ST. MARY'S UNIVERSITY

SCHOOL OF GRADUATE STUDY

Morpheme Based Bi-Directional Machine Translation

The Case of Ge'ez to Tigrigna

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

By: Helen Akelew Nega
Addis Ababa, Ethiopia

Advisor: Hafte Abera
Addis Ababa, Ethiopia

January 2023

Dedicated to:

My Mother Tsige G/kidan, I always remembers all the sufferings, pain and dedications that you lived to educate me. I offer you this thesis work as a memorial work for invaluable experiences and courage you built in me.

ACCEPTANCE

Morpheme Based Bi-Directional Machine Translation

The Case of Ge'ez to Tigrigna

By

Helen Akelew Nega

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

Thesis Examination Committee:

_____

Internal Examiner

_____

External Examiner

_____

Dean, Faculty of Informatics

January 13, 2023, 2022

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.

Helen Akelew Nega

_____ Signature

Addis Ababa,
Ethiopia

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

Hafte Abera (Mr.)

_____ Signature

Addis Ababa Ethiopia

January 13, 2023

Acknowledgments

This research work would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their invaluable assistance in time of need. First and foremost, I would like to express my gratitude to my advisor Mr. Hafte Abera for offering me the invaluable advice, guidance, monitoring and support during each step of my work. Second, I would to thank St. Mary's University School of Graduates Program staff members who contributed their genuine roles to accomplish my tasks. Besides, my thanks goes to Memehire Diakon Eng. Bisrat Halefom, and Gezai Aberha, for their assistance of help to understand the grammar of Geez and Tigrigna. My reserved acknowledge goes to all my colleagues for the entire moral supports they rendered me. In this regard, special consideration would be forwarded for Hawi Tamiru and Senaiyet Shermolo. There are also individuals that I am unable to list but have directly or indirectly help me in completion of my thesis that I want to express my heart-felt thanks. In addition, my appreciation also goes to Mr. Tamiru Geleta whose advice and support had been with me from start to this end. My deepest gratitude goes to my family for their unfailing support, patience and prayers, to reach this end. My dearest brother, Tewodros Akelew, you are not only my brother you are also the substitute of my Mother! Without you I couldn't finish this Study thank you very much. And no words of thanks could be found to express my feelings about your multifaceted supports and encouragement.

Last but not least, and the foremost thanks goes to the one above all, the Almighty God, for answering my prayers and for giving me the strength to be able to accomplish this study "መድኃኒዓለም የለም የሚሳነው".

Table of Contents

LIST OF ACRONYMS

ALPAC - Automatic Language Processing Advisory Committee

BLEU - Bilingual Evaluation Understudy

CBMT – Corpus Based Machine Translation

DMT – Direct Machine Translation

EM – Expectation Maximization

EBMT - Example-based machine translation

GPU – Graphic Process Unit

IBM - International Business Machines

IRSTLM –Institute of Research LM Language Model MT – Machine Translation

MAP - Maximum A Posteriori

METEOR – Metric for Evaluation if Translation with Explicit Ordering

ML – Maximum Likelihood

MT – Machine Translation

NIST- National Institute of Standards and Technology

NMT – Neural Mechanical Translation

OOV – Out of Vocabulary

OVS – Object Verb Subject

RBMT - Rule-Based Machine Translation

RNN – Recurrent Neural Network

SL – Source Language

SOV – Subject Object Verb

SMT – Statistical Machine Translation

TER – Translation Error Rate

TL – Target Language

VSO – Verb Subject Object

LIST OF FIGURES

LIST OF TABLES

Tables ***Page***

Abstract

Both Ge'ez and Tigrigna languages, which the native Ethiopian languages, are morphological rich and complex for bi-directional machine translation. To overcome this machine translation problem, this study explored the effect of morpheme-based translation unit for bidirectional Ge'ez and Tigrigna languages. The corpus was taken from Ten Bible Books that contained 384 that contained 9189 verses. The corpus was used both for developing pre-trained model and for validation. Accordingly, to train the morfessor, 12173 simple Ge'ez and 16708 Tigrigna words were taken from SQlite database. Explicitly, from the total of 7290 verses data, 80%, that is 7290 Verses were used to develop the pre-trained model and 20% which is 1899 Verses were used for testing or validation purposes. we used Mosses for translation process, MGIZA++ for alignment of word and morpheme, morfessor and IRSTLM techniques for the language modeling. After preparing and designing the prototype and the corpus, different experiments were conducted. BLUE score which is standard for automatic machine translation evaluation was used to measure how much of the system output is correct. Experimental results showed a better performance of 9.23% and 8.67% BLEU scores using morpheme-based from Geez to Tigrigna and from Tigrigna to Geez translation, respectively. That is, it was found out that the model or the system output was correct. Regarding the BLUE metrics evaluation tool, it was also found to show proper validation scores or results. As to the alignment challenges, many-to-many alignment is the major challenge. Hence, there is a need to conduct further research to handle the issue of many-to-many alignment challenge.

Keywords: Bi-directional Machine Translation, Bilingual Evaluation Understudy, Ge'ez Language, Tigrigna Language

CHAPTER ONE

1. INTRODUCTION

1.1. Background

Machine translation (MT) is defined by Daniel and James [1] "as a technology that enables the use of computers to automate the process of translating from one language to another." They further elaborated that translation is a difficult, fascinating, and intensely human endeavor which is as rich as any other area of human creativity. Machine translation has undergone nearly more than half a century period of development to reach its current status. When one refers to the history of machine translation, as is described by Jonathan Slocum [2], is traced from early systems of the 1950s and 1960s which is the impact of the Automatic Language Processing Advisory Committee (ALPAC) report in the mid-1960s, the revival in the 1970s, commercial and operational systems of the 1980s, and research during the 1980s.

Machine translation has various advantages: One of the advantages is that translation by machine takes a fraction of the time which is very short when compared to human translation that takes much longer time. As is underlined by Caitlin, M. [3], the rate of machine translation is highly rapid than that of human translation. He further explained that the average human translator can translate around 2,000 words a day. Multiple translators could be assigned to increase the project output, but it takes much longer time in comparison to the translation via machine. Machines can generate thousands of words each minute [3]. The second advantage is lower cost which makes MT comparatively cheaper. In many investments the use of MT is cost effective and beneficiary. In case, when expertise professional translator is used the charge as per page will be extremely costly. Thirdly, Machine Translation improves consistency which is based on MT engines that could be customized with your preferred business terms. That is, computer program can be trained to use the same term for the same concept every time [4]. The other advantage that makes machine translation favorable is its confidentiality. Giving sensitive data to a translator might be risky while with machine translation information is protected.

MT approaches are rule based, corpus based and hybrid [1]. Rule-Based Machine Translation (RBMT), also known as Knowledge-Based MT, is a general term that describes machine translation systems based on linguistic information about source and target languages. Corpus

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

Machine translation has its own challenges even if it is active current research area [1]. Several well-known problems are, fundamentally, problems of scarce bitext. The first challenge in MT is translation of low-resource language pairs. The most straightforward example of scarce bitext covers most of the world's language pairs. The second one is translation across domains. Translation systems are not robust across different types of data, performing poorly on text whose underlying properties differ from those of the system's training data. The third challenge is translation into morphologically rich languages. Much of human communication is oral. Even ignoring speech recognition errors, the substance and quality of oral communication differs greatly from that found in most bitext [5].

Ethiopian is one of the countries in Africa that has its own alphabets called "Fidel' and Numbers. This scripting method is the identity of the country not only in African but also in the international

Arena. The word Ge'ez means first in the Alphabet, first in reading style and first in Zema (Gloss) teaching of the Ethiopian orthodox Tewahedo Church. Ge'ez (ግዕዝ) is an ancient South Semitic language and is a member of the Ethiopian Semitic group. The language originated in southern regions of Eritrea and the northern region of Ethiopia in the Horn of Africa. It later became the official language of the Kingdom of Aksum and Ethiopian imperial court. Today, Ge'ez remains only as the main language used in the liturgy of the Ethiopian Orthodox Tewahedo Church, the Eritrean Orthodox Tewahedo Church, the Ethiopian Catholic Church, the Eritrean Catholic Church, and the Beta Israel Jewish community [6].

Ge’ez is fairly massive in size, with its 182 alphabets. Though in order to make a fair comparison it must be said that there are essentially 26 main alphabets, all consonants, in Ge’ez; while the rest are essentially those with additional strokes and modifications added on to the main forms to indicate a vowel sound associated with it or to make aural adjustments in the basic consonant sound. The common writing surface of ancient Ge’ez is “birana” a parchment made from animal skin, because of its organic nature it is subject to de gradation over long periods of time [7].

In Ethiopia, Tigrinya is the third most spoken language and the "Tigray" are the third largest ethnic group, after the Oromo and Amhara. In Eritrea, Tigrigna is by far the most spoken language, and they represent 55% of the population. Tigrinian, Tigrinyan, is a Semitic language spoken in the Tigray Region of Ethiopia (its speakers there are called "Tigraway") by the Tigrinya people, where it has official status, and in central Eritrea, where it is one of the two main languages of Eritrea, and, and among groups of emigrants from these regions, including some of the Beta Israel now living in Israel. There is no generally agreed upon name for the people who speak Tigrinya. A native of Tigray is referred to in Tigrinya as Tigraway (male), tigaweyt" (female), tigrawot or tegaru (plural). In Eritrea, Tigrinya speakers are officially known as the Bihér-Tigrigna which means nation of Tigrinya speakers. Bihér roughly means nation in the ethnic sense of the word in Tigrinya Muslim native Tigrigna speakers are known as the Jeberti, an Arabic name which implies conversion to Islam among Horn Africans. Tigrinya is the third most spoken language in Ethiopia, after Amharic and Oromo, and by far the most spoken in Eritrea. It is also spoken by large immigrant communities around the world, in countries including Sudan, Saudi Arabia, Germany, Italy, Sweden, the United Kingdom, Canada and the United States [8].

Tigrinya is a Semitic language spoken in Eritrea and in the Tigray Region of Northern Ethiopia. Tigrigna used the whole Geez alphabets and eight additional (Que-ቖ፣ She-ሸ፣ Che-ቸ፣ Gne-ኘ፣ Zye-ዠ፣ Je-ጀ፣ ጨ and Ve-ቨ) Fidels which is 34*7 (238) size syllables and 4*5 (20) labialized. Each of the columns are labeled as ግእዝ /ge‘‘ez/ (first order), ካእብ /ka’b/ (second order), ሳልስ /salɨs/ (third order), ራብእ /rabɨ’/ (fourth order), ሓምስ /hamɨs/ (fifth order), ሳዴስ /sadɨs/ (sixth order), and ሳብእ /sabɨ’/ (seventh order) of alphabets. The orders represent the tones of each of the vowels. This shows the combination of consonants and vowels [7]. Like English, Tigrinya is written from left to right.

1.2. Statement of the Problem

Geez and Tigrigna are closely related languages. The reason is, because both languages have similar alphabets, similar sentence structure and writing system, have similar phrasal categories, use similar punctuation. However, there are several differences between the two languages. For instance, there are several Tigrigna phrases that differ in their order of words in Geez phrases. So to solve this problem, syntactical reordering rules are proposed to change the order of words in a given Tigrigna phrase in a sentence to have more similar structural order of words as the target language which can be considered as a pre-processing step to statistical approach. There are times when human translations are used. However, they tend to be slower as compared to machine translations. Sometimes it can be hard to get a precise translation that reveals what the text is about without everything being translated word-to-word. In addition, it can be more important to get the result without delay which is hard to accomplish with human translators. That is, when machine translation comes in, that solves most of the problems caused by a human translator. As far as the researcher knowledge is concerned, there is no prior study conducted on system development of Geez-to-Tigrigna machine translation system. Nevertheless, there are researches done on machine translations regarding some of the languages spoken in Ethiopia at the national or local levels. For instance some of the researches done are *Ge'ez-Amharic automatic machine translation, Bidirectional Ge'ez-Amharic neural machine translation, Morpheme based bidirectional Ge'ez-Amharic machine translation, Bidirectional Tigrigna–English machine translation, Bidirectional Amharic-Afaan Oromo machine translation using statistical approach, , etc.* To this end, this study

"Morpheme based bidirectional Ge'ez-Tigrigna machine translation" strives to answer the following research questions.

1.3. Research questions

To realize the intents of this research the following basic research questions which need to be addressed are listed as hereunder.

- What optimal language translation model is required to facilitate effective machine translation process Geez and Tigrigna languages?
- What evaluation mechanisms need to be used?

- What syntactic relationships do exist between Geez and Tigrigna languages?
- What are the challenges observed between the morphology and Syntax of Geez and Tigrigna languages?

1.4. Objectives of the Study

The following general and specific objectives are designed as follows.

1.4.1. General Objective

The general objective of this research is to design morpheme-based Geez-Tigrigna Machine Translation model and implement the translation.

1.4.2. Specific Objectives

The general objective would be realized by the following specific objectives. The Specific objectives are:

- To design an optimal language and translation model.
- To evaluate the performance of the prototype
- To identify the syntactic relationship between Geez and Tigrigna languages.
- To identify the syntax and morpheme gaps between Ge'ez and Tigrigna languages

1.5. Scope of the Study

Bi-directional Ge'ez to Tigrigna machine translation is designed to translate a sentence written in Ge'ez into Tigrigna and vice versa. In this research, speech to speech translation, text to speech translation and speech to text translation are not included. Machine translation has different approaches such as, example-based approach, rule-based approach, statistical approach and hybrid approach. To conduct the research, the statistical MT approach which involves preparing parallel corpus for both target and source language was used. Aligning the prepared parallel corpus and training the system in both direction and the finally performing a bi-directional machine translation from source to target language and from target to source language were implemented.

1.6. Significance of the Study

As described above machine translation is a design to translate text from one language (source language) to another language (target language) without the help of human and the translation express the same meaning as it is in source language. Ge'ez is a Semitic language of the southern

peripheral group, to which also belong the south Arabic dialects and Amharic, one of the principal languages of Ethiopia. Tigrinya has its own alphabet **of 32 letters adopted from Ge'ez,** a language which exists with a very limited function within the Coptic Orthodox and Catholic Churches. Like English, Tigrinya is written from left to right. Although the Tigrinya script might look difficult, pronunciation is simple and straightforward, as the phonetic symbols closely resemble pronunciation. The closest living languages to Ge'ez are Tigre and Tigrigna with lexical similarity at 71% and 68% respectively [9].

As a result there is a need for a rule that can translate Ge'ez and Tigrigna morpheme based texts that have more than one meaning due to their part of speech. Collecting the corpus was difficult since there was not prearranged data for the bi-directional Ge'ez -Tigrigna corpus. In the context of such gaps, it is of paramount importance to undertake the study of morpheme based bidirectional machine translation of Geez-Tigrigna bulky contents. Besides getting meaningful translation to the bulky contents it has also has an advantage of reducing delayed time span and manual labor invested on the translation system.

1.7. Research Methodology

Research methodology is a systematic way of solving research questions scientifically by following various steps along with the logic behind them [10]. It is the general principle by which a researcher is guided [11]. Accordingly, the methodology of this research includes the research design and methods that are presented as follows. According to Janet [12], "The arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure" is called a research design. It is a conceptual structure that includes the collection, measurement and analysis of the corpus. Based on this general notion, this study used corpus preparation, analysis, tools and techniques as well as the evaluation mechanisms. Moreover, this study, a bidirectional Ge'ez-Tigrigna machine translation, also used corpus based (statistical) machine translation approach. Each element of the design is discussed as follows.

1.7.1. Corpus preparation

The process to develop translation model for morpheme based bidirectional Ge'ez-Tigrigna machine translation has followed the following procedure. From the total collected verses, 80% and the rest 20% used for validation or evaluation purposes.

1.7.2. Tools and Experiments

Machine translation has different approaches such as, example-based approach, and rule-based approach, statistical approach and hybrid approach. Statistical approach is economically wise i.e., doesn't need linguist professionals, the translation process is done by only from parallel corpus and also recommended by different researchers [6, 13, 14]. The basic tools used for accomplishing the machine translation task is Moses for Mere Mortal; free available open-source software which is used for statistical machine translation and integrates different toolkits which used for translation purpose such as IRSTLM for language model, Decoder for translation, MGIZA++ for word and morpheme alignment.

1.8. Thesis Organization

The following procedures are the logical organization of the thesis work. The first chapter, which is the introductory part, deals with the statement of the problem to be addressed and the objectives to be attained. Besides, the scope as well as the significance or the study and the methodology to be followed are also included. Chapter two incorporates the reviewed related literatures that include the relevant theoretical and the technical issues as well as the results of prior works done on morpheme based bidirectional machine translation study areas. Chapter three presents the overview of Ge'ez language and its relationship with Tigrigna language and discussion of alignment challenge between the two languages. Next comes chapter four that deals with designing processes of the prototype including corpus preparation, types of corpus used for the study, corpus alignment, and discussions about the prototype of the system. Chapter five deals with experiment of the study, which includes different experimentations and their outcomes, that is followed by interpretations and results. The last chapter, that is chapter six, incorporates the findings, conclusions and the way forward.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Machine translation

Machine translation (MT) is defined by Amine [16] as a translation of information from one natural language (source language) to another language (target language) using computerized systems; automatic or semi-automatic. It is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. As Clark, et.al [17] described MT was conceived as one of the first applications of the newly invented electronic computers back in 1940's. MT is an applied research that draws ideas and techniques from linguistics, computer science, artificial intelligence, translation theory and statistics.

One of the major importance of MT is that it reduces the language difficulties in information access and promotes multi-lingual real-time communications. According to Tadesse [6], translation is not just only word-to-word substitution, rather the translator has to interpret and analyze all elements of a text. It also needs to know how each word may influence another and this requires extensive expertise in grammar, syntax (sentence structure/word order), semantics, etc., in the source and target languages, as well as familiarity with each local region in which syntax and semantic means of sentence structure and meanings respectively.

According to Jabessa, et al and Daniel, et al [18, 19], machine translation systems can function as bilingual systems or multilingual systems based on the number of languages used in the translation process. They further explained that bilingual systems are designed specifically for two languages (single pair of languages) and multilingual systems are designed for more than two languages. It should also be noted that the translation can be unidirectional or bidirectional [16]. The system translates from the source language into the target language only in one direction, in the case of unidirectional [20]. Bidirectional systems function in both directions in such a way that one language can stand either as source language or a target language [19]. Bilingual systems can be unidirectional or they can be bidirectional, but multilingual systems are usually designed to be bidirectional.

2.2. Machine Translation Approaches

The first task of MT is to analyze the source language input and to create an internal representation. Such a representation is operated and transferred to a form that is suitable for the target language. Then at last output is generated in the target language [9]. It is further elaborated that MT systems can be classified according to their core methodology. The rule-based approach and the corpus-based approach are the two main paradigms that are found under this classification.

In the rule-based approach, a set of rules to describe the translation process so that an enormous amount of input from human experts is required [21] [22].

The other approach is corpus-based approach. In the corpus-based approach the knowledge is automatically extracted by analyzing translation examples from a parallel corpus built by human experts. The Hybrid Machine Translation Approach was created as the result of combination of the features of the two major approaches [9]. The aforementioned three MT approaches, namely, the rule-based, the corpus-based and the hybrid machine translation approaches are presented in detail as follows.

2.2.1. Rule-based Machine Translation Approach

According to Okpor [9], Rule-Based Machine Translation (RBMT) is a machine translation system based on linguistic information about source and target languages basically retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological, and syntactic regularities of each language respectively. RBMT is also known as Knowledge-Based Machine Translation or Classical Approach of MT. RBMT system generates input sentences (in some source language) to output sentences (in some target language) on the basis of morphological, syntactic, and semantic analysis of both the source and the target languages involved in a concrete translation task [21].

A set of linguistic rules of RBMT methodology is applied in three different phases, namely, analysis, transfer and generation [22]. Hence, RBMT system requires the following four steps, namely, syntax analysis, semantic analysis, syntax generation and semantic generation that are shown in Figure 2-1.

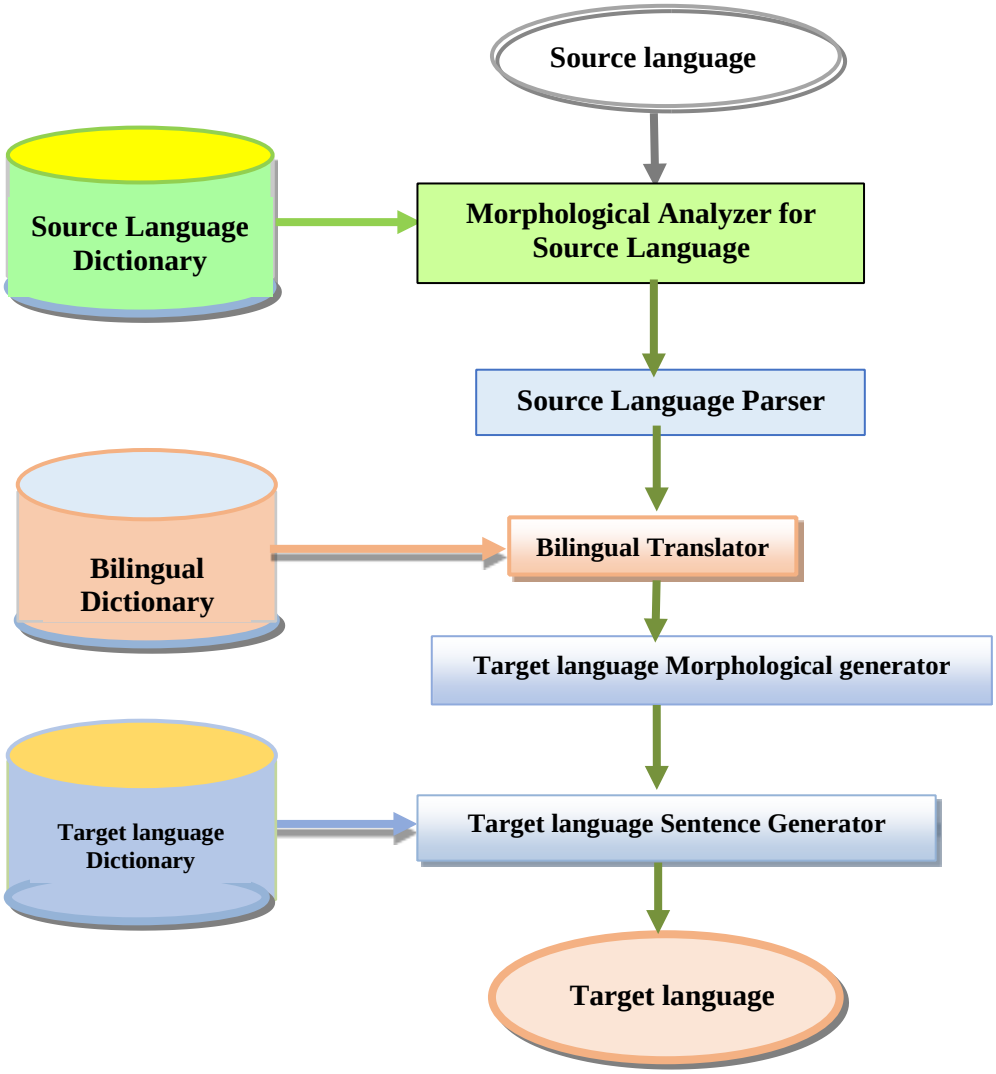

Figure 2.1: Architecture for a Rule-based Machine Translation System

Source: Taken from [23]

Okpor [9] has listed the following shortcomings that are inherent in RBMT approach;

- Insufficient number of good dictionaries. building new dictionaries is expensive;
- There is some linguistic information that needs to be set manually,
- Regarding the systems, it is hard to adjust to new fields to rule interactions that may result in ambiguity as well as failure.

- The last shortcoming is its failure to adapt to new fields or domains. RBMT systems usually provide a mechanism to create new rules and extend and adapt the lexicon, nevertheless, changes is usually very costly and the results, frequently, do not pay off.

According MOSSES [24], there are number of sub-approaches under the rule-based machine translation approach, namely: Direct, Transfer, and Interlingua machine translation approaches. The sub-approaches differ in the depth of analysis of the source language and the extent to which they attempt to reach a language-independent representation of meaning between the source and target languages.

Accordingly, the aforementioned sub-approaches used in Rule-based MT are briefly discussed as follows.

2.2.1.1.Direct Machine Translation Approach

Though it is the oldest, direct machine translation (DMT) approach is less popular approach [24]. DMT is made at the word level. Machine translation systems that use this approach are capable of translating a language, called source language (SL) directly to another language, called target language (TL). There is no need to pass the translated words through an additional/intermediary representation. Words of the SL are translated directly. The analysis of SL texts is oriented to only one TL. Direct translation systems are basically bilingual but uni-directional. Direct translation approach needs only a little syntactic and semantic analysis. SL analysis is oriented specifically to the production of representations appropriate for one particular TL. DMT is an approach that uses some simple grammatical adjustments and applying a word-by-word translation approach.

2.2.1.2. Interlingua Machine Translation Approach

For the translation of more than one language, Inter-lingual MT approach is used to translate source language text. Such a translation is from source language to an intermediate form called inter-lingual and then from inter-lingual to target language [25]. The rule-based machine translation approaches have the Inter-lingual machine translation as one of their instance. In the Inter-lingual machine translation approach, the source language text that is to be translated would be transformed into an inter-lingual language, that is, a language neutral representation. In this

case the inter-lingual generates the target language. The inter-lingual becomes more valuable as the amount of target languages it can be turned into increases. This is one of the major advantages of this system that makes it to become the most attractive for multilingual systems [1] [9].

2.2.1.3. Transfer-based Machine Translation

According to Jurafsky, et. al and Woin [1] [9], the transfer-based machine translation creates a translation from an intermediate representation that relates the meaning of the original sentence. This is what makes this approach similar with inter-lingual MT. However, unlike inter-lingual MT, it depends partially on the language pair involved in the translation. They [1] [9] further elaborate that on the basis of the structural differences between the source and target language, a transfer system can be broken down into three different stages: i) Analysis, ii) Transfer and iii) Generation. In the first stage, the SL parser is used to produce the syntactic representation of a SL sentence. the result of the first stage is converted into equivalent TL-oriented representations in the next stage. A TL morphological analyzer is used to generate the final TL texts, which is the final stage.

2.2.2. Corpus-based Machine Translation Approach

The dominance of the rule-based approach has been broken by the emergence of new methods and a strategy which is called the Corpus-based Machine Translation Approach (CBMT). CBMT that is referred as an alternative approach for machine translation to overcome problem of knowledge acquisition of rule-based machine translation [26]. It emerged as a dominant new method and strategy over the two preceding approaches. CBMT uses, as it names indicates, a bilingual parallel corpus to obtain knowledge for new incoming translation. A large amount of raw data in the form of parallel corpora is used by this approach. Text and their translations are included in this raw data. These corpora are used for acquiring translation knowledge [9]. Corpus-based approach is classified in to two approaches namely, Example-Based Machine Translation (EBMT), and Statistical Machine Translation (SMT) are the two classifications of CBMT. The approaches are briefly explained in the following section.

2.2.3. Example-Based Machine Translation

Memory based translation is another name for Example-based Translation (EBMT). EBMT is based on recalling/finding analogous examples (of the language pairs). The EBMT system is given a set of sentences in the source language (from which one is translating) and corresponding translations of each sentence in the target language with point to point mapping. These examples are used to translate similar type of sentences of source-language to the target language. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again [24]**.**

The fact that EBMT avoids the need for manually derived rules makes it an attractive approach to translation. However, to produce the dependency trees needed for the examples database and for analyzing the sentence it requires analysis and generation modules. A designated drawback in EBMT is computational efficiency, especially for large databases, although parallel computation techniques can be applied [9].

Accordingly, there are three major components of EBMT, as were indicated by Nagao, et al [27], are matching fragments against a database of real examples; identifying the corresponding translation fragments; and then recombining these to give the target text.

2.2.4. Statistical Machine Translation

According to Burnings [28], statistical machine translation (SMT), which is one of the corpus based translation classifications, is generated on the basis of statistical models. The general objective of SMT is to extract general translation rules from a given corpus consisting of sufficient number of sentence pairs which are aligned to each other [29]. Burnings [28] further describes that the parameters for SMT are derived from the analysis of bilingual text corpora. Brown *et al.* [30] proposed that the initial model of SMT is based on Bayes Theorem. The Theorem takes the view that every sentence in one language is a possible translation of sentence in the other and the most appropriate is the translation that is assigned the highest probability by the system. Parallel corpus that uses human produced translations is applied in SMT machine translation approach [18]. According to Lopez [31], the SMT translation process is considered as a machine learning problem. SMT algorithms automatically learn how to translate new sentences after examining the parallel corpus. The machine learning algorithms learn how to translate new

sentences from the parallel corpus which is a collection of previously translated texts. The translation accuracy of these algorithms mainly depends on the parallel corpus regarding its domain, quantity and quality. So, a consistent preprocessing of the data yields a good translation quality.

The probabilistic models of faithfulness and fluency are built by SMT to select the most probable translation by combining models [18, 18, 28]. The main focus of SMT is not on the process but on the result of the translation to produce true translation which is both, faithful to the channel equation shows that two components are needed. These components are a translation model *P(F|E),* and a language model *P(E).* SMT works based on the Bayesian model which translates foreign language F to English (E) or source language and the best translation is selected depending on the highest value of the translation model (*P (E|F)*) [19, 30]. Therefore, the noisy channel via Bayesian rule is given as shown below.

$$E = argmax_E p(E|F)$$
$$= argmax_E \frac{p(F|E)p(E)}{p(F)}$$
$$E = argmax_E p(F|E)p(E)$$

Where P (E\F) = the translation model for foreign to English language

P (F/E) = the translation model for English to foreign translation

P (E) = language model for English

Thinking of things backwards, according to [19], is a requirement for applying the noisy channel model to machine translation. There is a need to pretend that the foreign (source language) input F must be translated in a corrupted version of some target (e.g. English) sentence *E,* and that the task is to discover the hidden (target language) sentence *E* that generates the observation sentence *F*. There are three components to translate from a foreign sentence *F* to an English sentence *E* as a requirement for the noisy channel model of statistical MT [18, 28,31]. These are the language models to compute *P (E),* translation model to compute *P (F |E)* and decoder, which is given *F* and produces the most probable *E*.

2.2.4.1. Architecture of Statistical Machine Translation

According to Abdullah [32], the SMT approaches have three components, namely, decoder, language model and translation models. These models attempt to process the source text and finally translated to target language text. In the case of a monolingual, the goal of language modeling is to assign n-gram that is, unigram, bigram, etc. to a sentence of target language. The translation model, on the other hand, is bilingual probability that is computed from a given source language sentence to generate target language sentence.

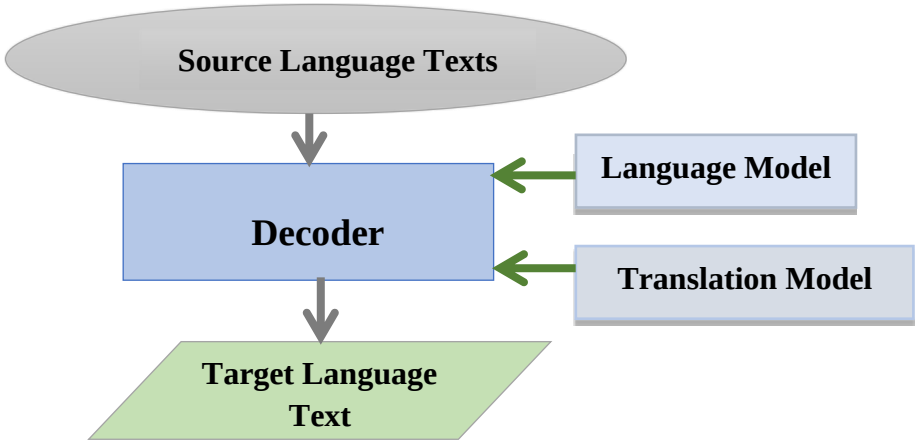

Figure 2.2: Statistical Machine Translation Architecture

Source: Taken and adopted from [33]

As depicted by this architecture, the noisy channel model of statistical MT thus requires three components to translate from a foreign sentence *F* to an English sentence E [34].

- A **language model** to compute *P(E)*
- A **translation model** to compute *P(F|E)*
- A **decoder**, which is given F and produces the most probable

2.2.4.2. Statistical Machine Translation Models

There are two statistical Machine translation models, namely, the language model and the translation model.

- **A Language model**

According to Maučec, and Donaj [35] language model is usually formulated as a probability p(s) over strings s that attempts to reflect how frequently a string occurs as a sentence. Given such a sequence with length m, it assigns a probability, P(w1,w2,w3…….,wm) to the whole sequence. The most widely-used language models, by far, are n-gram language models. N-gram language models are usually estimated over 3 to 5 grams. For example, trigram model means two words history are considered for predicting the third word. Bigrams model requires just one word to estimate the next one while unigram model disregards the previous words the unigram model is easy to estimate but it is not a good language model. Two similar sentences with different word order will have the same probability. N-gram probability can be computed as follow [32]:

$$p(ei|ei\text{-}n,\ldots..,ei\text{-}1) = \frac{count(ei-n,\ldots,ei-1,ei}{\sum_e count(ei-n,\ldots,ei-1,e}$$

Any corpus will not have all the possible sentences. Therefore, a language model based on sentence frequency might assign zero probability to a fluent sentence because it did not occur in the corpus. N-gram models manage to avoid assigning zero probability to unseen sentences by breaking up the estimation process into n-gram. However, if there is one n-gram in a given sentence that was not in the training data, the model will assign the sentence zero probability since the estimation is based on the product of all n-grams [32].

- **Translation model:**

It states that the most likely translation of a given sentence **G** is the sentence that maximizes the product of language model *p*(**T**) and translation model *p*(**G**|**T**) [32]. Therefore, the job of the translation model is to assign a probability that a given source language sentence (Geez) generates a target language (Tigrigna). As mentioned above, for a given source and target sentences **G** and **T**, it is the way sentences in **G** get converted to sentences in **T** which is denoted by (G|T) calculated as follows:

$$P(G|T) = \frac{Count(G,T)}{Count(G)}$$

Translation model assures suitable meaning while language model assures fluent output. In language modeling section, breaking the sentences into smaller parts enables us to collect sufficient statistics. The same approach will be applied in translation modeling.

The above equation may be difficult to achieve, if the sentences are too long. To overcome this problem the sentence is decomposed into words and sub-words called morpheme, as in language modeling [32].

$$p(G|T) = \sum_{X} P(G, X|T)$$

Where the variable X represents alignments between the individual chunks in the sentence pair where the chunks in the sentence pair can be morphemes, words or phrases. The variable X represents alignments between the individual chunks in the sentence pair where the chunks in the sentence pair can be morphemes or words or phrases. In morpheme-based translation, the fundamental unit of translation is a morpheme. Phrase-based translations, most commonly used, translates whole sequences of words, where the lengths may differ in which blocks are not linguistic phrases but, phrases found using statistical methods from corpus [6].

Translation models are generally divided into three types [32]: word-based (input sentence are translated word by word individually, and these words finally are arranged in a specific way to get the target sentence), phrase-based (each source and target sentence is divided into separate phrases instead of words before translation) and hierarchical phrase-based (hierarchical phrases have recursive structures instead of simple phrases).

Decoding: searches for the best sequence of transformations that translates source sentence to the target sentence [9]. It looks up all translations of every source morphemes, words, phrases, using word or phrase translation table and recombine the target language phrases that maximize the translation model probability multiplied by the language model probability can be computed as follow:

$$P(g|t) = \underset{t}{\operatorname{argmax}}(p(t|g) * p(g))$$

By following the above procedures, the decoder performs the translations of the input text for both languages. Decoders in MT are based on best-first search, a kind of heuristic or informed search; these are search algorithms that are informed by knowledge from the problem domain [1].

2.2.4.3. Challenges of Statistical Machine Translation

According to M. D. Okpor [9], there are issues on statistical machine translation these are:-

- **Sentence Alignment**: In parallel corpora single sentences in one language can be found translated into several sentences in the other and vice versa. Sentence aligning can be performed through the Gale-Church alignment algorithm.
- **Statistical Anomalies:** Real-world training sets may override translations of, say, proper nouns. An example would be that "I took the train to Berlin" gets mis-translated as "I took the train to Paris" due to an abundance of "train to Paris" in the training set.
- **Data Dilution**: This is a common anomaly caused when attempting to construct a new statistical model (engine) to represent a distinct terminology (for a specific corporate brand or domain). Training sets used from alternative sources to the specific brand to compensate for a limited quantity of brand-specific corpora may ‘dilute’ brand terminology, choice of words, text format and style.
- **Idioms**: Depending on the corpora used, idioms may not translate "idiomatically".
- **Different word orders:** Word orders in languages differ. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence and one can talk, for instance, of SVO or VSO languages. There are also additional differences in word orders, for instance, where modifiers for nouns are located, or where the same words are used as a question or a statement.

2.2.5. Neural Machine Translation Approaches

The state of the art that is used until very recently is called neural machine translation which is a new breed of corpus-based machine translation. It is similar to the statistical machine translation technology but completely different by their computational approach: neural networks it uses [37] . Sequence-to-sequence models or encoder-decoder networks are the alternative names for the neural machine translation systems. The systems were initially fairly simple neural network

models made out of two recurrent parts [39]. That is, is an approach to machine translation that uses an artificial neural network to predict the likelihood of sequences of words. It also consists of many small sub-components (words) that are tuned separately. Neural machine translation attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation. According to [37] [39], most of the proposed neural machine translation models belong to a family of encoder–decoders. Encoder is used by the neural network to encode a source sentence into a fixed vector and decoder, used to predict words in the target language are the two components of recurrent neural networks (RNN).

The main advantage of the encoder-decoder approach is that a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector and There is no separate language model, translation model, and reordering model, but just a single sequence model that predicts one word at a time [39]. Nevertheless, the neural machine translation has its own inherent disadvantages. The main disadvantages of neural machine translation (NMT) That are, they are time-consuming if target vocabulary is large, weak to OOV (out of vocabulary) problem, difficult to debug the errors, and needs high perform computing devices (GPU - graphic process unit [39].

2.2.6. Hybrid Machine Translation Approach

Hybrid machine translation uses both Rule-based Machine Translation (RBMT) and Statistical Machine Translation (SMT) to translate from Source languages to Target language [9, 21]. The hybrid approach can be used in a number of different ways. In some cases, translations are performed in the first stage using a rule-based approach followed by adjusting or correcting the output using statistical information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system [41].

2.3. Alignment

The usual approach to building a statistical machine translation system is to first build a model of alignment between the input and output languages. According to J. Brunning [28] alignment is the arrangement of something in an orderly manner in relation to something else. An alignment is a parallel segmentation of the two texts, typically into sentences, such that the n^{th} segment of the first text and the n^{th} segment of the second are mutual translations [40].

One of the limitations of current word alignment models for statistical machine translation is that they do not address morphology beyond merely splitting. However, current alignment models do not take into account the morpheme, the smallest unit of syntax, beyond merely splitting words. That is, it can be performed at different levels, from paragraphs, sentences, segments, words and characters [28]. Since morphology has not been addressed explicitly in word alignment models, researchers have resorted to tweaking SMT systems by manipulating the content and the form of what should be the so-called "word" [6].

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

Sentence alignment represents the basis for computer-assisted translation is represented by sentence alignment, terminology management, word alignment and cross linguistic information retrieval [39]. In a parallel text context, sentence alignment is the problem of finding a bipartite graph matching minimal groups of sentences in one language to their translated counterparts. Due to the fact that sentences do not always align one-to-one, the sentence alignment task is important [44]. Sentence alignment means identifying which sentence in the target language is a translation of which one in the source language [45]. Robustness and accuracy are two kinds of difficulties in automatic sentence alignment methods [46].

The size and domain of the parallel corpus used strongly influences the quality of translations produced in any statistical machine translation system [47]. Sentence-aligned parallel bilingual corpora, which originate in sentence aligned form, are not proved to have very useful for applying machine learning to machine translation. This makes the task of aligning such a corpus of considerable interest, and several methods have been developed to solve this problem. Ideally, a sentence alignment method should be fast, highly accurate, and require no special knowledge about the corpus of the two languages [48]. Sentence alignment of parallel corpus affect the performance of the machine translation especially on statistical machine translation based on the above concepts. Following the standard alignment models of Brown et al. [49], we assume one-to-many alignment for both words and morphemes. This function of mapping a set of word

positions in a source language sentence to a set of word positions in a target language sentence is known as word alignment ***aw*** [50].

On the other hand, a morpheme alignment ***am*** is a function mapping a set of morpheme positions in a source language sentence to a set of morpheme positions in a target language sentence. A morpheme position is a pair of integers (j, k), which defines a word position j and a relative morpheme position k in the word at position j [51].

2.3.1. Tools used for alignment

There are different tools developed for aligning corpus for different purpose of text processing according to Andre and William, et al [47, 48, 49, 50], The following are some common tools:

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

GIZA++ is a SMT toolkit freely available for research purposes. The original program called GIZA was part of the SMT toolkit EGYPT, developed at the center of language and speech processing at Johns Hopkins University by Liang, et al [50]. GIZA++ is part of the statistical machine translation toolkit used to train IBM Model 1 to Model 5 and the Hidden Markov Model.

- **IBM Model 1**

IBM Model 1is the simplest and the most widely used word alignment model among the models that the IBM group has proposed. , The other name for IBM 1 Model is a lexical translation model that uses an Expectation Maximization (EM) algorithm which works in an iterative fashion to estimate the optimal value for each alignment and translation probabilities in parallel texts. The IBM Model 1, given a Geez sentence G = *(g1, . . . gl)* of length *l* and Tigrigna sentence T = *(t1, . . . , tn*) of length n, ignores the order of the words in the source and target sentence and the probability of aligning word and is independent of their positions in string G and T, j and i respectively [24].

IBM Model 1 tries to identify a position j in the source sentence from which to generate the ^{h} target word according to the distribution in the context of noisy channel.

$$Pr(g|t) = \frac{\in}{(j+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{1} t(g_{j|t_i})$$

It denotes the translation probability of given and *denotes.* It is also assumed that all positions in the source sentence, including position zero for the null word, are equally likely to be chosen and there are acceptable alignments.

- **IBM Model 2:**

The IBM Model 2 has an additional model for alignment that is not present in Model 1 [24]. The IBM Model 2 addressed this issue by modeling the translation of a foreign input word in position to a native language word in position using an alignment probability distribution defined as: a(i V j, l_e ,l_f) in this equation, the length of the input sentence f is denoted as l_f , and the length of the translated sentence e is l_e .

Assuming t(e | f) is the translation probability and a(i V j, l_e ,l_f) is the alignment probability, IBM Model 2 can be defined as:

P(e, $a|f$) = $\in \prod_{j=1}^{l_e} t(ej\ V\ f_{a|j}) a(j) V\ j, l_{e,}\ l_f$),

The alignment function maps each output word to a foreign input position ➢

IBM Model 3

A Single word in the source language may not map to exactly one word in the target language [24]. Model 3 adds the fertility probability n(s_j) which is equal to the likelihood of each source word translated to one word, two words, three words, and so on, on top of Model 2 parameters Modeled by distribution $n(\emptyset|f)$. The number of inserted words depends on sentence length. This is why the NULL token insertion is modeled as an additional step to the fertility step.

IBM Model 3 can be mathematically expressed as:

$$p(S|E,A) = \prod_{i=1}^{I} \varphi i\ !\ n(\varphi|ej) * \prod_{i=1}^{J} t\ (fi|eaj) * \prod_{i:a(i)\neq 0}^{J} d(j|aj, I, J) * \binom{J-\varphi 0}{\varphi 0} P_0^{\varphi 0} P_1^{j},$$

Where represents the fertility of each source word is assigned a fertility distribution , and refer to the absolute lengths of the target and source sentences, respectively.

➢ **IBM Model 4**

The set of distortion probabilities for each source and target position (i.e., the probability of a word in the source sentence change its position in the target sentence). As opposed to Model 2 which does absolute reordering, model 4 does relative reordering. ➢ **IBM Model 5**

Model 5 removes the deficiencies of the previous models [1-4]. For example, Model 4 can stack several words on top of one another. It can also place words before the first position or beyond the last position in the target string. Therefore, Model 5 fixes deficiencies like this one that the previous models have not handled [24].

2.4. Morphological Segmentation

A Linguistic Operation wherein words are separated into their composite morphemes is called morphological segmentation. The smallest possible building blocks of language that also have meaning when alone are called morphemes [49]. Morphemes are usually divided into two groups, i.e. stems and affixes; stem defines the basic meaning of a word, whereas affixes define the various forms of meaning of the word.

Morphemes are usually divided into two groups, that is, stems and affixes. The stem defines the basic meaning of a word, whereas affixes define the various forms of meaning of the word. For instance, consider the word 'unsegmented'. This word is consisted of 3 morphemes - 'un', 'segment' and 'ed'. Morphemes are used in a variety of linguistic tasks. They are used in understanding word structure and word formation., Morphology is used in text preprocessing tasks in Natural Language Processing (word stemming and lemmatization) and generating vector-space representations of words [49].

2.4.1. Segmentation tool

Morfessor model is to discover as compact a description of the input text data as possible. Substrings occurring frequently enough in several different word forms are proposed as morphs and the words are then represented as a concatenation of morphs, e.g., ‘hand, hand+s, left+hand+ed, hand+ful’. *The model* uses unsupervised training but still gives better results in most cases than other rule based natural language models and supervised machine learning models [49]. From the alignment tools mentioned above we used MGIZA++ and Morfessor for

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

2.4.2. Identifying Morphemes

In identifying morphemes, Morfessor Baseline takes a corpus as input and segments its words into a set of morphs without labeling the corpus [50]. Maximum Aposteriori estimate (MAP) is the basis for Morfessor algorithm.

The probability of the model of language P(M) and the maximum likelihood (ML) estimate of the corpus conditioned on the given model of language, written as P(corpus | M) are the two MAP estimate components. The algorithm looks for a much that has the highest probability in the given the corpus:

$$\underset{\mathrm{m}}{\operatorname{argmax}} P(\mathrm{M}|\ \mathrm{corpus}) \ = \ \underset{M}{\operatorname{argmax}} P(\mathrm{corpus}|\mathrm{M}) \ \cdot \ \mathrm{P}(\mathrm{M}))$$

P(M) = P(Lexicon, grammar): is the joint probability of the probability of the induced lexicon and grammar.

Where (Lexicon) = {μ1, μ2 ,… … … . . , $\mu|L|$} is the morph lexicon,
“Lexicon” refers to an inventory of whatever information one might want to store regarding a set of morphs. It also includes a set of morphs interrelations [50]. Suppose that the lexicon consists of M distinct morphs, the probability of coming up with a particular set of M morphs µ1 . . . µM making up the lexicon can be written as:

P(lexicon) = M! · P(properties(µ1), . . . , properties(µM))
M! is explained by the fact that there are M! Possible orderings of a set of M items and the lexicon is the same regardless of the order in which the M morphs emerged.

In the Baseline versions of Morfessor, the only properties stored for a morph in the lexicon is the frequency (number of occurrences) of the morph in the corpus and the string of letters that the morph consists of. Assuming independence of strings and frequencies, we can write:

P(properties(µ1), . . . , properties(µM)) = P(fµ1 , . . . , fµM) · P(sµ1 , . . . , sµM), where f represents the morph frequency and s the morph string.

To estimate probability distribution of the morph frequencies Morfessor Baseline uses the noninformative prior:

$P(f\ \mu 1, \ldots f_{\mu|L|}) = \frac{1}{\binom{N-1}{|L|-1}}$, Where $N = \sum| JL = |1\ f\ \mu j$ (number of morph tokens in the corpus).

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

$P(S\mu 1, \ldots . S\mu|L|) = \prod_{K=1}^{|L|} P(S_{\mu k})$ and all the characters within the morph are also independent:

$P(S\mu k|) = \prod_{K=1}^{l_k} P(C_{ik})$,

Where $s\ \mu k\ = C1, \ldots\ Clk$, and $P(Cik)$ is the character probability distribution over the alphabet estimated by counting its frequency in the corpus.

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

$(l) = (1 - (\#))\ (\#)$, Where # is a special end-of-morph character.

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

$P((\text{Corpus}|M)) = \prod_{j=1}^{w} \prod_{k=1}^{nj} P(\mu_{jk})$, Where W is the number of tokens in the corpus and (μi) is estimated by counting its frequency:

$$P(\mu_i) = \frac{f_{\mu j}}{\sum_{j=1}^{|L|} f_{\mu j}}$$

The algorithm uses the following data structure [50].

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

2.5. Machine Translation Evaluation

The evaluation of machine translation systems is a vital field of research. It both determines the effectiveness of existing MT systems and optimizes the performance of MT systems. Judging machine translation quality is defined as machine translation evaluation, there are two common types of MT evaluation:

- **Manual Evaluation also called Human Raters:**

The most accurate evaluations to evaluate each translation along the two dimensions use human raters. Example, along the dimension of fluency, we can ask how intelligible, how clear, how readable, or how natural the MT output (the target text) is [54]. Human evaluations of machine translation are extensive but expensive. Human evaluation is laborious. It can take months to finish and involve human labor that cannot be reused.

There are two methods to use human raters to answer the questions [54]. One method is to give the raters a scale, for example from 1 (totally unintelligible) to 5 (totally intelligible), and ask them to rate each sentence or paragraph of the MT output. They can use distinct scales for any of the aspects of fluency, such as clarity, naturalness, or style. The second methods relies less on the conscious decisions of the participants. For instance, we can measure the time it takes for the raters to read each output sentence or paragraph. Clearer or more fluent sentences should be faster or easier to read.

The two different perspectives by which the quality of MT output is judged by experts in translation and linguistics are accuracy and fluency [55].

In accuracy, source text adherence is judged to the source text norms and meaning, in terms of how well the target text represents the information content of the source text. The source text and translation being judged are accessed by the evaluators. Frequently, the context of a sentence is also taken into account. The evaluation requires to be bilingual in both the source and target languages.

In fluency, the degree of adhere to the target text and target languages norms, referring, for example, to features such as grammatical and clarity., The source text is not relevant when judging fluency. In fluency, the evaluators have access to only the translation being judged and

not the source data. Fluency demands a fluent expert only in the target language. The adequacy and fluency are usually judged on a Likert 5-point scale [55].

➢ **Automatic Evaluation**

Automatic evaluation metrics are cost-free or cost-effective alternatives to human evaluation and are used in the development of MT system to estimate improvement [55].

Bilingual Evaluation Understudy (BLEU), National Institute of Standards and Technology (NIST), Translation Error Rate (TER), Precision and Recall, and Metric for Evaluation if Translation with Explicit ORdering (METEOR) are different types of heuristic evaluation methods [54]. All heuristic methods except BLEU require human translation and time consuming. In BLEU each MT output is evaluated by a weighted average of the number of *N*gram overlaps with the human translation.

The Other score metrics widely used for automatic evaluation of machine translation output is BLEU score [55]. The basic assumption is that a translation of a piece of text is better if it is close to a high-quality translation produced by a professional translator. The translation hypothesis is compared to the reference translation, or multiple reference translations, by counting how many of the n-grams in the hypothesis appear in the reference sentence(s); better translations will have a larger number of matches.

According to M. S. Mirjam and D. Gergor [55], BLEU is based on precision and is starting computed with just unigrams. Unigram precision is calculated by finding the number of words in the candidate sentence (MT output) that occur in any reference transcription and dividing by the total number of words in the candidate sentence. Unigram is not an accurate measurement of translation quality as the system can generate many words that occur in the references but not output grammatical or meaningful sentences. Bleu uses a modified N-gram precision metric. Ngrams in the test set to avoid this problem.

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

$$Pn = \frac{\left(\sum_{C\in\{candidates\}}\sum_{n-gram\in C} Count_{clip^{(n-gram)}}\right)}{\sum_{C'\in\{candidates\}}\sum_{n-gram'\in C'} Count_{(n-gram')}}$$

2.6. Related Prior Works

The following are some researches that are related to machine translation models. And their methods, results and the way forward are presented in brief as follows.

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

The research was conducted by Tadesse Kassa in 2018 with the purpose to design morphemebased bi-directional machine translation for Ge'ez-Amharic textual documents.

Corpus preparation and preprocessing was collected from online sources. Such Online sources include Old Testament of Holy bible and anaphora (or Kidase). The corpus includes manually prepared bitext from Wedase Maryam, Anketse Berhane, yewedesewa melahekete, Kidan and Liton. To make the corpus suitable for the system, different preprocessing tasks such as tokenization, cleaning and normalization have been done. The data set contains a total of 13,833 simple and complex sentences, out of which 90% and 10% are used for training and testing, respectively. To build a language model for both languages we used 12, 450 parallel sentences. For both statistical and rule-based approaches we used Mosses for translation process, MGIZA++ for alignment of word and morpheme, morfessor and rules were used for morphological segmentation and IRSTLM for language modeling. After preparing and designing the prototype and the corpus, different experiments were conducted. Dataset being prepared using unsupervised morpheme segmentation performs 14.54% and 14. 88% BLEU score from Geez to Amharic and from Amharic to Geez respectively. And also dataset prepared using rule-based segmentation performs 15.14% and 16. 15% BLEU score from Geez to Amharic and from Amharic to Geez respectively. As we compare the result rule-based morpheme segmentation performs better than unsupervised morphological segmentation. This is due to rule-based morpheme segmentation uses rules well-crafted by linguist that directs to the morphemes of the language.

This study achieves a promising result that identifies morpheme as an optimal unit of translation and it enhances the performance of bi-directional Ge'ez-Amharic machine translation.Rule-based morpheme segmentation requires linguistic knowledge to generate well-crafted rules, time taking, resources incentive and it is long term work plan. On the other hand the unsupervised morpheme segmentation technique generates the rules from corpus of the language, which is economical and doesn't need linguistic knowledge.

- **English-Afaan Oromoo Machine Translation: An Experiment Using Statistical Approach**

The research was conducted by Sisay Adugna Chala in 2009, with the aim to develop a prototype English-*Afaan Oromoo* machine translation system using statistical approach, i.e, without explicit formulation of linguistic rules

There are two possibilities to translation, namely: Manual Translation (in which any translation task is carried out by human translators) and Automatic or Machine Translation (in which any translation task is carried out by computer software). Thus, the focus of this research is on automatic or machine translation from English to Afaan Oromoo.

Evaluation is done using the BLEU (Papineni et al., 2002) scoring tool. Using a reference translation prepared manually from the parallel corpus, the translation quality of the system output which was translated can be evaluated.

In this research, experimentation of statistical machine translation of English to Afaan Oromoo was conducted and a score of 17.74% was found. Although Afaan Oromoo is among resourcescarce languages (Kula et. al., 2008) of the world, the result of this experiment shows that the amount of data available can be used as a good starting point to build machine translation system from English to Afaan Oromoo. The researcher believes that these tools and techniques should be applied for other languages in Ethiopia to help the speakers of the languages reap the benefits of getting documents available in English without renouncing their own language.

- **Geez to Amharic Automatic Machine Translation: A Statistical Approach**

The thesis was conducted by Dawit Mulugeta in 2015, with the objective to investigate the application of Statistical Machine learning technique to Machine Translation from Geez to Amharic.

The required amount of parallel data, a Holy Bible Geez-Amharic translation and some other religious books (Wedase Mariam and Arganon) are used. 12860 parallel sentences are used for the training and testing. The collected data were divided in to training and testing set in such a way that more than 90% of the collected data was used as a training set.

The collected data are further preprocessed so as to make the data fit to the modeling tools requirement. These include breaking of the documents into sentence level in such a way that separate sentences appear on a separate line and corresponding Geez and Amharic documents being on different files with corresponding sentences on corresponding lines. With some expectation in the Geez versions, most of materials were inherently verse level aligned and sentence level alignment was not required. Some document (Widase Mariam and part of Arganon), which are not aligned at sentence level were aligned manually.

SMT uses different tools in order to build the language model, the word alignment model and decoding. Language modeling (LM) is the attempt to capture regularities of natural language for the purpose of improving the performance of various natural language applications. The word alignment tries to model word-to-word correspondences between source and target words using an alignment modeling. Whereas, decoding is the process of searching among all possible translation for a given source sentence from the huge different possible translation for each word (phrase) with different ordering in sentence.

The common statistical MT platform, namely Moses, is used for the translation. Moses is selected due to the familiarity of the researcher to the tool and because of its accessibility, processing capability and language independent features. Moses consists of all the components needed to preprocess data, train the language models and the translation models (decoding) (Och, 2003).

3

Although Moses integrates both the IRSTLM and SRILM language modeling toolkits, the IRSTLM, which requires about half memory than SRILM for storing an equivalent LM during decoding (Federico et.al, 2007), is used in this research.

The BLUE score for Hebrew to Arabic translation (Shilon, 2012), which are both morphologically rich languages, is 14.3%. As well, the BLUE score for English to Affaan Oromo (Sisay, 2009) was 17.74%.

Accordingly, the average result that was achieved at the end of the experimentation was 8.26%. We have found that increasing the Amharic monolingual corpus can enhance the accuracy of the language modeling and the translation result. The performance of the system appears relatively low as compared to the performance of other experiments performed on huge amount of data. First reason for the low performance is the morphological complexity of the two languages. Geez and Amharic are related but with scarce parallel corpora. Machine translation between the two languages is therefore challenging and requires exploring different approaches. Due to time constraints the researcher was not able to test the approach. The researcher recommends future research of Geez – Amharic translation should be undertaken using Example-based Machine Translation approach which is the other corpus based machine translation approach and requires relatively small amount of bilingual data for training (Dandapat, 2010).

- **Bidirectional Tigrigna-English Statistical Machine Translation**

This thesis was conducted by Mulubrahan Hailegebreal in 2017 with the aim to develop a bidirectional Tigrigna–English machine translation system using statistical machine translation approach.

In this work, experimental quantitative research method is used. This research has been conducted by developing thirty types of experiments all based on Tigrigna - English and English – Tigrigna Statistical based Machine Translation

The direction should be in the application of methods that help to get semi-supervised segmentation model to segment Tigrigna morphology as the processed segmentation experiment outperformed the other experiments (baseline and morph-based experiments). Since this method

only segments prepositions and conjunctions, there should be a mechanism to apply more techniques to segment the other partsof-speech as well.

CHAPTER THEREE

3. GE'EZ AND TIGRIGNA LANGUAGES

3.1. Geez and Tigrigna Languages Writing System

Writing system is a set of rules for using one or more scripts, to represent human language in written form. The Tigrigna writing system uses Ge'ez syllable or alphabet called "Fidel /"ፈደል" meaning letter, which was adapted from Ge'ez, the extinct classical language of Ethiopia.

3.2. Syntax

Tigrigna sentences should have at least two components the subject and a finite verb. Tigrigna sentence structure follows subject ("በዓል-ቤት) (beOel-Biet)" object ("ተስሓቢ) (tesHebi)" verb ("ግሲ) (gsi)" word order (SOV) (Tsehaye, 1979), (Tewelde, 2002), (Teklu, 2008), (Tegay, 2014) [57], whereas, the syntax of Ge'ez follows SVO, VSO and OVS [6].

	ሥርዉ ፊደል	ነድነቱ ሠራዊተ ፊደላት					
	ግእዝ	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	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
፲፪	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
፲፫	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
፲፬	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
፲፭	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
፲፮	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
፲፯	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
፲፰	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
፲፱	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
፳	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
፳፩	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
፳፪	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
፳፫	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
፳፬	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
፳፭	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
፳፮	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
፳፯	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Table 3.2.1.A: Previous Ge'ez Alphabet

	ሥርው ፊደል	ስድስቱ ሠራዊተ ፊደላት					
	ግእዝ	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	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
፲፪	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
፲፫	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ

	ሥርው ፊደል	ስድስቱ ሠራዊተ ፊደላት					
	ግእዝ	ካዕብ	ሣልስ	ራብዕ	ሐምስ	ሳድስ	ሳብዕ
፲፬	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
፲፭	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
፲፮	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
፲፯	የ	ዩ	ይ	ያ	ዬ	ይ	ዮ
፲፰	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
፳	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
፳፩	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
፳፪	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
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 3.2.1.B Current Ge'ez Alphabets

ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
..		ቊ	ቋ	ቌ	ቍ	
ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ
ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
ኰ		ኵ	ኳ	ኴ	ኵ	

ኰ		ኵ	ኳ	ኴ	ኵ	
ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ጐ		ጒ	ጓ	ጔ	ጕ	
ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Table 3.2.2: Tigrigna Alphabets

3.3. Number system (አሃዝ)

Geez has its own numerals for designating numbers. Tigrigna number system uses Ge'ez numbering systems. It has twenty characters. They represent numbers from one to ten (፩-፲), twenty to ninety (፳-፺), hundred (፻) and thousand (፼). However, these are not suitable for arithmetic computation purposes because there is no representation for zero (0), decimal points. Tigrigna" numbering system can be classified as ordinal numbers, cardinal numbers and fraction. The cardinal numbers are numbers like "ሓደ" (one), "ክልተ" (two), "ሰለስተ" (three), "ዓሰርተ" (ten), etc…, the ordinal numbers are "ቀዳማይ" (first), "ካልኣይ" (second), "ሳልሳይ" (third), "ዓስራይ" (tenth), etc.… and fraction numbers are also special numerals in Tigrigna that correspond to the English like:

"ፍርቂ" (half), "ርብዒ" (quarter), "ሲሶ" (one-third) etc.

Ge'ez no.	-	፩	፪	፫	፬	፭	፮	፯	፰	፱	፲
Arabic no.	0	1	2	3	4	5	6	7	8	9	1
Ge'ez	አልቦ	አሐዱ	ክልኤቱ	ሠለስቱ	አርባዕቱ	ሐምስቱ	ስድስቱ	ስብዓቱ	ስመንቱ	ተሰዓቱ	አ
Tigrigna	ባዶ	ሓደ	ክልተ	ሰለስተ	ኣርባዕተ	ሓሙሽተ	ሽዱሽተ	ሸውዓተ	ሸሞንተ	ትሽዓተ	ዓ
Ge'ez no.	፳	፴	፵	፶	፷	፸	፹	፺	፻	፼ ፼	፲
Arabic no.	20	30	40	50	60	70	80	90	100	10,000	1
Ge'ez	ዕሥራ	ሠላሳ	አርብዓ	ሃምሳ	ስሳ	ሰብዓ	ሠማንያ	ተስዓ	ምዕት	እልፍ	አ እ
Tigrigna	ዒስራ	ሰላሳ	ኣርብዓ	ሓምሳ	ስልሳ	ሰብዓ	ሰማንያ	ቴስዓ	ሚኢቲ	ዓሰርተሽሕ	ሚ ሽ
Ge'ez no.	፼፼	፲፼፼	፼፼፼	፲፼፼፼							

Ge’ez	አእላፍት	ትእልፊት	ትልፊታት	ምእልፊት
Tigrigna	ሚሊየን	ዓሰርተሚሊየን	ሚኢቲሚሊየን	ቢልየን

Table 3.3: Ge’ez and Tigrigna Numeral

3.4. Similar Letters

Similar letters are letters that have similar sounds but they are different in shape Orthographically.

Ge’ez language has similar letters but in Tigrigna language there is no similar letters that have similar sounds like Ge’ez language.

Similar letters in Ge’ez	
Sound	**Letters**
hä	ሀ፣ሐ፣ኀ
ṣä ፣ śä	ሰ፣ሠ
ʾä ፣ ʿä	አ፣ዐ
ṣä ፣ ʦä	ፀ፣ጸ

. Figure 3.4 Similar Letters of Ge’ez

3.5. Word Classes

Word class category or simply lexical category refers to classes in which a given word can be resided. The term word class is used to identify word behavior in the sentence or corpus. Each word that we use for speech as well as writing has its own part of speech. Based on parts of speech a word of grammarians classified words in to eight major parts in both Ge’ez and Tigrigna [58]

[59]. These are Nouns/ “ሹም”, Verbs/”ግሳት”, Adjectives/”ቅፅላት”, Adverbs/”ተውሳኽ-ግሳት, Pronoun/ተውላጠ-ሹም/ክንዲ-ሹም, Preposition/መስተዋድድ, Conjunction/መስተፃምር and

Interjection/ቃል አጋኖ” [58][59]. The syntactic structure is formed by combining different words in sequence. The syntactic structure of Tigrigna is generally SOV whereas Geez follows SVO, VSO and OVS word order for declarative sentences. The Tigrigna equivalent for the Geez sentence “ውእቱ መጸአ እምቤቱ [weetu metsa embet]” is “ንሱ ገዛ መፂኡ [nsu geza metsie’u]” meaning “He came home” where “ንሱ [nsu]” is the subject of the Tigrigna sentence equivalent to “ውእቱ [weetu] in the Geez , “ገዛ [geza]” is the object of the Tigrigna sentence equivalent to “እምቤት [embet]” in the Geez and “መፂኡ [metsie’u]” is the verb of the Tigrigna sentence which is equivalent to “መጸ /metsa” in Geez . But usually pronouns are omitted in both Geez and Tigrigna sentences and become part of the verb when they used as a subject “መጽአ እምቤቱ [metsa embet]” equivalent to “ገዛ መፂኡ [geza metsie’u]”.

3.5.1. Parts of Speech

3.5.1.1. Noun /“ሹም”

Noun in Ge’ez and Tigrigna are name given for people, place, animal, feeling, quality, action and idea. Noun can be also sub divided as common noun, proper noun, concrete noun, abstract, countable. Most nouns in both languages ends with the sixth letter, “Sadese Fidel”, It doesn’t mean that it never ends by other letters or Fidel. Both Tigrigna and Ge’ez nouns have plural forms to represent a number of things that share common characteristics. However, in Both Tigrigna and

Ge’ez, the most complex and difficult part of the languages is there is no common system of converting a singular form to its plural forms. Even though there is no common system of converting a singular form to its plural forms there are two ways forming plural forms of a nouns in both the languages. These are the following:

- Pattern replacement (broken plurals):

- Ge’ez: - “ደብር dabr”-------- “አድባር adbar”

“ሀገር hager” ---------“ አህጉር ahagur”

“ቤት bet” -------- “አብያት”

- Tigrigna:- “ ወዲ wɵdi” ……”አወዳት awɵdat”

“ጊመል gimɐl”…….” ኣግማል /agmal”

“መንበር mɐnbɐr” ……. “መናብር mɐnabɨr”

- Addition of an ending (Internal plurals):

- Ge’ez: - “አመት” ---------- “አመታት ፣ ስዕል”……“ስዕላት”
- Tigrigna “ዓመት” ---------- “ዓመታት ፣ ሰብ”……“ሰባት”

Plural Nouns formed by pattern replacement are often referred to as ‘broken plurals’ or ‘internal’ plurals; those formed by adding suffixes, as ‘ external' plurals [6]. The two endings used to form external plurals are **-ãn (ኦነ)** and -“**ãt /ኦት**”. -“**ãn**” is, for the most part, restricted to nouns denoting male human being. Most Ge’ez nouns form their plural form using broken plural or internal plural ways. In Ge’ez languages we use “አ ፤ አ……ት ፤ ን ፣ት” to inflect a singular noun to Plural and also in Tigrigna we use some of them.

Ge’ez			Tigrigna		
using	Original word	Inflicted word	Using	Original word	Inflicted word
አ	ልብ	አልባብ	ኣ	ልቢ	ልብታት
አ……ት	ባሕር	አብሕርት	ኣ	ባሕሪ	ባሕርታት
	ገብር	አግበርት	ታት or ት	ባርያ	ባርያታት፣ባሮት
ት	ገዳም	ገዳማት	ት	ገዳም	ገዳማት
	እም	እማት	ታት	አዶ	አዶታት/አዴታት
ል	ኪሩብ	ኪሩቤል	ል	ኪሩብ	ኪሩቤል
	ሱራፊ	ሱራፌል	ል	ሱራፊ	ሱራፌል
ን	ጻድቅ	ጻድቃን	ን or ናት	ጻድቅ	ጻድቃን/ጻድቃናት
ው	እኁ	አኀው	?????	?????	?????
	አብ	አበው		ኣቦ	ኣቦታት

Table 3.5. Example of infliction in numerals in Ge'ez and Tigrigna (Adopted from Tadesse, 2018)

3.5.1.2. Adjective (“ቅፅል”)

Adjectives are one of the four major word classes, and its main purpose is to give clear explanation for a noun (i.e., talk about things behavior or characteristics, like shape, size, color, type, property).

Adjectives in Ge’ez and Tigrigna are based on property, size, shape, color. Most Tigrigna adjective was found in front of a noun where as In Ge’ez language adjectives are used before and after noun.

For example: - “ፍንዋን እደው ይነግሩ መልእክተ ።።” (Geez)

“ዝተልአኩ አወዳት መልእኽቲ ይዛረቡ።።” (Tigrigna)

“እደው ፍንዋን ይነግሩ መልእክተ።።” (Geez) ፣

“ዝተልአኩ አወዳት መልእኽቲ ይዛረቡ።።” (Tigrigna).

“ብእሲ ሐጺር” (Geez) ፣

“ሓፂር ወዲ” (Tigrigna)

There are many ways of creating plural form of adjectives in Tigrigna and Ge’ez language. One way of making plural in Tigrigna is by adding affixes (“-ኦ፣-ኣት፤-ኦት፤ቲ”) to a given word [57]. and also creating plural form of adjective in Ge’ez language by adding prefix “እለ ፣ አ” at the beginning and adding suffix “ን ፣ ዊ/ይ ፣ ያ ፣ ት ፣ ሙ ፣ ው …” at the end [59][6]. A detail explanation was given in Table 3.5.1.2.

Tigrigna				**Ge’ez**			
Singular	Plural	Prefix	Suffix	Singular	Plural	Prefix	Suffix
ኣቦ	ኣቦታት		… ኣት	አብ	አበው		…ው
ሓፂር	ሓፀርቲ		… ቲ	ሐጺር			
ዘበናይ(fashion)	ዘበነኞት		… ኦት				
ሸቃሊ (labor)	ሸቃሎ		… ኦ				

Table 3.6.: Tigrigna and Ge’ez singular plural prefix and suffix

3.5.1.3. Verb /“ግስ”

Verb is a word used to describe an action, state, or occurrence, and forming the main part of the predicate of a sentence [59]. Verbs in Tigrigna mostly are placed at the end of the sentence whereas in most Geez sentences the verbs are placed in the middle. In both language verbs have two types of ending: one relating to the subject and one relating to the object. Thus, the affix attached to the verb can simultaneously agree with the subject or the object.

In Ge’ez and Tigrigna there are two types of verbs regular and irregular verbs based on the affix used to form. Tadesse Kassa [6] argued that, Regular verbs are main verbs that have four types; namely, “ቀዳማይ/ሓላፊ” past tense/perfect, “ካልአይ/ናይ ሕጂ ን መፃኢን” present and future /imperfect, command and “ዘንድ” to verbs. “ትዕዛዝ” command and “ዘንድ” to verbs are the same.

Perfect verb shows the action is past or completed, which include past-perfect, past-continuous, past-participle with relative pronoun ዘ (of), whereas imperfect verb includes present-continuous and future action. The end of all perfect verbs is the first order while all imperfect verbs ends with the 6th order when the noun is “ውእቱ” he. Morphology of verbs starts with perfect verbs. To change imperfect verbs, it has its own rules which is expressed by the root verbs /“ግስ አርእስቲ” [59] [60].

Root verb in Ge’ez are eight and have their own characteristics [59].
These are: -

Head	**Number of radicals**	**Pronunciation**
ቀተለ	Tri-radical, 1-1-1	/*ḱətələ*/
ቀደሰ	Tri-radical, 1-1-1	/ḱəddəsə/
ኀህለ	Tri-radical, 6-6-1	/*ḫiḫilə*/
አእመረ	Quadric-radical ,1-6-1-1	/ʔəʔmərə/
ሴሠየ	Bi-radical, 5-1-1	/śemə/

ባረከ	Tri-radical, 3-1-1	/*barəkə*/
ቆመ	Bi-radical, 7-1	/*ḱomə*/

Table 3.7፡Root verb of Ge’ez

The two main characteristics of Ge’ez and Tigrigna verb are፡

- How they are written; and
- In a given sentence verbs indicates an action done by subject of a sentences and also it is always agreed with the doer of the action.

As Tadesse Kassa [6] and Mulugeta Atsebeha [57] discussed both language verbs are using affixes [prefix, suffixes, infixes, and circumfix] for inflectional morphology. Affixes are morphemes that are sub words of a word. Based on affixes usage two types of morphemes exist called **Inflectional Morphemes** and **Derivational morphemes**. The one that inflect verbs in number, gender, tense and if the newly formed word class is same as that the first such a morpheme is called **Inflectional Morphemes**. **Derivational morphemes** are responsible not only for the formation of new word but also the word class of the new word also different from that of the previous one. Let us discuss each of the types of affixes in both languages.

3.5.1.4. Adverb /”ተውሳኽ ግስ”

Adverb is a word used to describe the property of a verb.

3.5.1.5. The Stems of verb /”አዕማዲ ግስ”

The Stems of verb pillars or bases of verbs are those that support the conjugations of verbs. Ge’ez and Tigrigna have five stem patterns [60] and all stems have prefixes. These are

- ***Perfective stems ”***ገቢር***”***

e.g., Ge’ez …….. “ቀተለ”

Tigrigna …….. “ቀተለ”

- ***Causative stems “***አገብሮ***”***

e.g., Ge’ez …….. “አቅተለ”

Tigrigna ……..”አቐተለ”

- ***Causative-reciprocal stems*** ”አስተጋብሮ ”

e.g., Ge’ez …….. “አስተቃተለ”

Tigrigna …….. “አቀታተለ”

- ***Reflexive stems “ተገብሮ”***

e.g., Ge’ez …….. “ተቀትለ”

Tigrigna …….. “ተቐተለ”

- ***Reciprocal stems “ተጋብሮ”***

e.g., Ge’ez …….. “ተቃለተ”

Tigrigna …….. “ተቓተለ”

3.5.2. Minor Parts of Speech

3.5.2.1. Pronoun (“ተውላጠ ስም/ክንዲ-ሹም”)

Any word that replaces noun and utilized in the noun place is a pronoun. Pronoun provides the same functionality like that of noun functionality provides. There are here are different types of pronouns these are: personal, reflexive, relative, reciprocal, demonstrative, interrogative, indefinite, and possessive pronoun.

3.5.2.1.1. Personal Pronoun

In Ge’ez and Tigrigna pronouns can be classified as singular and plural, masculine and feminine, and near and far.

	Pronoun		**Gender**			
1st person	**Ge’ez**	**Tigrigna**	**Masculine**	**Feminine**	**Singular**	**Plural**

	አነ	ኣነ	✓	✓	✓	
	ንሕነ	ንሕና	✓	✓		✓
2nd peerson	አንተ	አንታ/ንስኻ	✓		✓	
	አንቲ	አንቲ/ንስኺ		✓	✓	
	አንትሙ	ንስኹም/ንስኻትኩም	✓			✓
	አንትን	ንስኽን/ንስኻትክን		✓		
3rd peerson	ውእቱ	ንሱ	✓		✓	✓ **?**
	ይእቲ	ንሳ		✓	✓	✓ **?**
	ውእቶሙ	ንሳቶም	✓			✓
	ውእቶን	ንሳተን		✓		✓

Table 3.8፡ Tigrigna and Geez pronoun

Pronoun in Ge'ez and Tigrigna can be used being Subject in leading the sentence as singular and plural, near and far, and Masculine and feminine.

Example: (አነ - ንሕነ) አነ = as described at the above table አነ and ንሕነ can be used for both genders().

- አነ ኤፍሬም ሖርኩ ኀበ ቤተ መጻሕፍት / ኣነ ኤፍሬም ናብ ቤተ-መጻሕፍቲ ከይደ። አነ ሖርኩ ኀበ ቤተ መጻሕፍት / ኣነ ናብ ቤተ-መጻሕፍቲ ኸይደ።
- ንሕነ **(**ንሕና**)** = ኤፍሬም ወኤልያስ ሖርነ ኀበ ቤተ መጻሕፍት / ንሕና ኤፍሬምን ኤልያስን ናብ ቤተ-መጻሕፍቲ ኼድና:: ንሕነ ሖርነ ኀበ ቤተ መጻሕፍት / ንሕና ናብ ቤተ-መጻሕፍቲ ኼድና::
- አንተ **(**ኣንታ/ንስኻ**)** = ኤፍሬም ሰተይከ ወይነ / ኤፍሬም ወይኒ ሰቲኻ።

  አንተ ሰተይከ ወይነ / ንስኻ ወይኒ ሰቲኻ።

- አንትሙ **(**እናንተ**)** = ኤፍሬም ወተመስገን ሰተይክሙ ወይነ / ኤፍሬምን ተመስገንን ወይኒ ሰቲኹም። አንትሙ ሰተይክሙ ወይነ / ንስኻትኩም ወይኒ ሰቲኹም።

- አንቲ **(**አንቲ/ንስኺ**)** = አስቴር ሰተይኪ ወይነ / አስቴር ወይኒ ሰቲኺ። አንቲ ሰተይኪ ወይነ / ንስኺ/አንቲ ወይኒ ሰቲኺ።

- አንትን **(**ንስኻትክን**)** = አስቴር ወአልማዝ ሰተይክን ወይነ / አስቴርን አልማዝን ወይኒ ሰቲኽን። አንትን ሰተይክን ወይነ / ንስኻትክን ወይኒ ሰቲኽን።

- ውእቱ **(**እርሱ**)** = ኤፍሬም ሰትየ ወይነ / ኤፍሬም ወይኒ ሰትዩ።

  ውእቱ ሰትየ ወይነ / ንሱ ወይኒ ሰትዩ።

- ውእቶሙ **(**ንሳቶም**)** = ኤፍሬም ወተመስገን ሰትዩ ወይነ / ኤፍሬምን ተመስገንን ወይኒ ሰተዩ/ሰትዮም።

  ውእቶሙ ሰትዩ ወይነ / ንሳቶም ወይኒ ሰትዮም።

- ይእቲ **(**ንሳ**)** = አስቴር ሰትየት ወይነ / አስቴር ወይኒ ሰትያ። ይእቲ ሰትየት ወይነ / ንሳ ወይኒ ሰትያ።

- ውእቶን **(**ንሳተን**)** = አስቴር ወአልማዝ ሰትያ ወይነ / አስቴርን አልማዝን ወይኒ ሰትየን/ሰተያ።

  ውእቶን ሰትያ ወይነ / ንሳተን ወይኒ ሰትየን/ሰተያ።

3.5.2.1.2. Demonstrative Pronoun (“አስተኣማሪ (አመልካቲ/ጠቋሚ) ተውላጠ ስም”)

Demonstrative pronoun is a pronoun that is used to point something specific within a sentence. These pronouns can be used in place of a noun, so long as the noun being replaced can be understood from the pronoun’s context and used before a verb of a sentence. These pronouns can identify either the sentence is Near or Far. These are:

Demonstrative pronoun (Near)

Singular		Plural		Gender	
Ge'ez	**Tigrigna**	**Ge'ez**	**Tigrigna**	**Masculine**	**Femi-nine**
ዝንቱ	እዙይ፣ ነዙይ (this)	እሉ ፣እሎንቱ	እዚአም፣ እዚአቶም (these)	✓	
ዛ	ይቺ፡ ይቺው ? ነዚኣ	እላ	ነዚኣተን	✓	
ዛቲ	እዚኣ ፣ (this)	እሎን	እኒኹ፡እኒኹና ? እዚአን፣እዚአተን		✓
	እዚኣ እያ	እላንቱ	እኒኽ ናቸው ? እዚአተን እየን		

Table3.9.A Demonstrative pronoun (Near) in Ge'ez and Tigrigna

Demonstrative pronoun (Far)					
Singular		**Plural**		**Gender**	
Ge'ez	**Tigrigna**	**Ge'ez**	**Tigrigna**	**Masculine**	**Feminine**
ዝኩ፣ዝክቱ፣ዝስኩ	ንሱ፣እቱይ (that)	እልኩ፣እልክቱ፣እሙንቱ	ንሳቶም፣ ንሶም፣ንሱታት (those)	✓	
እንታከቲ፣እንታክቲ፣እንትኩ	ንሳ፣እቲኣ፣እዚኣ (that)	እልኮን፣እልክቶን	እቲኣተን፣ንሰን፣ነቲኣተን		✓

Table 3.9.B Demonstrative pronoun (Far) in Ge'ez and Tigrigna

Possessive pronoun / "አገናዛቢ ተውላጠ ስም"

Possessive pronoun is a pronoun that takes the place of a noun to show the ownership of someone or something. It can be used instead of a noun phrase to avoid repetition in a sentence.

	Possessive pronoun			
	Singular		Plural	
	Ge'ez	Tigrigna	Ge'ez	Tigrigna
1st person	ዚኣ-የ	ናተይ	ዚኣ-ነ	ናህና
2ndperson	ዚኣ-ከ	ናትካ	ዚኣ-ክሙ	ናትኩም ናትክን
	ዚኣ-ኪ	ናትኪ	ዚኣ-ክን	
3rdperson	ዚኣ-ሁ	ናቱ	ዚኣ-ሆሙ	ናታቶም ናታተን
	ዚኣ-ሃ(ብእሴ)	ናታ	ዚኣ-ሆን(ብእሴ)	

Table 3.10 Possessive pronoun in Tigrigna and Ge'ez

When Ge'ez pronouns are used as verb to be each pronoun express their own meaning as translated into Tigrigna.

Pronoun	The translated meaning of Ge'ez in Tigrigna
ይእቲ	እያ፣ነይራ
ውእቶሙ	እዮም፣ነይሮም፣ነበሩ፣ይንበሩ
ውእቶን	እዮም
አንተ	ኢኻ፣ኔርካ፣ኮይንካ፣ነቢርካ፣ንበር
አንቲ	ኢኺ፣ነይርኪ፣ኮይንኪ፣ንበሪ
አንትሙ	ኢኹም፣ኮይንኩም፣ነይርኩም፣ንበሩ
አንትን	ኢኽን፣ኮይንክን፣ነይርክን፣ንበራ
ንሕነ	ኢና፣ኮይንና፣ኔርና፣ንንበር
አነ	እየ፣ኮንኩ፣ነይረ፣ክነብር

Table 3.11 Translated meaning of Pronouns from Ge'ez to Tigrigna

3.5.2.2. Conjunction “መስተፃምር”

As Tadesse [6] discussed about conjunction: Conjunction is a word used to connect clauses or sentences or to coordinate words in the same clause. In Ge’ez “ወ ፤ አው ፤ and ዳዕሙ ፣ አላ ፣ ባሕቱ” and in “Tigrigna ን ፣ ወይ ፣ and ነገር ግን” are conjunction used. ወ in Ge’ez has 27 meaning. The most commonly used meaning of “ወ” used as “ን”.

3.5.2.3. Punctuation Mark

In Ge’ez there is no question mark whereas Tigrigna has. The interrogative is placed at the end of the sentences. It is pronounced with a low level and the style of pronunciation by itself also shows an interrogation. In most cases, Ge’ez interrogatives are preceded by a radical which has the same order to the interrogative. These two languages have the same punctuation mark except question mark as we explained at the above. For example,”ሁ፣ ኑ፣ ኡ፣ ኢ፣ ት፣ ኣ፣ ኣይት? ሶበኑ?” (When?) “ተአምሩኑ?” (Do you know?) ፣ “አንትሙሁ” (are you?) ፣ “ተአምረኒኢ” (do you know me?).

3.6. Morphology

Morphologically, languages are often characterized along two dimensions of variation. The first is the number of morphemes per word, ranging from isolating languages in which each word generally has one morpheme, to polysynthetic languages in which a single word may have very many morphemes. The second dimension is the degree to which morphemes are segment able, ranging from agglutinative languages. Ge’ez and Tigrigna exhibit such character that the performance of the SMT system difficult. Inflectional morphemes include the grammatical functions of the word. These are number, tense/aspects, possession and comparison []. Number: - Ge’ez and Tigrigna has singular and plural numbers. The number marker in Ge’ez and Tigrigna language usually exists noun, adjectives, and verb conjunctions. It exists in either of prefix, infix, suffix and super-fix. The number markers in pronouns, demonstratives, prepositions are the same but numbers in nouns are complex with exception of every conjunction. In Ge’ez, -yan, -an, yat, and -at are suffix plural number marks in Ge’ez. Gender: -in Ge’ez the gender markers are not limited. They may vary from time to time accordingly to the part of speech. The gender markers are the feminine markers. Gender is distinguishable in both singular and plural. Gender is nouns,

adjectives, some adverbs, prepositions, demonstratives, possessive, verbs are marked by the following "-ኦች-" at plural, "-ተ -" as person profile as personal suffix, "-ን -" in pronoun plural, -ኢ" in pronoun possessive, and aspect… "ሃ" - as objective markers in personal names in possession preposition. ኣ -in gerund, infinitive and derivational morphemes

3.7. Challenges of Ge'ez and Tigrigna During Machine Translation

There are different challenges that we noticed when trying to do machine translation between Ge'ez and Tigrigna language. Some of the challenges are described below: -

- **Morphological challenges**

Translating between two morphologically rich languages poses challenges in analysis, transfer and generation. The complex morphology induces an inherent data scarcity problem, and the limitation imposed by the dearth of available parallel corpora is magnified. Both Ge'ez and Tigrigna are ploy syntactic languages which is the number of morphemes per word is not always one. Most of the researches conducted in SMT are using morphologically rich language as a source language and target language is morphologically poor. Nevertheless, both Ge'ez and Tigrigna, which have rich language morphemes, are used interchangeably in the context of bidirectional morpheme-based machine translation [6, 57, and 61].

- **Syntactical challenges**

Syntactically, both Ge'ez and Tigrigna languages are perhaps most saliently different in the basic word order of verbs, subjects, and objects in simple declarative clauses. The syntactic structure of Tigrigna is generally SOV whereas Geez follows SVO, VSO and OVS word order for declarative sentences. This makes the translation most challenging [60].

- **Alignment challenge**

In the case of conducting bidirectional statistical machine translation, two morphologically rich languages, Ge'ez and Tigrigna Languages, there exist critical alignment challenge due to the variation of alignments between the languages. That is, in some sentences there could be one to one, one to many or many to one or many to many [62].

3.8. Number system

Geez has its own numerals for designating numbers [58]. Tigrigna number system uses Ge”ez numbering systems. It has twenty characters. They represent numbers from one to ten “፩-፲”, twenty to ninety “፳-፺”, hundred “፻” and thousand “፼”. However, these are not suitable for arithmetic computation purposes because there is no representation for zero (0), decimal points. Tigrigna numbering system can be classified as ordinal numbers, cardinal numbers and fraction. The cardinal numbers are numbers like “ሓደ (one), “ክልተ (two), ሰለስተ (three), “ዓሰርተ” (ten), etc…, the ordinal numbers are “ቀዳማይ” (first), ካልኣይ (second), “ሳልሳይ” (third), “ዓስራይ” (tenth), etc.… and fraction numbers are also special numerals in Tigrigna that correspond to the English like:

“ፍርቂ” (half), “ርብዒ” (quarter) “ሲሶ” (one-third) etc.

CHAPTER FOUR

4. METHDOLOGY

4.1. Introduction

A research methodology is a way to systematically solve the research problem [62]. In this section, the procedure for Ge'ez to Tigrigna morpheme based bi-directional machine translation is presented. Here included are the corpus preparation, data description, methods, procedures, the model and the evaluation techniques which are presented respectively as follows.

4.2. The Methods

A research method is the procedures to be undertaken that involve the forms of data collection, analysis, and interpretation that researcher proposes for the study [63]. The method followed is morpheme based bi-directional machine translation in case of Ge'ez to Tigrigna and vis-versa. In this study statistical machine translation approach was used.

4.3. Data Description

The dataset is composed of 9 books of Bibles, which consist of 384 chapters for each Ge'ez and Tigrigna Languages. The Bible Books are Genesis, Exodus, Leviticus, Numbers, Deuteronomy, Judge, Joshua, Ruth, and Psalms. The 9189 verses of the 384 chapters of the nine books were used for the experimentation purpose.

4.3.1. Corpus Preparation

A parallel corpus was collected from https://www.stepbible.org/ digitally available Bible in Ge'ez and Tigrigna. To train the Morfessor, 12173 Ge'ez and 16708 Tigrigna words were taken. The corpus dataset was divided into two parts: train and test, with 7290 verses for training and 1899 verses for testing for both languages. These data were used to develop the model. For the translation purpose, the toolkits such IRSTLM was used for language model and MGIZA++ for word and morpheme alignment.

Algorithm for collecting Corpus by using web scrapping:

S. N.	Books of the Bibles	Nº. of Chapters	Nº. of Verses
1.	Genesis	50	1527
2.	Exodus	38	1221
3.	Leviticus	27	806
4.	Numbers	36	1279
5.	Deuteronomy	34	780
6.	Joshua	24	555
7.	Judges	21	574
8.	Ruth	4	85
9.	Psalms	150	2362
Total		**384**	**9189**

Table 4.1: Books of the Bibles and their respective Chapters used as dataset

Web scrapping : Extracting geez bible corpus from website

```
import numpy as np
import requests
from bs4 import  BeautifulSoup as bs

import timeit

start = timeit.default_timer()

next_links'/?q=version=Geez|reference=Gen.1'
geez_bible=[]

for i in range(236):
    req = requests.get("https://www.stepbible.org/"+next_link)
    soup = bs(req.content,"html.parser")

    # get next link
    links = soup.select('a')
    for i in range(len(links)):
        link = soup.select('a')[i]
        if link.get('class') == ['nextChapter']:
            next_link = link.get('href')

            # extract body
            body=soup.find('div', class_='passageContent')

            # find chapter title in body
            chapter_title = body.find('h2', class_='xgen')
            # append chapter title and body text to our list
            geez_bible.append(chapter_title.text)
            geez_bible.append(body.text)

stop = timeit.default_timer()

print('Time taken to scrap: ', stop - start)
```

4.4. Language Model

A language model uses machine learning to conduct a probability distribution over words used to predict the most likely next word in a sentence based on the previous entry. Language models

learn from text and can be used for producing original text, predicting the next word in a text, speech recognition, optical character recognition and handwriting recognition. As described at section 4.3.1 to perform the training and testing procedures. From all corpus 80% which is 7290 of both Ge'ez and Tigrigna verses were used to train the model.

For Example when translating morpheme based bi-directional a Tigrigna Bible verse phrase "ወይቤሎን" into Geez, the translator can give several choices as output:

Most likely Tigrigna translations for geez word ወይቤሎን after1 iteration:

['ገበርክን፡', 'ኻልኣይቲ', 'እትሐድጋኦምሲ፡', 'ኣንስቲ', 'ብህይወት', 'ከሎኽን፡', 'ጹዓ', 'እብራውያን', 'ኼነ', 'ዝ ስማ፡']

Most likely Tigrigna translations for geez word ወይቤሎን after 5 iterations:

['ጸዊዑ፡', 'ዝስማ፡', 'ብህይወታ', 'ከተሕርሳኤን', 'እብራውያን፡', 'ጓል', 'ኼነት', 'ሺፍራ',

'እትሐድጋኦምሲ፡', 'ጀመሩ።'] Most likely Tigrigna translations for geez word ወይቤሎን after

10 iterations: ['ገበርክን፡', 'ርኣያ፡', 'ጹዓ', 'ሺፍራ', 'ቅተላኡ፡', 'ነተን', 'ዝስማ፡', 'በለን።', 'ከተሕርሳኤን',

'ብህይወታ'] Most likely Tigrigna translations for geez word ወይቤሎን after 20 iterations:

['ነተን', 'ሺፍራ', 'ጸዊዑ፡', 'ኼነት', 'ገበርክን፡', 'ከተሕርሳኤን', 'እብራውያን፡', 'ብስም', 'ተዛረበን፡', 'ከሎኽን ፡']

Here, the language model tells that the translation "ነተን" sounds natural and will suggest the same as output.

4.5. Translation Model

Translation models describe the mathematical relationship between two or more languages. We call them models of translational equivalence because the main thing that they aim to predict is whether expressions in different languages have equivalent meanings [64].

4.5.1. 4.5.1 Decoder

A decoder searches for the best sequence of transformations that translates input (source) sentence to the corresponding output (target) sentence. It looks up all translations of every source word or phrase, using word or phrase translation table and recombine the target language phrases that maximizes the translation model probability multiplied by the language model probability [9][34]. By following the above procedure, the decoder performs the translation process from both directions.

- **GIZA ++ Tool**

GIZA++ is a statistical machine translation toolkit that is used to train IBM Models 1-5 and an HMM word alignment model. This package also contains the source for the mkcls tool which generates the word classes necessary for training some of the alignment models.

- **Morfessor Tool**

Is a family of probabilistic machine learning methods for finding the morphological segmentation from raw text data that were indicated in sub-unit 4.3.1, Corpus preparation.

- **The Viterbi Algorithm**

The Viterbi Algorithm is a dynamic programming solution for finding the most probable hidden state sequence [65]. If there are G and T is the number of observations in the sequence P(G|T) to P(G, T)/P(T) can be transformed, but there is no need in finding P(T) as P(T) does not pertain to changes in state sequences.

P(G,T) = P(T|G)P(G)

$$=\mathrm{P}(T_1 \dots T_t|G_1 \dots G_t)\prod_{i=1}^{t} P(_{Gi|G_1 \dots G_{i-1}})$$

$$=\mathrm{P}(T_1 \dots T_t|G_1 \dots G_t)\prod_{i=1}^{t} P(_{Gi|G_{i-1}})$$

$$=\prod_{i=1}^{t} p(T_i\,|T_1 \dots T_{i-1}, G_1 \dots G_{i-1})\prod_{i=1}^{t} P(_{Gi|G_{i-1}})$$

$=\prod_{i=1}^{t} p(T_i\,|G_i|P(_{Gi|G_{i-1}})$, where t is the number of observations in the sequence.

4.6. Evaluation

The final output of the translation systems needs to be evaluated. The evaluation is made by comparing the translations of a set of sentences (output of the system) to the correct translations. As it was discussed in section 2.5, we can evaluate machine translation systems using human evaluation and automatic evaluation, but human evaluation is expensive, too slow, and subjective, therefore automatic evaluation is reliable. BLUE score is one of the popular automatic evaluation systems and which is standard for automatic machine translation evaluation and it is a precision oriented metric in that it measures how much of the system output is correct.

CHAPTER FIVE

5. DESIGN, IMPLEMENTATIONN AND EXPERIMET RESULT

To perform the experiment, we design architecture of system design and apply morphological segmentation, construct language and translation model.

5.1. System Design Architecture

The system design and architecture are framed for the implementation of morpheme based on bidirectional Geez to Tigrigna machine translation are presented as follows

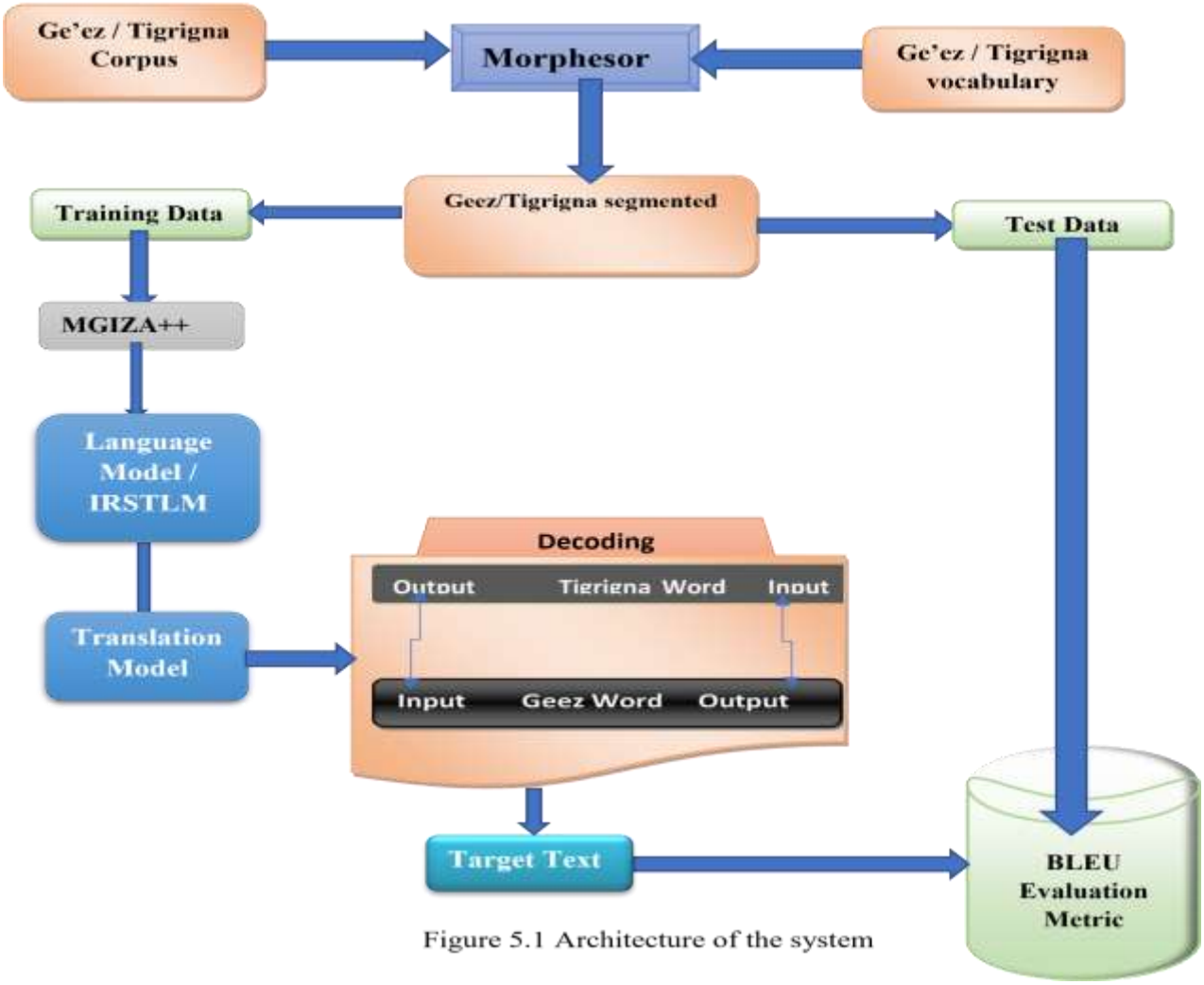

Figure 5.1 Architecture of the system

5.2. Implementation

The implementation process is presented as follows.

5.2.1. Preprocessing

Preprocessing began with the removal of any unnecessary or unusual characters from the sentence. Then total the frequency of the ten most frequent words in both Tigrinya and Ge'ez.

```
Most common Tigrignawords: [('ድማ', 166), ('ኣብ', 147), ('ከኣ', 105), ('ምድሪ', 102), ('ኣ
ምላኽ', 80), ('ካብ', 69), ('እቲ', 49), ('ኽኣ', 45), ('እግዚኣብሄር', 44), ('ከሉ', 43)]
Most common Geez words: [('፡', 3130), ('።', 260), ('እግዚኣብሔር', 106), ('ምድር', 91), ('ከ
ሙ', 49), ('ከሉ', 46), ('ውስተ', 44), ('ወከነ', 42), ('እምነ', 36), ('፡]', 31)]
```

5.2.2. Training the model

We have used the data set described in Section 4.3.1 to perform the training and testing procedures. From corpus 80% which is 7290 of both Ge'ez and Tigrigna verses were used to train the model.

5.2.3. Training the system

The training process includes creating language model, translation model, and conducting decoding using the help of GIZA++ and IRSTLM. The created language model is built with the target language model, that is, for Tigrigna as well as Ge'ez separately; both the languages become a target and source language at some point.

5.2.4. Tokenizer and Frequency

Tokenization is the process of breaking a stream of text up into words, phrases, symbols. The Tigrigna and Ge'ez corpora becomes input to this tokenizer and frequency calculator component, and the component generates list of words with their frequency of occurrence. Some noise like, Punctuation, digits and whitespaces are not included in the resulting list of tokens. The list of tokens becomes input to the segmentation learner for further processing.

```
from collections import Counter

# Clean strings from special characters
tigrigna_str = remove_noise(tigrigna_str)
geez_str = remove_noise(geez_str)

# Count word frequency
tig_word_counter = Counter(tigrigna_str.split())
ge_word_coutner = Counter(geez_str.split())

# 10 most common words
print(f'Most common Tigrigna words: {tig_word_counter.most_common(10)}')
print(f'Most common Geez words: {ge_word_coutner.most_common(10)}')
```

```
Most common Tigrigna words: [('ንዝናጥን', 1), ('ብምሳሩ', 1), ('መምሬ', 1), ('ቀቕጣጠ', 1), ('ንግበር', 1), ('እተብህሃ', 1), ('ሸሸብዓት', 1), ('ኩ
ላ', 1), ('ብልዛ', 1), ('ተባህለ', 1)]
Most common Geez words: [('ኣልኩ', 1), ('ጠሐለየ', 1), ('ወሰመ', 1), ('ንግበር', 1), ('እምነበ', 1), ('ወእትጋሠሥ', 1), ('ይኩን', 1), ('መጎና', 1),
('ለመፍሙ', 1), ('ባሕቱ', 1)]
```

Figure 5.2: Frequency

5.2.5. Segmentation with Morfessor

To conduct this experiment, we need to segment the words in each of the language based on unsupervised segmentation tool called Morfessor. This unsupervised segmentation system learns the segmentation from a list of corpora with their frequency that are not annotated or pre-processed in any way that helps the system to predict the correct segmentation. The output from this segmentation learner is morph segmentations of the corpus in the input. This morph segmentation output becomes a segmentation model that can be used as a model to segment Tigrigna and vice versa.

In this segmentation model of the Morfessor program no model learning takes place. Each input is segmented into morphs by the Viterbi algorithm, which finds the most likely segmentation of the word into a sequence of morphs.

```
def gezSegment(seg):
    segmented = [geez_model.viterbi_segment(i)[0] for i in seg.split()]
    result = [j for i in segmented for j in i]
    return result
```

```
geez_seg = [" ".join([str(item) for item in gezSegment(i)]) for i in geezCleaned ]
geez_seg
```

```
['Genesis 1',
 '1 በ ቀዳሚ ፡ ፣ ብ ረ ፡ እግዚኣብሔር ፡ ሰ ማ የ ፡ ወ ም ድ ረ ።',
 '2 ወ ምድር ሰ ፡ ኢ ታስተርኢ ፡ ወ እ ኮነ ት ፡ ድሉተ ፡ ወ ጽልሞት ፡ ሞል ዕ ል ተ ፡ ቀላይ ፡ መሞጌለ ፡ እግዚኣብሔር ፡ ይሕልል ፡ ሞል ዕ ል ተ ፡ ማይ ።',
 '3 ወ ይቤ ፡ እግዚኣብሔር ፡ ለይ ኩ ን ፡ ብርሃን ፡ ወ ኮነ ፡ ብርሃን ።',
 '4 ወ ር እ ዮ ፡ እግዚኣብሔር ፡ ለ ብርሃን ፡ ከሞ ፡ ፡ናይ ፡ ወ ፈ ለ ጠ ፡ እግዚኣብሔር ፡ ማክለ ፡ ብርሃን ፡ ወ ማክለ ፡ ጽልሞት ።',
 '5 ወ ሰ ሞ ዮ ፡ እግዚኣብሔር ፡ ሰ ብርሃን ፡ ዕ ለ ተ ፡ ወ ለ ጽልሞት ፡ ሌ ሊታ ፡ ወ ኮነ ፡ ሌ ሊታ ፡ ወ ጸ ብ ሐ ፡ ወ ኮነ ፡ መ ዓ ል ተ ፡ ፩ ።',
 '6 ወ ይቤ ፡ እግዚኣብሔር ፡ ለይ ኩ ን ፡ ጠለለ ፡ ማክለ ፡ ማይ ፡ ከሞ ፡ ይ ፍ ል ተ ፡ ማክለ ፡ ማይ ፡ ወ ኮነ ፡ ከሞኡ ።',
 '7 ወገብ ረ ፡ እግዚኣብሔር ፡ ጠ ለረ ፡ ወ ፈ ለ ጠ ፡ እግዚኣብሔር ፡ ማ እ ብ ለ ፡ ማይ ፡ ዘ ታሕተ ፡ ጠለረ ፡ ወ ማክለ ፡ ማይ ፡ ዘ ማል ዕ ል ተ ፡ ጠለረ ።',
 '8 ወ ሰ ሞ ዮ ፡ እግዚኣብሔር ፡ ለ ሙሉቱ ፡ ጠለረ ፡ ሰ ማ የ ፡ ወ ር እ የ ፡ እግዚኣብሔር ፡ ከሞ ፡ ሠናይ ፡ ወ ኮነ ፡ ሌ ሊታ ፡ ወ ጸ ብ ሐ ፡ ወ ኮነ ፡ ካልኣ ተ ፡ ዕ ለ ተ ።',
 '9 ወ ይቤ ፡ እግዚኣብሔር ፡ ለይ ት ጋ ባእ ፡ ማይ ፡ ዘ ማክለ ፡ ለማይ ፡ ውስተ ፡ አሐዱ ፡ መካን ፡ መሰከርኢ ፡ የ ብ ስ ፡ ወ ኮነ ፡ ከሞሁ ፡ ው ተጋብኢ ፡ ማይ ፡ ውስተ ፡ ምስላዊሁ ፡ ወ አስ ተርኢየ
```

```
def tigSegment(seg):
    segmented = [Tigrigna_model.viterbi_segment(i)[0] for i in seg.split()]
    result = [j for i in segmented for j in i]
    return result
```

```
tig_seg = [" ".join([str(item) for item in tigSegment(i)]) for i in tigrignaCleaned ]

tig_seg
```

```
['ኣራት ዘ ፍጥረት ምዕራፍ 1',
 '1 ኣምላኽ ብ መ ጀ መር ታ ሰማይ ን ም ድር ን ፈጠረ',
 '2 ምድሪ ድማ ባረኻን ጥ ራ ያ ን ነበረ ት ጸል ማ ት ከኣ ኣብ ል ዕሊ መኣሙቐ ነበረ መንፈስ ኣምላኽ ድማ ኣብ ል ዕሊ ማ ያ ት ይ ዝ ም ቢ ነበረ',
 '3 ኣምላኽ ከኣ ብርሃን ይኹን በለ ብርሃን ድማ ኾነ',
 '4 ኣምላኽ ድማ እ ቲ ብርሃን ጽቡቕ ከም ዝኾነ ረአየ ኣምላኽ ከኣ ን ቲ ብርሃን ካብ ጸል ማ ት ፈ ለ ዮ',
 '5 ኣምላኽ ን ቲ ብርሃን መዓልቲ ኣውጽኡ ን ቲ ጸል ማ ት ከኣ ለይቲ ኣውጽኡ ም ሸ ት ኮነ ብጌጥኡን ኮነ ሓንቲ መዓልቲ',
 '6 ኣምላኽ ድማ ን ማ ያ ት ካብ ማ ያ ት ጥ ፈ ሊ ጠፈር ኣብ መንፅ ማ ያ ት ይኹን በለ',
 '7 ኣምላኽ ን ቲ ጠፈር ገበ ን ቲ ኣብ ትሕቲ ጠፈር ዘ ሎ ማ ያ ት ድማ ኸብቲ ኣብ ል ዕሊ ጠፈር ዘ ሎ ማ ያ ት ፈ ለ ዮ ከምኡ ውን ኮነ',
 '8 ኣምላኽ ከኣ ን ቲ ጠፈር ሰማይ ኣውጽኡ ም ሸ ት ኮነ ብጌጥኡን ኮነ ካልኣይ ቲ መዓልቲ',
 '9 ኣምላኽ ድማ እ ቲ ንቹጽ ምእንቲ ኼርኤስ እ ቲ ኣብ ትሕቲ ሰማይ ዘ ሎ ማ ያ ት ናብ ሓንቲ ቦታ ይተኣክብ በለ ከምኡ ድማ ኾነ',
 '10 ኣምላኽ ከኣ ን ቲ ንቹጽ ምድሪ ኣውጽኡ ን ቲ እኩብ ማ ያ ት ድማ ባሕሪ ኣውጽኡ ኣምላኽ ከኣ ጽቡቕ ከም ዝኾነ ረአየ',
 '11 ኣምላኽ ድማ ኣታ ምድሪ ሳዕ ር ን ዘርኢ ዘሀ ብ ብቑልን ዘርኢ ኣብ ርእሱ ዘለዎ ፍረ ከ ከም ዓይነቱ ኣብ ምድሪ ዘ ፈ ሪ ኦም ተው ጽ እ በለ ከምኡ ድማ ኾነ',
```

Figure 5.4: Morphological segmentation result

5.3. Experiment Result

This section presents experimental results of morpheme based bidirectional Ge'ez-Tigrigna statistical machine translation.

- **Morpheme based translation from Tigrigna to Ge'ez**

For this experiment, Tigrigna is the source language and Ge'ez is the target language. (10792 sample Ge'ez and 10792 Tigrigna were taken)

```
1 ድ ም ጸ ይ ናብ ኣምላኽ ኣልዒለ ኤ እ ዊ ድ ም ጸ ይ ናብ ኣምላኽ ኤ ል ዕ ል ኣሎኹ ንሱ ውን ይሰምዓኒ እዩ
2 ብ መዓልቲ ጸበባ ይ ንእግዚኣብሄር ደሌኽዎ ብ ለይቲ ኢ ደ ይ ተ ዘርግሔ ት ኣይሰልከየትን ከኣ ነፍሰ ይ ምጽንናዕ ኣበየት
3 ን ኣምላኽ እዝክሮን እ ቕ ዝ ም ን ኣሎኹ ኤ ስ ተን ት ንን መን ፈ ሰ ይ ከኣ ይሕለልን ኣሎ
4 ኦዒንተይ ቄላሕ ላሕ ተ ብ ለን እሽ ገር ምዝራብ ውን እስ እን ኣሎኹ
5 ናይ ቀደም መዓል ት ታት ናይ ጥንቲ ዓመታ ት እ ዝክር ኣሎኹ
6 ብ ለይቲ ን በገና ይ እሐስብ ምስ ል በ ይ ኤ ስ ተን ት ን ኣሎኹ መን ፈ ሰ ይ ውን ይ ምርምር ኣሎ
7 እግዚኣብሄርሲ ንዘለኣለምዶ ይ ድር ቢ ድሕርዚኽ ጸጋኡዶ ኣየርእን እዩ
8 ጸጋ ኡስ ንዘለኣለምዶ ፈጺሙ ተወዲኡ ተስፋ ኡኽ ን ውሉድ ወለዶ ዶ ተሪፉ እዩ
9 ኣምላኽሲ ሳህሉዶ ርሲዕዎ ን ም ሕ ረ ቱስ ብኽሪራዶ ዐጽይዎ እዩ
10 ኣነ በል ኩ እዚ እዩ ድኻ መ ይ ዓመታ ት የ ማ ነ ይቲ እ ቲ ልዑል እ ዝክር ኣሎኹ
11 ናይ ቀደም ተ ኣም ራ ት ካ ኽሐስብ እዩ እሞ ን ግ ብር ታት እግዚኣብሄር ክዝክሮ እዩ
12 ን ብ ዘ ሎ ግ ብር ታት ካ ድማ ከስ ተን ት ኖ ንዕዮኻውን ክምርምሮ እዩ
13 ዎ ኣም ላ ኽ መ ን ድኻ ብቕድስና እዩ ከም ኣም ላ ኽ ዝ በለ ዓብዪ ኣም ላ ኽ ከ መን ኣሎ
14 ተ ኣም ራ ት እ ት ገብር ኣም ላ ኽ ንስኻ ኢኻ ንሓይልኻ ኣብ ማእከል ኣህዛብ ኣፍሊጥካዮ
```

Fig 5.4.A. Sample morpheme-based Translation intput from Tigrigna to Ge'ez(Ge'ez)

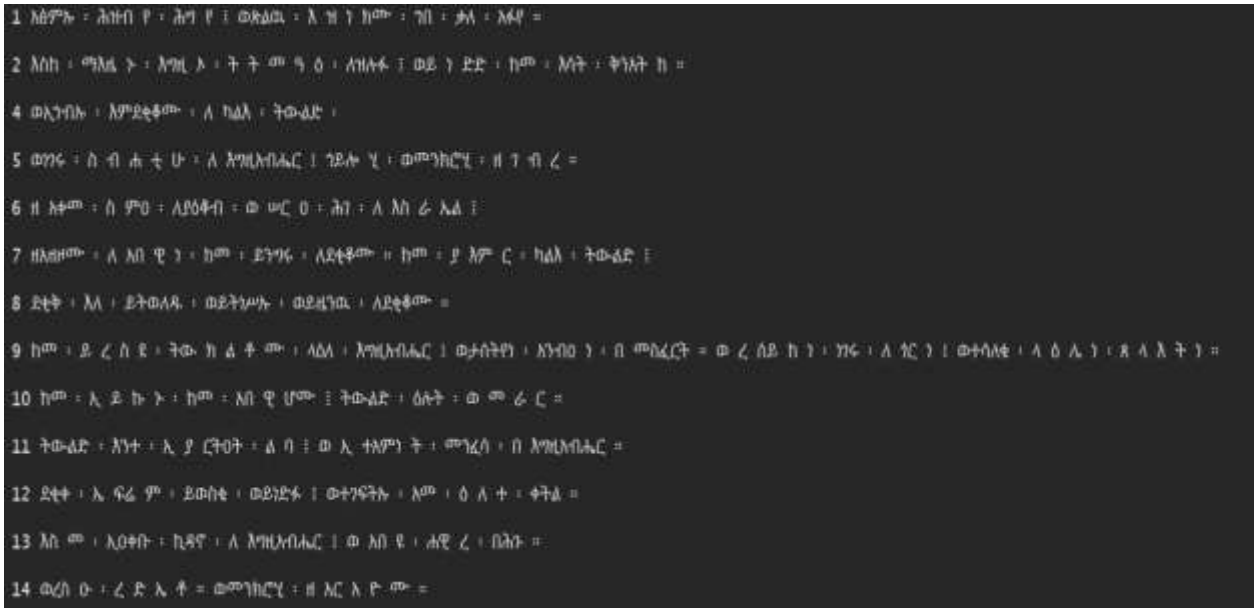

Figure 5.4.B: Sample morpheme-based Translation output from Tigrigna to Ge’ez (Ge’ez)

Result of the translation

```
from nltk.translate.bleu_score import sentence_bleu,corpus_bleu

with open ('dataset/test.ti','r',encoding='utf-8') as g:
    ref = g.readlines()
with open ('dataset/translatedd.ti','r',encoding='utf-8') as f:
    can = f.readlines()
```

```
reference = [j for i in range (len(ref)) for j in ref[i].split()]
candidate = [j for i in range (len(can)) for j in can[i].split()]
```

```
score = sentence_bleu(reference, candidate)
```

```
print('BLEU score -> {}'.format(round(score,2)))
```

```
BLEU score -> 9.23
```

Figure: 5.5 Evaluation of the translation from Tigrigna to Ge’ez

To assess the system's performance in terms of translation accuracy for a single Tigrigna to Ge'ez sentence, 1899 Tigrigna and 1899 Ge'ez verses were employed. In order to do this, the BLEU score technique is employed to determine how well the translation process worked. According to the BLEU score methodology's results, 9.23 percent of the translations from Ge'ez to Tigrigna were done correctly.

➢ **Morpheme based translation from Ge’ez to Tigrigna**

For this experiment we use, Ge’ez language as an input for source language and Tigrigna is target language.

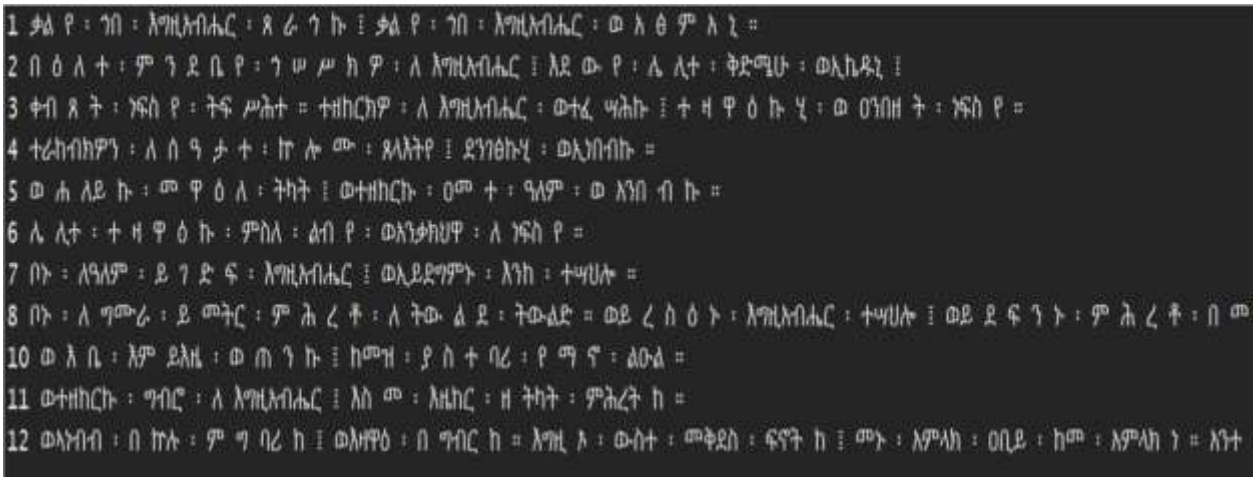

Figure 5. 5.A.: Sample morpheme-based Translation input from Ge’ez, to Tigrigna (Ge’ez)

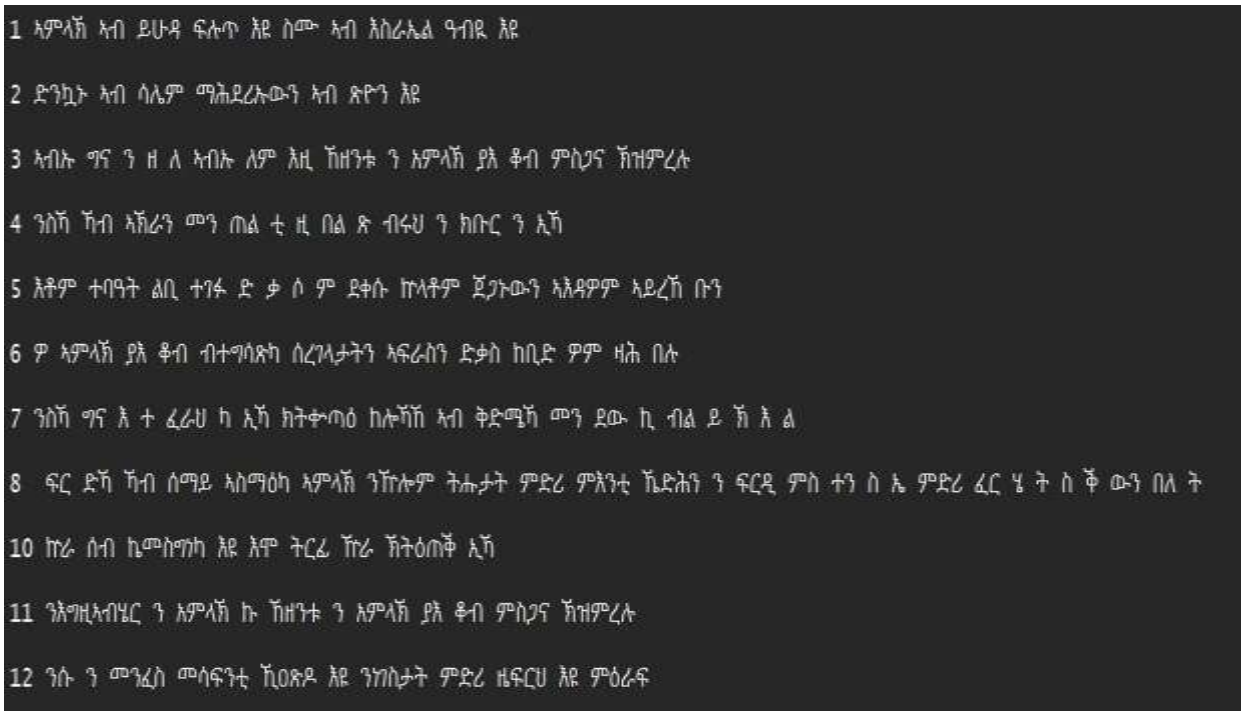

Figure 5. 5. B.: Sample morpheme-based Translation input from Ge’ez, to Tigrigna (Tigrigna)

Result of the translation

```
from nltk.translate.bleu_score import sentence_bleu,corpus_bleu

with open ('testt.ge','r',encoding='utf-8') as g:
    ref = g.readlines()
with open ('translatedd.ge','r',encoding='utf-8') as f:
    can = f.readlines()
```

```
reference = [j for i in range (len(ref)) for j in ref[i].split()]
candidate = [j for i in range (len(can)) for j in can[i].split()]
```

```
score = sentence_bleu(reference, candidate)
```

```
print('BLEU score -> {}'.format(round(score,2)))
```

```
BLEU score -> 8.67
```

To assess the system's performance in terms of translation accuracy for a single Ge'ez to Tigrigna sentence, 1899 Ge'ez and 1899 Tigrigna verses were employed. In order to do this, the BLEU score technique is employed to determine how well the translation process worked. According to the BLEU score methodology's results, 8.67 percent of the translations from Ge'ez to Tigrigna were done correctly.

Findings

BLUE score is one of the popular automatic evaluation systems and which is standard for automatic machine translation evaluation and it is a precision-oriented metric in that it measures how much of the system output is correct. The results are shown below:

Experiment Conducted	**Result of experiment in BLUE From both Direction**	
Morphem Based Translation	**Ge'ez to Tigrigna**	**Tigrigna to Ge'ez**
	8.67	**9.23**

Table BLUE Score evaluation results:

Here the BLUE evaluation results for both the Languages are discussed.

Generally, regarding the relevance of the summary outputs, the Tigrigna to Ge'ez translation output is BLUE Score = 9.23 and Ge'ez to Tigrigna translation output is BLUE Score = 8.67. Morphological richness of the two languages requires lack of standard corpus especially for machine learning algorithms, both languages are perhaps most saliently different in the basic word order of verbs, subjects, and objects in simple declarative clauses this made the translation difficult and also In these Languages, there exist critical alignment challenge due to the variation of alignments between the languages. According to this we get poor evaluation score.

Chapter SIX

6. CONCLUSIONS AND RECOMMENDATIONS

6.1. Conclusions

Morphologically rich languages like Ge'ez and Tigrigna pose a challenge for statistical machine translation, as these languages possess a large set of morphological features producing many rich surface forms. Morphologically complex languages are well known to cause problems for contemporary statistical machine translation (SMT) systems. This is because of a single word consists of one or more sub-words called morpheme. Therefore, this study aimed to explore an optimal translation unit for Ge'ez- Tigrigna bi-directional translation. To achieve this goal, the first researcher studied the morphology and syntax of both Geez and Tigrigna language. Accordingly, it was identified that both languages have equivalent morphological richness and Ge'ez is a free grammar language regarding the syntax being SVO, VSO, or VOS. The position of the adverb and adjectives also in Geez is any place before or after a verb and a noun respectively. The design process of bidirectional Geez-Tigrigna machine translation involved the collection of Ge'ez and Tigrigna parallel corpus. The corpus collected from freely available online sources such as Old Testament Holly Bible and SQLite digital database. Corpus preparation involved activities of preprocessing the corpus such as tokenization and character normalization. Morfessor and morphological rules were used to segment morpheme of Ge'ez and Tigrigna respectively. And they were used to find morpheme of Geez and Tigrigna. MGIZA++ used for word and morpheme level alignment. Moses was used for translation process which integrates all necessary tools for machine translation such as IRSTLM, MGIZA++ and decoder. To identify an optimal translation unit, different experiment on each translation unit called word and morpheme were conducted. Based on unsupervised morpheme segmentation using morfessor the study creates morpheme-based datasets which achieved 9.23 % from Tigrigna to Ge'ez and 8.67% Ge'ez to Tigrigna BLEU score respectively. These results showed that the identified morpheme was an optimal unit of translation and it enhanced the performance of bi-directional Ge'ez-Tigrigna machine translation and vice versa. However, being conducting machine translation between morphologically rich languages, there are a number challenges observed. One of the challenges was alignment challenge due to the multiple syntactic order used in Geez

writing system. In addition, handling morphological richness of the two languages requires lack of standard corpus especially for machine learning algorithms.

6.2. Recommendation

Bidirectional statistical machine translation of corpus-based approach was used. It trained and translated the corpus prepare for the purpose. Based on the aforementioned conclusions, the following recommendations were forwarded. In our study we focus only on morpheme as a translation unit, further research can be done on other unit of translation like phrase, sentence and word.

- The corpus used for this study was solely collected from the Holly Bible books, chapters and verses. To prove the current results, it is essential to undertake further ample corpus from different disciplines.
- To exploit the strength of the two major machine learning approaches, further research needs to be conducted on Ge'ez and Tigrigna using Neural machine translation.
- Better results could be obtained by increasing the size and domain of the data set used for training the system.
- To minimize the prevalent challenge in preparing the corpus and then to develop a full-fledged bidirectional Ge'ez -Tigrigna machine translation there is a need to increasing the size of the data set for validation and integrating the linguistic information is of paramount importance.
- Finally, it could be recommended that the professionals in the fields of language, in the case of this study, need to avail pre-prepared standard corpus file to bypass the challenges imposed by the complexities inherent in the morphology of the languages.

REFERENCES

[1] J. Daniel and H. M. James, *An introduction to natural language processing, computational linguistics, and speech recognition*, United States of America: Prentice-Hall Inc. 2000.

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

[3] M. Caitlin, “Man VS Machine Interpretation the Ambiguous in Diplomatic Negotiations” Syracuse University Honor Program Capstone Project, 381, 2010

[4] A. Douglas, B. Lorna, M. Siety, H. R. Lee and S. Louisa, *Machine Translation An Introductory Guide.* London: NCC Blackwell Ltd, pp.234, 1994

[5] L. Adam and P. Matt, "Beyond bitext: Five open problems in machine translation," Human Language Technology Center of Excellence Johns Hopkins University, 2013.

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

[7] M. Maučec and G. Donaj, “Morphology in statistical machine translation from English to highly inflectional language.” *Information Technology and Control*, vol. 47 (1): 63-74, 2018 [8] R. M. Steven and M. R. Gary, "Experimental Research Method," in Experimental Research Methods. Memphis: Wayane, 2003, p. 25.

[9] M. Bulakh, and L. Kogan, “The Genealogical Position of Tigre and the Problem of North Ethio- Semitic Unity.” *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 160 (2010): 273–302

[10] C. R. Kothari. *Research Methodology: Methods and Techniques*, New Delhi: New Age International (P) Limited, 2nd revised edition, 2004

[11] Catherine Dawson. A Practical Guide to Research Methods, U K: Oxford OX5 2DQ, 1st edition, 2007

[12] M. R. Janet. Essentials of Research Methods, U K: Oxford Ox4 IRX, 3rd revised edition, 2007

[13] Biruk Abel, “Geez to Amharic Machine Translation”, MSc Thesis, submitted to Addis Ababa University, Addis Ababa, Ethiopia, 2018

[14] Dawit Mulugeta **"**Geez to Amharic Automatic Machine Translation: A Statistical Approach" MSc Thesis, submitted to Addis Ababa University, Addis Ababa, Ethiopia, 2015

[15] Akubazgi Gebremariam, "Amharic-to-Tigrigna Machine Translation Using Hybrid Approach", MSc Thesis submitted to Addis Ababa University, Addis Ababa, Ethiopia, 2017

[16] M. C. Amine, "Theoretical Overview of Machine Translation," in *Proceedings ICWIT*, African University, Adrar, Algeria, 2012

[17] A. Clark, C. Fox, and S. Lappin, *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons, 2010

[18] Jabesa Daba and Yaregal Assabie, "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, Warsaw, Poland, 2014 pp. 228-235.

[19] SYSTRAN, [Online], Available: http://www.systransoft.com/systran/corporate-profile/translation technology/ *what-is-machine-translation*/, last visited March 7, 2017

[20] Jurafsky, D. and Martin, J. H. *Speech and Language Processing: An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Harlow: Prentice Hall, 2006

[21] T. Sneha and K. S. Juran, "Approaches to machine translation", *Annuals of Library and Information Studies*, vol. 57, pp. 388-393, December 2010

[22] J. K. Sarkhel, "Approaches to machine translation," *Annals of Library and Information Studies*, vol. 57, pp. 388-393, 2010

[23] H. Budditha, "A Computational Grammar of Sinhala for English-Sinhala Machine Translation", A Thesis for Degree of Master of Philosophy, Department of Information Technology, University of Moratuwa. Sri Lanka, 2010

[24] MOSES, "Moses: open source toolkit for statistical machine translation system," [Online] 2017, Available: http://www.statmt.org/moses/

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

[26] A. Lopez and M. Post, "Beyond bitext: Five open problems in machine translation," Human Language Technology Center of Excellence, Johns Hopkins Uni., 2013

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

Processing, Manchester, 1992

[28] J. Brunning, "Alignment models and algorithms for statistical machine translation". Diss. Uni. of Cambridge, 2010

[29] G. Mukesh, S. Vatsa, J. Nikita, and G. Sumit, (2010), "Statistical Machine Translation' *DESIDOC Journal of Library & Information Technology*, vol. 30, no. 4, pp. 25-32

[30] F. B. Peter, C. S. John, A. P. Della, J. Vincent, P. Della, J. Fredrick, D. L. John, L. M. Robert, and S. R. Paul, "A statistical approach to machine translation" *Computational linguistics.* vol. 16, no. 2, pp, 79-85, 1990

[31] A. Lopez, "Statistical machine translation". *ACM Computing Surveys (CSUR),* vol. 40.3, p. 8, 2008

[32] S. A. Abdullah**,** "Large-scale Reordering Models for Statistical Machine Translation", degree of Doctor of Philosophy, University of Southampton**,** May, 2015

[33] Yitayew Solomon, "Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation" A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Masters of Science in Information Science, Addis Ababa Univer., Ethiopia, 2017

[34] M. Collins, "Statistical machine translation: IBM models 1 and 2." Columbia Univ (2011).

[35] M. Maučec, and G. Donaj, "Morphology in statistical machine translation from English to highly inflectional language". *Information Technology and Control*. Vol. 47 (1): 63-74, 2018

[36] S. Magnolini, N.P. An Vo and O. Popescu, "Learning the Impact of Machine Translation Evaluation Metrics for Semantic Textual Similarity," in Proceedings of Recent Advances in Natural Language Processing, Bulgaria, Sep 7–9 2015.

[37] M. L. Forcada, "Making sense of neural machine translation," Translation Spaces 6:2 (2017) 291–309, DOI 10.1075/ts.6.2.06for, 2017, [Online]. Available: https://www.dlsi.ua.es/~mlf/docum/forcada17j2. pdf

[38] M. Mara, “English-Wolaytta Machine Translation Using Statistical Approach,” St. Mary’s University School of Graduate Studies, Addis Ababa, Ethiopia, 2018.

[39] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation By Jointly Learning to Align and Translate” May 2016, [Online]. Available: https://arxiv.org/pdf/1409.0473.pdf

[40] S. Tripathi and S.J. Krishna, "Approaches to machine translation," Annals of library and Information Studies, pp. 388-393, Dec 2010

[41] M. Simard, P. Plamondon, “Bilingual Sentence Alignment: Balancing Robustness and Accuracy”, *Machine Translation*. Vol. 13, pp. 59–80, 1998

[42] E. Elif, G. Daniel and O. Kemal, “Simultaneous Word-Morpheme Alignment for Statistical Machine Translation”, Computer Science University of Rochester, Rochester, NY 14627

SSS[43] S. Sanja, G. Angelina and P. Damir, "Sentence Alignment as the Basis for Translation Memory Database," *Digital Information and Heritage*. Vol. 7, pp. 299-311, 2007

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

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

[46] S. Michel and P. Pierre, "Bilingual Sentence Alignment: Balancing Robustness and Accuracy," Centre for Information Technology Innovation, pp. 135-144, 1998

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

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

[49] S. Andre, “A Survey on Parallel Corpora Alignment”, Conference: *MI-STAR 2011*, Braga, Portugal , Pages 117–128

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

[51] T. Liang, W. Fai, and C. Sam, "Word Alignment Using Giza++ And Cygwin On Windows," *International Journal of Engineering Research & Technology (IJERT),* vol. 2, no. 5, pp. 17621765, 2013

[52] M. Creutz and K. Lagus, "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0", *Computer and Information Science*. Report A81,

Helsinki University of Technology, 2005

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

[54] K. Papineni, R. Salim, Todd Ward and Wei-Jing Zhu, "BLEU: A Method for Automatic Evaluation of Machine Translation." *In Proceedings of the 40th annual meeting on association for computational linguistics,* pp. 311-318. Association for Computational Linguistics, 2002

[55] M. S. Mirjam and D. Gergor, "Machine Translation and the Evaluation of its quality", September 7th, 2019, [ONLINE], Available: *DOI: http://dx.doi.org/10.5772/intechopen.89063*

[56] Mulubrhan Hailegebreal, "A Bidirectional Tigrigna – English Statistical Machine Translation", A Thesis Submitted to the School of Information Science in Partial Fulfillment for the Degree of Master of Science in Information Science, Addis Ababa University, Ethiopia, 2017

[57] Mulugeta Atsbaha, "Automatic Part-Of-Speech Tagger For Tigrigna Language Using Hybrid Approach", A Thesis Submitted To The School Of Information Science in Partial Fulfillment for the Degree of Master of Science in Information Science, Addis Ababa University, October 2016

[58] Kasa Gebrehiwot. "ሰዋሰው ትግርኛ, ኣዲስ ኣበባ": Mega printing enterprise, 2004

[59] ግዛቸው ደጀኑ መኩንን, "ሙዳየ ግእዝ", አ .አ ኢትዮጵያ ሚያዚያ 2014

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

[61] መምህር መስፍን ተከሥተ,የግእዝ ቋንቋ መራሕያን (ተውላጠ ስም) ምንነትና ትርጉም አገልግሎትና ዐይነቶች, አዲስ አባባ ኢትዮጵያ ጥቅምት 2015

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

[62] Kothari, C. R. Research Methodology – Methods and Techniques (Second Revised Edition) New Delhi: New Age

[63] Creswell, J. W. Research Design: Qualitative, Quantitative and Mixed Methods Approaches. New Delhi: SAGE PUBLICATIONS Ltd. P. 31, 2014

[64] Albert, G. Group of New Yorkers, New York State Legislation, New York: New York University, 2004.

[65] Paul Butler, Introduction to Viterbi Algorithm, March 2, 2021

Appendices

Appendix I: Extracting Ge'ez Bible Corpus from Website I:

Web scrapping : Extracting geez bible corpus from website

```
import numpy as np
import requests
from bs4 import  BeautifulSoup as bs

import timeit

start = timeit.default_timer()

next_link='/?q=version=Geez|reference=Gen.1'
geez_bible=[]

for i in range(236):
    req = requests.get("https://www.stepbible.org/"+next_link)
    soup = bs(req.content,"html.parser")

    # get next link
    links = soup.select('a')
    for i in range(len(links)):
        link = soup.select('a')[i]
        if link.get('class') == ['nextChapter']:
            next_link = link.get('href')

            # extract body
            body=soup.find('div', class_='passageContent')

            # find chapter title in body
            chapter_title = body.find('h2', class_='xgen')
            # append chapter title and body text to our list
            geez_bible.append(chapter_title.text)
            geez_bible.append(body.text)

stop = timeit.default_timer()

print('Time taken to scrap: ', stop - start)
```

Appendix II: Tokenizer and Frequency calculator

```
from collections import Counter

# Clean strings from special characters
tigrigna_str = remove_noise(tigrigna_str)
geez_str = remove_noise(geez_str)

# Count word frequency
tig_word_counter = Counter(tigrigna_str.split())
ge_word_coutner = Counter(geez_str.split())

# 10 most common words
print(f'Most common Tigrigna words: {tig_word_counter.most_common(10)}')
print(f'Most common Geez words: {ge_word_coutner.most_common(10)}')
```

```
Most common Tigrigna words: [('ንመበርትን', 1), ('ብምባኡ', 1), ('መምሬ', 1), ('ተቕመጠ', 1), ('ንግበር', 1), ('እተብህማ', 1), ('ሽሞብዓተ', 1), ('ኮ
ላ', 1), ('ብስዓ', 1), ('ተባሀለ', 1)]
Most common Geez words: [('ኣሕዝ', 1), ('ወሐለቮ', 1), ('ወሰመ', 1), ('ንግበር', 1), ('እምነ', 1), ('ወእትጋሥሥ', 1), ('ይኩን', 1), ('መካና', 1),
('ሰማሙ', 1), ('ዓሰቦ', 1)]
```

Appendix III: Ge’ez Morphological Segemntation

Geez Morphological segmentation

[3]:
```
import morfessor
import math
# function for adjusting the counts of each compound
def log_func(x):
    return int(round(math.log(x + 1, 2)))

infile = "geez.txt"
io = morfessor.MorfessorIO()
# io.read_corpus_list_files
train_data = list(io.read_corpus_list_file(infile))
model = morfessor.BaselineModel()
model.load_data(train_data, count_modifier=log_func)
model.train_batch()
io.write_binary_model_file("geez_model.bin", model)
```

```
.....
.....
.....
.....
```

[9]:
```
# test
geez_model_file = "geez_model.bin"

geez_model = io.read_binary_model_file(geez_model_file)

word = 'መምህር'
# for segmenting new words we use the viterbi_segment(compound) method
print (geez_model.viterbi_segment(word)[0])
```

```
['መ', 'ምህር', 'ኧ']
```

Appendix IV: Tigrigna Morphological Segmentation

Tigrigna Morphological segmentation

[5]:
```
# function for adjusting the counts of each compound
def log_func(x):
    return int(round(math.log(x + 1, 2)))

infile = "Vocabulary/tigrignaVocabulary.txt"
io = morfessor.MorfessorIO()
# io.read_corpus_list_files
train_data = list(io.read_corpus_list_file(infile))
model = morfessor.BaselineModel()
model.load_data(train_data, count_modifier=log_func)
model.train_batch()
io.write_binary_model_file("tg.bin", model)
```

```
.....
.....
.....
.....
.....
```

[6]:
```
# test
Tigrigna_model_file = "tg.bin"

Tigrigna_model = io.read_binary_model_file(Tigrigna_model_file)

word = 'ሰማይን'
# for segmenting new words we use the viterbi_segment(compound) method
print (Tigrigna_model.viterbi_segment(word)[0])
```

```
['ሰማይ', 'ን']
```