎ ያነብብ መዕልተ ወሌሊተ	ሕጉንም በቀንና በሌሊት ያስባል
ወየከውን ከመ ዕፅ እንተ ትክልት ኀበ ሙሓዘ ማይ	እርሱም በውኃ ፈሳሾች ዳር እንደ ተተከለች
እንተ ትሁብ ፍሬሃ በበጊዜሃ	ፍሬዋን በየጊዜዋ እንደምትሰጥ
ወቈጽላኒ ኢይትነገፍ	ቅጠልዋም እንደማይረግፍ ዛፍ ይሆናል
ወኵሎ ዘገብረ ይፈጽም	የሚሠራውም ሁሉ ይከናወንለታል
አኮ ከመዝ ኃጥኣንሰ	ክፉዎች እንዲህ አይደሉም
አኮ መሬት ዘይግሕፍ ነፍስ እምገጸ ምድር	ነገር ግን ነፋስ ጠርጎ እንደሚወስደው ትቢያ ናቸው
ወበእንተዝ ኢይትነሥኡ ረሲዓን እምደይን ወኢኃጥኣን ውስተ ምክረ ጻድቃን	ስለዚህ ክፉዎች በፍርድ ኃጢአተኞችም በጻድቃን ማኅበር አይቆሙም
እስመ ያአምር እግዚአብሔር ፍኖቶሙ ለጻድቃን	እግዚአብሔር የጻድቃንን መንገድ ያውቃልና

ወፍኖቶሙ ለኃጥኣን ትጠፍእ	የክፉዎች መንገድ ግን ትጠፋለች
ለምንት አንገለጉ አሕዛብ	አሕዛብ ለምን ያጕረመርማሉ
ወሕዘብኒ ነበቡ ከንቶ	ወገኖችስ ለምን ከንቱን ይናገራሉ
ወተንሥኡ ነገሥተ ምድር	የምድር ነገሥታት ተነሡ
ወመላእክትኒ ተጋብኡ ላዕለ እግዚአብሔር ወላዕለ መሲሑ	አለቆችም በእግዚአብሔርና በመሢሑ ላይ እንዲህ ሲሉ ተማከሩ
ንበትክ መኣሥሪሆሙ	ማሰርያቸውን እንበጥስ
ወንገድፍ እምላዕሌነ አርዑቶሙ	ገመዳቸውንም ከእኛ እንጣል
ዘይነብር ውስተ ሰማይ ይሥሕቆሙ	በሰማይ የሚኖር እርሱ ይሥቃል
ወእግዚአብሔር ይሳለቅ	ጌታም ይሣለቅባቸዋል
ሶበ ይነቦሙ በመዐቱ	በዚያን ጊዜ በቍጣው ይናገራቸዋል
ወበመዐቱ የሀውኮሙ	በመዓቱም ያውካቸዋል
ወአንሰ ተሠየምኩ ንጉሥ በጽዮን በደብረ መቅደሱ	እኔ ግን ንጉሤን ሾምሁ በተቀደሰው ተራራዬ በጽዮን ላይ
ከመ እንግር ትእዛዘ ለእግዚአብሔር	ትእዛዙን እናገራለሁ
እግዚአብሔር ይቤለኒ	እግዚአብሔር አለኝ
ወልዱየ እንተ	አንተ ልጄ ነህ
ወአነ ዮም ወለድኩከ	እኔ ዛሬ ወለድሁህ
ሰአል እምነየ	ለምነኝ
ወእሁብከ አሕዛበ ለርስትከ ወምኵናኒከኒ እስከ አጽናፈ ምድር	አሕዛብን ለርስትህ የምድርንም ዳርቻ ለግዛትህ እሰጥሃለሁ
ወትሬዕዮሙ በበትረ ኀጺን	በብረት በትር ትጠብቃቸዋለህ
ወከመ ንዋየ ለብሓ ትቀጠቅጦሙ	እንደ ሸክላ ሠሪ ዕቃዎች ትቀጠቅጣቸዋለህ
ወይእዜኒ ነገሥት ለብዉ	አሁንም እናንት ነገሥታት ልብ አድርጉ
ወተገሠጹ ኵልክሙ እለ ትኬንንዋ ለምድር	እናንት የምድር ፈራጆችም ተገሠጹ
ተቀነዩ ለእግዚአብሔር በፍርሀት	ለእግዚአብሔር በፍርሃት ተገዙ
ወተሐሠዩ በረዐድ	በረዓድም ደስ ይበላችሁ
አጽንዕዋ ለጥበብ ከመ ኢይትመዓዕ እግዚአብሔር ወኢትትሐጐሉ እምፍኖተ ጽድቅ	ጌታ እንዳይቈጣ እናንተም በመንገድ እንዳትጠፉ ተግሣጹን ተቀበሉ
ሶበ ነደት ፍጡነ መዐቱ	ቍጣው ፈጥና ትነድዳለችና

ብፁዓን ኵሎሙ እለ ተወከሉ ቦቱ	በእርሱ የታመኑ ሁሉ የተመሰገኑ ናቸው
አግዚአ ሚበዝኑ እለ ይሣቅዩኒ	አቤቱ የሚያስጨንቁኝ ምንኛ በዙ
ብዙኃን ቆሙ ላዕሌየ	በኔ ላይ የሚቆሙት ብዙ ናቸው
ብዙኃን ይቤልዋ ለነፍስየ ኢያድኅነኪ አምላክኪ	ብዙ ሰዎች ነፍሴን አምላክሽ አያድንሽም አልዋት
አንተሰ እግዚአ ምስካይየ	አንተ ግን አቤቱ መጠጊያዬ ነህ
አንተ ክብርየ ወመልዕለ ርእስየ	ክብሬንና ራሴንም ከፍ ከፍ የምታደርገው አንተ ነህ
ቃልየ ኀበ እግዚአብሔር ጸራኅኩ	በቃሌ ወደ እግዚአብሔር እጮሃለሁ
ወሰምዐኒ እምደብረ መቅደሱ	ከተቀደሰ ተራራውም ይሰማኛል
አንሰ ሰከብኩ ወኖምኩ	እኔ ተኛሁ አንቀላፋሁም
ወተንሣእኩ እስመ እግዚአብሔር አንሥአኒ	እግዚአብሔርም ደግፎኛልና ነቃሁ
ኢይፈርህ እምአእላፍ አሕዛብ እለ ዐገቱኒ	ከሚከብቡኝ ከአእላፍ ሕዝብ አልፈራም
ተንሥእ	ተነሥ
እግዚአ አምላኪየ ወአድኅነኒ	አቤቱ አምላኬ ሆይ አድነኝ
እስመ አንተ ቀሠፍኮሙ ለኵሎሙ እለ ይፃረሩኒ	አንተ የጠላቶቼን መንጋጋ መትተሃልና
ስነኒሆሙ ለኃጥኣን ሰበርከ	የክፉዎችንም ጥርስ ሰብረሃልና
ዘእግዚአብሔር አድኅኖ	ማዳን የእግዚአብሔር ነው
ወላዕለ ሕዝብከ በረከትከ	በረከትህም በሕዝብህ ላይ ነው
ሶበ ጸዋዕኩዎ ለእግዚአብሔር ሰምዐኒ ጽድቅየ	የጽድቄ አምላክ በጠራሁት ጊዜ መለሰልኝ
ወእምንዳቤየ አርሐበ	በጭንቀቴም አሰፋህልኝ
ተሥህለኒ ወሰምዐኒ ጸሎትየ	ማረኝ ጸሎቴንም ስማ
ደቂቀ እጓለእመሕያው	የሰው ልጆች
እስከ ማእዜኑ ታከብዱ ልብከሙ	እስከ መቼ ድረስ ልባችሁን ታከብዳላችሁ
ላእምሩ ከመ ተሰብሐ እግዚአብሔር በጻድቁ	እግዚአብሔር በጻድቁ እንደ ተገለጠ እወቁ
እግዚአብሔር ይሰምዐኒ ሶበ ጸራኅኩ ኀቤሁ	እግዚአብሔር ወደ እርሱ በተጣራሁ ጊዜ ይሰማኛል
ተምዑ	ተቈጡ
ወኢተአብሱ	ኃጢአትን አታድርጉ

ወዘትሔልዩ በልብከሙ ውስተ መስካቢከሙ	በመኻታችሁ ሳላችሁ በልባችሁ አስቡ
ሥዑ መሥዋዕተ ጽድቅ	የጽድቅን መሥዋዕት ሠዉ
ወተወከሉ ለእግዚአብሔር	በእግዚአብሔርም ታመኑ
መኑ ያርእየነ ሠናይቶ	በጎውን ማን ያሳየናል
ብዙኃን እለ ይቤሉ	የሚሉ ብዙ ናቸው
ተዐውቀ በላዕሌነ ብርሃነ ገጽከ እግዚኦ	አቤቱ የፊትህ ብርሃን በላያችን ታወቀ
ወወደይከ ትፍሥሕተ ውስተ ልብነ	በልቤ ደስታን ጨመርህ
እምፍሬ ስርናይ ወወይን ወቅብእ በዝኃ	ከስንዴ ፍሬና ከወይን ከዘይትም ይልቅ በዛ
በሰላም እሰክብ ወእነውም	በሰላም እተኛለሁ አንቀላፋለሁም
እስመ እንተ እግዚኦ በተስፋ ባሕቲትከ ላኅደርከኒ	አቤቱ አንተ ብቻህን በእምነት አሳድረኸኛልና
ቃልየ አፅምእ እግዚኦ	አቤቱ ቃሌን አድምጥ
ወለቡ ጽራኅየ	ጩኸቴንም አስተውል
ወአፅምአኒ ቃለ ስእለትየ	የልመናዬን ቃል አድምጥ
ንጉሥየኒ ወአምላክየኒ እስመ ኀቤከ እጼሊ	ንጉሤና አምላኬ ሆይ አቤቱ ወደ አንተ እጸልያለሁና
እግዚኦ በጽባሕ ስምዐኒ ቃልየ	በማለዳ ድምፄን ትሰማለህ
በጽባሕ እቀውም ቅድሜከ	በማለዳ በፊትህ እቆማለሁ
ወኣስተርኢ	እጠብቃለሁም
እስመ ኢኮንከ አምላከ ዘዐመፃ	አንተ በደልን የማትወድድ አምላክ ነህና
ወኢየኀድሩ እኩያን ምስሌከ	ክፉ ከአንተ ጋር አያድርም
ወኢይነብሩ ዐማፅያን ቅድመ አዕይንቲከ	በከንቱ የሚመኩ በዓይኖችህ ፊት አይኖሩም
ጸላእከ ኵሎሙ ገበርተ ዐመፃ	ክፉ አድራጊዎችን ሁሉ ጠላህ
ወትገድፎሙ ይነቡ ሐሰተ	ሐሰትን የሚናገሩትን ታጠፋቸዋለህ
ብእሴ ደም ወጉሕላዌ ይስቆርር እግዚአብሔር	ደም አፍሳሹንና ሸንጋዩን ሰው እግዚአብሔር ይጸየፋል
ወአንሰ በብዝኀ ምሕረትከ እበውእ ቤተከ	እኔ ግን በምሕረትህ ብዛት ወደ ቤትህ እገባለሁ
ወእሰግድ ውስተ ጽርሐ መቅደስከ በፍሪሆትከ	አንተን በመፍራት ወደ ቅድስናህ መቅደስ እሰግዳለሁ
እግዚኦ ምርሐኒ በጽድቅከ ወበእንተ ጸላእትየ	አቤቱ ስለ ጠላቶቼ በጽድቅህ ምራኝ

አርትዕ ፍኖትየ ቅድሜከ	መንገዴን በፊትህ አቅና
እስመ አልቦ ጽድቀ ውስተ አፉሆሙ	በአፋቸው እውነት የለምና
ወልቦሙኒ ከንቱ	ልባቸውም ከንቱ ነው
መቃብር ክሡት ጐራዲቶሙ	ጉሮሮአቸው የተከፈተ መቃብር ነው
ወጸልሕዉ በልሳናቲሆሙ	በምላሳቸው ይሸነግላሉ
ኰንኖሙ እግዚኦ	አቤቱ ፍረድባቸው

Annex II: Sample Parallel Corpus for Testing

Ge’ez Sentences	Amharic Sentences
ወጸልኡኒ ህየንተ ዘአፍቀርክዎሙ	በወደድኋቸውም ፋንታ ጠላትነትን
ሚም ላዕሌሁ ኃጥአ	በላዩ ኃጢአተኛን ሹም
ወሰይጣን ይቁም በየማኑ	ሰይጣንም በቀኙ ይቁም
ወሶበሂ ይትዋቀሥ ይፃእ ተመዊአ	በተምዋገተም ጊዜ ተረትቶ ይውጣ
ወጸሎቱሂ ትኩኖ ጌጋየ	ጸሎቱም ኃጢአት ትሁንበት
ወይኩና መዋዕሊሁ ጎዳጠ	ዘመኖቹም ጥቂት ይሁኑ
ወሢመቶሂ ይንሣእ ባዕድ	ሹመቱንም ሌላ ይውሰድ
ወይኩኑ ደቂቁ እጓለ ማውታ	ልጆቹም ድህ አደግ ይሁኑ
ወብእሲቱሂ ትኩን መበለተ	ሚስቱም መበለት ትሁን
ወይትሀውኩ ደቂቁ	ልጆቹም ይናወጡ
ወይፍልሱ	ይቅበዝበዙ
ወያስተፍእሙ	ይለምኑም
ወይስድድዎሙ እምአብያቲሆሙ	ከስፍራቸውም ይባረሩ
ወይበርበሮ ባዕለ ዕዳ ኵሎ ንዋዮ	ባለዕዳም ያለውን ሁሉ ይበርብረው
ወይሐብልዩ ነቢር ኵሎ ተግባሮ	እንግዶችም ድካሙን ሁሉ ይበዝብዙት
ወኢይርከብ ዘይረድኦ	የሚያግዘውንም አያግኝ
ወኢይምሐርዎሙ ለእጓለ ማውታሁ	ለድህ አደግ ልጆቹም የሚራራ አይኑር

ወይሠረዉ ደቂቁ	ልጆቹ ይጥፉ
በአሕቲ ትውልድ ትደምሰስ ስሙ	በአንድ ትውልድ ስሙ ይደምሰስ
ወትዘከር ኀጢአተ አቡሁ በቅድመ እግዚአብሔር	የአባቶቹ ኃጢአት በእግዚአብሔር ፊት ትታሰብ
ወኢይደምሰስ ጌጋያ ለእሙ	የእናቱም ኃጢአት አትደምሰስ
ወየሀሉ ቅድመ እግዚአብሔር በኵሉ ጊዜ	በእግዚአብሔር ፊት ሁልጊዜ ይኑሩ
ወይጥፋእ እምድር ዝክሩ	መታሰቢያቸው ከምድር ይጥፋ
እስመ ኢተዘከረ ይግበር ምጽዋተ	ምሕረትን ያደርግ ዘንድ አላሰበምና
ወሰደደ ብእሴ ነዳየ ወምስኪነ ወጥቡዕ ልቡ ለቀቲል	ችግረኛንና ምስኪንን ልቡ የተሰበረውንም ሰው ይገድል ዘንድ አሳደደ
ወአብደረ መርገመ	መርገምን ወደደ
ወትመጽአ	ወደ እርሱም መጣች
ወአበያ ለበረከት	በረከትንም አልመረጠም
ወትርሕቅ አምኔሁ	ከእርሱም ራቀች
ወለብሳ ለመርገም ከመ ልብስ	መርገምን እንደ ልብስ ለበሳት
ወቦአት ከመ ማይ ውስተ አማዑቱ	እንደ ውኃም ወደ አንጀቱ ገባች
ወከመ ቅብእ ውስተ አዕጽምቲሁ	እንደ ቅባትም ወደ አጥንቱ
ወትኩኖ ከመ ልብስ ዘይትዐጸፍ	እንደሚለብሰው ልብስ ይሁነው
ወከመ ቅናት ዘይቀንት ዘልፈ	ሁልጊዜም እንደሚታጠቀው ትጥቅ

Annex III: Sample Language Model for Amharic Language

\data

ngram 1=3850

ngram 2=7396

ngram 3=6998

\1-grams:

-0.7607266 </s>

-99 <s> -0.1418663

-0.3181287 <unk>

-4.5114 ሁለት -0.04584911

-1.881648 ሁሉ 0.01564619

-4.5114 ሁሉም -0.04564605

-3.816312 ሁሉን -0.04599274

-4.5114 ሁላቸው -0.04296102

-4.5114 ሁላችሁ -0.04600612

-4.5114 ሁላችሁም -0.04532008

-3.081609 ሁልጊዜ -0.04193328

-4.5114 ሁልጊዜም -0.04592644

-4.5114 ሁነኝ -0.0460195

-4.5114 ሄዱ 0.0367188

-4.5114 ሄጃለሁ 0.0367188

-4.5114 ሄጃለሁና -0.04598924

-3.373595 ሆነ -0.1547874

-3.540826 ሆነህ -0.302618

-4.5114 ሆነልኝ 0.0367188

-4.5114 ሆነች -0.0459456

-4.5114 ሆነችልኝ 0.0367188

-4.5114 ሆነችኝ 0.0367188

-4.5114 ሆነኝ 0.0367188

-4.5114 ሆነው -0.04584911

-4.5114 ሆኑ -0.04598924

-4.5114 ሆኑባቸው 0.0367188

-4.5114 ሆናቸው 0.0367188

-3.816312 ሆንሁ -0.09494096

-4.5114 ሆንን -0.04584911

-3.373595 ሆኖ -0.04573501

-4.5114 ሆኖለሁና 0.0367188

-3.081609 ሆይ -0.07060698

-4.5114 ሆዳቸውን -0.0460195

-4.5114 ሆዴን -0.0460195

-4.5114 ሆድ -0.0460195

-4.5114 ለሌሊት -0.04598924

-4.5114 ለልዑል -0.04525367

-3.081609 ለልጅ -0.6034188

-4.5114 ለልጆቹ -0.0460195

-4.5114 ለልጆችህ -0.0460195

-4.5114 ለሕዝቡ -0.04580545

-4.5114 ለሕይወቴ -0.04584911

-4.5114 ለሕፃናቶቻቸው -0.0460195

-4.5114 ለመማለድ -0.0460195

-4.5114 ለመሬትዋም -0.0460195

-4.5114 ለመርዳት -0.0460195

-4.5114 ለመሮጥ -0.04426009

-4.5114 ለመስጠት -0.04590785

-4.5114 ለመሸፈኛ -0.0460195

-4.5114 ለመነህ -0.0460195

-4.5114 ለመኑ 0.0367188

-4.5114 ለመንሁ 0.0367188

-4.5114 ለመንሁት -0.0460195

-4.5114 ለመንገዴ -0.04590785

-3.997929 ለመንጠቅ -0.04597586

-4.5114 ለመጠበቅ -0.0460195

-4.5114 ለመጽናናቴ -0.0460195

-4.5114 ለመፍረድ -0.04598924

-4.5114 ለሙሴ -0.04598924

-4.5114 ለሚመጣ -0.04561406

-4.5114 ለሚሰድቡኝ -0.0460195

-4.5114 ለሚታመኑት -0.04029167

-4.5114 ለሚታገሡት -0.0457316

-4.5114 ለሚኖር -0.0454377

-4.5114 ለሚወለደው -0.04567283

-3.997929 ለሚወድዱ -0.04560067

-4.5114 ለሚያስጨንቁኝ -0.0460195

-3.997929 ለሚጠሩት -0.2499411

-4.5114 ለሚጠብቁ 0.0367188

-3.997929 ለሚፈሩት-0.04596248

-4.5114 ለሚፋጠኑት -0.0460195

-4.5114 ለማመስገን -0.0460195