Addis Ababa University
School of Graduate Studies
College of Natural Science
Department of Computer Science

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

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

By

Eleni Teshome

Advisor: Yaregal Assabie (PhD)
Co-Advisor: Mulu Gebreegziabher

March 2013

Addis Ababa University
School of Graduate Studies
College of Natural Science

Department of Computer Science

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

Signature of the Board of Examiners for Approval

Name	Signature
Dr. Yaregal Assabie, Advisor	_____
Mulu Gebreegziabher, Co-Advisor	_____
_____	_____

March 2013

Declaration

I hereby declare that this thesis is my original work and has not been submitted as a partial requirement for a degree in any other university.

_____

Eleni Teshome

March 2013

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

Name	Signature
Dr. Yaregal Assabie, Advisor	_____
Mulu Gebreegziabher, Co-Advisor	_____
_____	_____

Acknowledgement

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 valuable assistance in the preparation and completion of this study.

First and foremost, my utmost gratitude goes to my advisor Dr. Yaregal Assabie whose sincerity and encouragement I will never forget. Dr. Yaregal has been my inspiration as I hurdle all the obstacles starting from the title up to the completion of this research work. He has been guiding me, providing all the necessary materials all the way through and giving me his significant feedback through every step of this study. My Co-advisor Mulu Gebreegziabher was also the one helping me whenever I face difficulties and leading me to face the difficulties so I could get to the solution.

I would like to thank the Moses support team for helping me through the entire process and for contributing their ideas at the time of failure or when some barriers occur. Their support was undeniable and it was what got me trough the hard times. Despite the distance, they have e-mailed the information I needed and the necessary steps that need to be taken for this study to come true.

My deepest gratitude goes to my family for their unfailing support and prayers, for their patience and encouragement to complete this research work.

Last but not the least, my gratefulness goes to the one above all of us, the almighty God, for answering my prayers and for giving me the strength to be able to perform this study.

Abstract

Natural language processing is the ability of computers to generate and interpret natural language. Machine translation is a sub-field of natural language processing that investigates the use of computer software to translate text or speech from one natural language to another. Ethiopia needs to cope up with the technology others are pursuing. Thus, the purpose of this study is to develop a bidirectional English-Amharic machine translation system using constrained corpus.

This research work implemented the statistical machine translation approach. In order to realize the goal, two different corpora were prepared and collected; the first corpus consisted of simple sentences and the other, complex sentences. Two language models were developed, one for Amharic and the other for English so as to ensure a bi-directional translation. Translation models were built which assigns a probability that a given source language text generates a target language text. A decoder was used which searches for the shortest path and expectation maximization algorithm was used for aligning words in the accurate order.

Experiments were carried out based on the dataset and results were recorded. The experiments were taken separately, one for the simple sentences and the other for complex sentences. The result obtained for the simple sentence using BLEU Score had an average of 82.22% accuracy for the English to Amharic, 90.59% for the Amharic to English and using the manual questionnaire preparation method, the accuracy from English to Amharic was 91% and from Amharic to English was 97%. For the complex sentences, the result acquired from the BLEU Score was approximately 73.38% for the English to Amharic, 84.12% for the Amharic to English and from the questionnaire method from English to Amharic was 87% and from Amharic to English was 89%. From this, we can see that the difference with the BLEU score and the questionnaire preparation method is not that visible so we can use both methods as reference.

As a result, with a corpus that is very large and appropriately examined, a better translation could be achieved since more words will be available in the provided corpus with higher probability of a particular word preceding another.

Table of Contents

Abbreviations

NLP – Natural Language Processing

NLU – Natural Language Unit

CPU – Central Processing Unit

MT – Machine Translation

RBMT – Rule Based Machine Translation

EBMT – Example Based Machine Translation

MRD – Machine Readable Dictionary

HAMT – Human Aided Machine Translation

MAHT - Machine Aided Human Translation

CAT – Computer Aided Translation

EASMT – English Amharic Statistical Machine Translation

BLEU - Bilingual Evaluation Understudy

VM- Virtual Machine

CMD - Command

CMU SLM - Carnegie Mellon Statistical Language Modeling

NP – Noun Phrase

VP – Verb Phrase

S - Subject

ART – Article

N - Noun

V – Verb

NP - Noun Phrase

VP - Verb Phrase

AdjP - Adjectival Phrase

AdvP - Adverbial Phrase

PP – Prepositional Phrase

Spec- Specifier

PC – Personal Computer

CHAPTER ONE

INTRODUCTION

Ever since the emergence of modern technology, most of the day-to-day activities of man have been performed with the assistance of different kinds of machines. Countries all over the world are utilizing these machines to make life easier. Ethiopia is one of the countries that have been trying to cope-up with the new technology the world has reached.

Ethiopia is a multi lingual country with over 80 distinct languages [1], and with a population of more than 84,734,262 as authorities estimated on the basis of the 2011 census taken from World Bank. Amharic being the official language of Ethiopia is spoken by a substantial segment of the population and it is ranked as the second language that has native speakers, which is preceded by the language Oromifa. In the 2007 census, 21.6 million speak Amharic which is 29.33% of the population. Owing to political and social conditions and the multiplicity of the languages, Amharic has gained ground throughout the country. The language is used in business, government and education. Newspapers are printed in Amharic as are numerous books on many subjects.

Although Ethiopia is the oldest independent nation in Africa [2], the country has not been using the technology as well as it should. Therefore, it becomes a necessity for the users to have a clear understanding of the communication media used in these devices in order to be competitive in the aspect of technology. Despite of the fact that almost all devices use English language, there must be a way to facilitate the implementation of every application.

Most important data such as literatures, e-books, research documents, papers and other raw facts are written in English. As a developing country, Ethiopia needs this information. Since it is inconvenient to document all the data in hard copies by translating them to Amharic, people need a way to access the data only when necessary. The people also demand to have clear understanding whenever they come across letters and personal e-mails written in English or whenever they want to express themselves in English. Nevertheless, it gives them a hard time looking for human translators as they might have private issues they usually do not want others to know about. Therefore, it is a necessity to come up with a solution that solves the problem they face especially with the confidentiality of their information. One way to meet their need is developing a bidirectional English Amharic machine translation system.

Translation is not a word-for-word substitution. A translator must interpret and analyze all of the elements in the text and know how each word may influence another. In order to achieve a better quality from a translation, a combined effort of the machine and the human mind is used. Machine translation is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another [3]. Machine translation performs substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed.

Translation as an art of rendering work of one language into another is as old as written literature. In this modern civilization of ours, the need for translation is ever growing and its importance in the field of business, economics and industrialization can't be ignored. These needs coupled with the modern scientific advancements paved the way to the conception of modern machine translation.

The term 'machine translation' refers to computerized systems responsible for the production of translation with or without human assistance. It excludes computer-based

translation tools which support translators by providing access to on-line dictionaries, remote terminology databanks, transmission and reception of texts, etc. The boundaries between MAHT (Machine-Aided Human Translation) and HAMT(Human-Aided Machine Translation) are often uncertain and the term CAT (Computer-Aided Machine Translation) can cover both, but the central core of machine translation itself is the automation of the full translation process.

With the emergence of Personal Computers, machine translation gained a strong momentum giving birth to commercially available software and hardware with translation tools and powerful dictionaries. But relentless research on the development of MT is going on to tackle problems in the various fields of application.

Ideally, machine translation is a batch process which is applied to a given text which produces a translated text which then only needs to be printed out. There are always the possibilities of multiple translations of the same text of requirements. MT is different; there cannot be a perfect automatic translation. The use of an MT system is contingent upon its cost effectiveness in practical situations [18].

Many countries are utilizing the machine translation system as a way of using and providing services for the public and paving a way to make themselves familiar to the ever growing technology. Ethiopia is a developing country that is working its way up to the top. Although the country is not accustoming to the technology as well, it's a necessity to follow the footsteps of others. And one feature to be acquainted to the technology is machine translation. It covers all branches of computational linguistics and language engineering, wherever they incorporate a multilingual aspect.

The main goal of this study, Bidirectional English-Amharic machine translation, is to translate English texts into Amharic or the vice versa by using a machine in a form that is understandable by human. Out of the approaches that are used in machine translation, this study plans to use a statistical machine translation.

1.1. Background

A given human language, whether written or spoken, is a fundamental part of human communication. Any hope of providing computer systems that claim intelligence approaching that of a human, therefore, rests on the hope of providing communication in natural language. Natural language is one of the fundamental aspects of human behavior and is a crucial component in our lives. It is a tool for communicating all around the world. Natural language processing can be described as the ability of computers to generate and interpret natural language [2].

Natural language processing, also called computational linguistics is widely regarded as a promising and critically important endeavor in the field of computer research. The goal of computational natural language processing is to create computational representations of the relationships that hold between language and some computational model (a knowledge base or a database schema) of the world; and to exploit those relationships to understand and generate language as appropriate to some set of tasks. The general goal for most computational linguists is to let the computer have the ability to understand and generate natural language so that eventually people can address their computers through text and speech as though they were addressing another person.

One of the reasons that have created a great interest in NLP is the fact that most human knowledge is recorded using natural language. Only computers that have the capability to understand natural language can access all the information contained in the natural language efficiently. The applications that will be possible when NLP capabilities are fully realized are impressive; computers would be able to understand and process natural language, translate languages accurately and in real time, or extract and summarize information from a variety of data sources, depending on the users' requests.

NLP demands deep NLU and modeling the natural language so that computer programs that act appropriately on the information contained in the text or utterance of the language can be developed.

In the natural language processing world, each and every language should be well understood by the machine so as to communicate very well. The process that lets machine understand the different languages used all around the world is called machine translation. Machine translation uses computer software to translate text or speech from one natural language to another [2]. Of the approaches followed in machine translation, this research work plans on using statistical machine translation so as the process of bidirectional translation could be performed.

English is a language that is mainly spoken on different parts of the world. Most of the materials, software or other applicable literatures are written in English. Amharic being the main spoken language in Ethiopia, it is undeniable that the people need the data written in English and they also need some sort of way to communicate with others. Since most countries use English language, the main language that is needed for communication with countries outside Ethiopia is English. That is why the bidirectional English Amharic machine translation is used. It translates the languages both ways so as it could be easier to know the language very well and it could also be applicable for those who want to translate English to Amharic or Amharic to English.

1.2. Statement of the Problem

A great deal of research has been conducted on NLP, and currently, there are systems that translate a text from natural language input texts for languages such as English, Chinese, German, Finish, etc. However, only one system, English Amharic Statistical Machine Translation, is being developed in Ethiopia.

The main factor that initiated for this study to be carried out is due to the fact that an application that translates English texts into Amharic or the other way is not available at hand for the time being. Because of this, people use human translation and they tend to be slower as compared to machines. Sometimes it can be hard to get a precise translation that reveals what the text is about without everything being translated word-by-word. In

addition, it can be more important to get the result without delay which is hard to accomplish with a human translator. That is when machine translation comes in, that solves most of the problems caused by a human translator.

Following the different approaches of machine translation, a satisfactory translation could be performed that will be used by numerous streams in our country. Everyone needs a well-organized and proficient translation; those who can speak both languages need it for confirmation purposes and those who only speak one of the languages need it for grasping knowledge if the translation is bidirectional. For various reasons, the translation of English texts in to Amharic, or the other way around becomes a necessity which is why this study is proposed. The English-Amharic machine translation mainly tries to accomplish an efficient translation with respect to accuracy as well as time.

1.3. Objectives

1.3.1. General Objectives

The general objective of this study is to design and develop a bidirectional English-Amharic machine translation system using constrained corpus.

1.3.2. Specific Objectives

The specific objectives of this study are to:

- Review the techniques on how to develop the system
- Collect a bilingual parallel corpus
- Design the architecture of the system
- Develop a prototype for the system; and
- Test the performance of the system both ways, that is from English to Amharic as well as from Amharic to English

1.4. Significance of the Study

The main contribution of this research work is:

- Translation of reading materials, such as scientific journals, e-books and other researches, so it could be used when necessary.
- Information Retrieval by using key words, that is, if a user wants to retrieve a document that has already been translated and at hand, the user can search for the document by using a key word.
- Speech to Speech translation, meaning if someone wants to work on speech to speech translation, it would be easier because the text to text translation is already at hand.
- Personal privacy because human translators won't interfere.
- Improving efficiency as compared to manual translation.

1.5. Scope of the Study

The scope of the study is on the implementing of bidirectional translation on the languages English and Amharic using constrained corpus. It is called 'Constrained' because some of the corpus used is prepared manually and the other is collected and examined carefully. Translation was performed mainly on simple sentences and based on those sentences it experiments on how the system translates from English to Amharic as well as Amharic to English. It was also tested on complex sentences to see and identify its applicability.

1.6. Limitation of the Study

The following are considered as the limitation of this research work:

- The main prototype developed in the study parses only simple Amharic sentences that are not more than four words, although a different corpus was prepared that entails complex sentences.
- This study uses a small sample prepared corpora due to a lack of large annotated corpora in the language pairs. Generating such large corpora is very expensive and time consuming. Although it is essential to produce such corpus, it was not generated for the purpose of this study due to time limitations and other constraints.

1.7. Methodology of the Study

1.7.1. Literature Review

Secondary data sources, like books, articles, publications and other previously written resources related to the topic were referred so that there could be more understanding about this particular subject matter. Researches related to this study were compiled so as to know the pros and cons of various machine translation techniques.

1.7.2. Data Collection

Sample text corpora were collected from relevant data sources with parallel text. The corresponding sentences, phrases and words in both halves were identified. Query preparation was performed so as to get ready for the testing process and alignment was also implemented in order for the sentences to make sense. Simple sentences were manually prepared and complex sentences were also collected to test its applicability.

For the simple sentences, 1020 sentences were manually prepared and for the complex sentences, 1951 were collected. That is, 414 from the Public Procurement Directive and 1537 sentences from the Bible. But these corpora are not that large because it was hard to find resources that mainly contain texts which are used in the day-to-day activities of human beings. Most of the resources are domain specific, for example, law, directives, Bible and so on.

1.7.3. Software Tools

The relevant tools for the developing of the English-Amharic machine translation were selected and used:

- VMware workstation, a workstation that enables users to set up multiple virtual machines (VMs) and use them simultaneously along with the actual machine.
- Ubuntu 11.04, an operating system which is suitable for the Moses environment
- Moses, a statistical machine translation system that automatically trains translation models for any language pair.
- Giza++, a word alignment tool
- MKCLS, a tool to train word classes
- IRSTLM, a language modeling kit
- BLEU Score, to evaluate the system
- Notepad, to make the corpus in system understandable way
- Microsoft Office 2007, software for the documentation of the study.

1.7.4. Experiment and Testing

The proposed system was tested to evaluate its performance. The query that has been prepared was used to test the system. To this end, the prototype of the system has been developed.

1.8. Organization of the Thesis

This section describes the organization of the rest of the research work. The next chapter briefly discusses about an overview of the Amharic language and how it differs from English. The third chapter will reveal the general applications of machine translation and the types are briefly identified. In this chapter, related works that has been done on machine translation are also generally described.

The fourth chapter, which is the main contribution of this research work, discusses about the Bidirectional English Amharic machine translation. Included in the discussion is the overall process that has been followed in order to make this study work.

The fifth chapter briefly describes the process of experiment and results. That is, it explains how the system is tested with the query provided.

Chapter Six, the last chapter, presents conclusions and recommendations based on the findings of the study. This chapter also indicates some pointers to future works. A reference list used for further reading is also included at the end of this chapter.

The appendices attached at the end provide additional information on some of the topics discussed in the different parts of this study and are found at the end of the last chapter.

CHAPTER TWO
THE AMHARIC LANGUAGE

2.1. Introduction

Ethiopia is a country that entails different nations and nationalities with over 80 distinct languages. Of all the languages being spoken, Amharic is a widely spoken language all over the country. It is a Semitic language mostly spoken in North Central Ethiopia. Amharic is the "official" language of the Federal Democratic Republic of Ethiopia and thus has official status nationwide. It is also the official or working language of several of the states within the federal system, including Amhara and the multi-ethnic Southern Nations, Nationalities and Peoples region. It has been the working language of government, the military, and the Ethiopian Orthodox Church throughout modern times. Outside Ethiopia, Amharic is the language of some 2.7 million emigrants (notably in Egypt, Israel and Sweden), and is spoken in Eritrea by Eritrean deportees from Ethiopia. It is written using a writing system called fidel or abugida, adapted from the one used for Ge'ez language [2].

This chapter briefly discusses on the different characteristics of Amharic words and sentences as compared to the English language. The major Amharic word classes, which are nouns, verbs, adjectives and conjunctions, are also discussed in this chapter. Pronouns are treated under the noun category. The various phrase structures of the Amharic language such as noun phrases, verb phrases, adjectival phrases, adverbial phrases and prepositional phrases, and sentence formalisms of the language, more importantly the features of Amharic simple and complex sentences, are all discussed in this chapter. To the better understanding of the material in this section, the chapter begins with a brief review of the language and Amharic alphabet.

2.2. A Brief Overview of Amharic Language

Ethiopia, officially known as the Federal Democratic Republic of Ethiopia, is a country located in the horn of Africa that has its own different ethnic groups and languages. Ge'ez is the ancient language, and was introduced as an official written language during the first Axumite kingdom in Ethiopia when the Sabeans sought refuge in Axum. The Axumite developed Ge'ez, a unique script derived from the Sabean alphabet, and it is still used by the Ethiopian Orthodox Tewahedo Church today [4]. Tigrigna and Amharic (Amharigna) are the modern languages which are derived from Ge'ez. Amharic is the official national language of Ethiopia. It belongs to the Afro-Asiatic language family which includes Arabic, Hebrew and Assyrian. Although other languages are spoken in Ethiopia, Amharic is the most widely used and well understood language.

Ethiopian Languages are divided into four major language groups. These are Semitic, Cushitic, Omotic and Nilo-Saharan. Amharic is a Semitic language. The majority of the 25 million or so speakers of Amharic can be found in Ethiopia, but there are also speakers in a number of other countries, particularly Eritrea, Canada, the USA and Sweden [5]. The name Amharic came from the district of Amhara in northern Ethiopia, which is thought to be the historic centre of the language.

2.3. Amharic Alphabet

Amharic is written with a version of the Ge'ez script known as Fidel. Unlike Latin language, Amharic script is far more complex because every first letter has six suffixes. The Amharic fidels are illustrated in the following figure. Other than those alphabets, there are also around forty labialized characters such as "ቋ ሟ ኗ ቧ ኳ…".

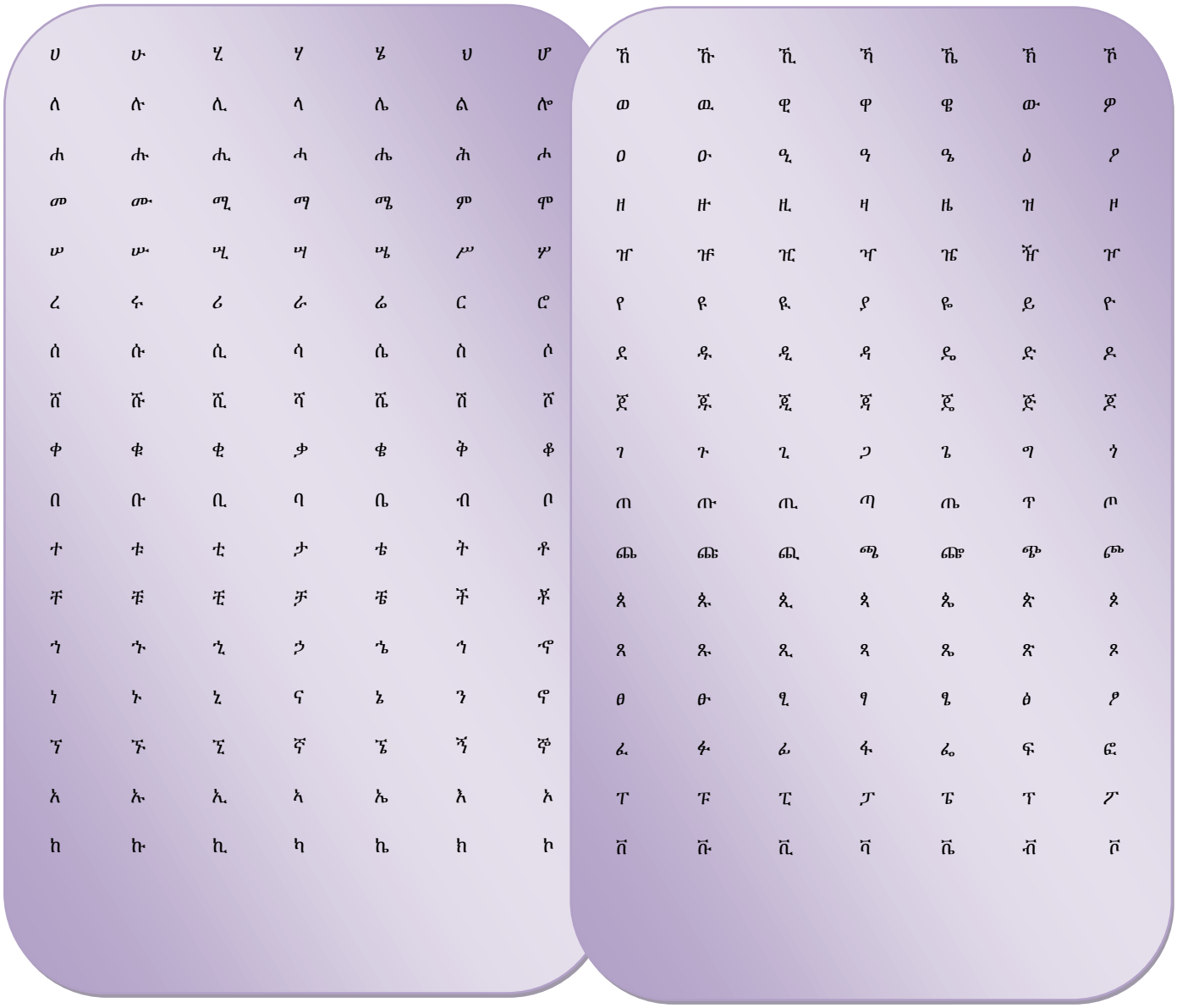


Figure 2.1. The Amharic Fidel

2.4. The Amharic Morphology

The roots of verbs and most nouns in the Semitic languages are characterized as a sequence of consonants or "radicals" (hence also the term consonantal root). Such abstract consonantal roots are used in the formation of actual words by adding the vowels and non-root consonants (or "transfixes") which go with a particular morphological category around the root consonants, in an appropriate way, generally following specific patterns. It is a peculiarity of Semitic linguistics that a large majority of these consonantal roots are triliterals (although there are a number of quadriliterals and in some languages, also biliterals). A triliteral or triconsonantal root is a root containing a sequence of three consonants.

As with many Semitic languages, Amharic uses triconsonantal roots in its verb morphology. The result of this is that a fluent speaker of Amharic can often decipher written text by observing the consonants, with the vowel variants being supplemental detail [2].

2.4.1. Personal Pronouns

In most languages, there is a small number of basic distinctions of person, number, and often gender that play a role within the grammar of the language. We see these distinctions within the basic set of independent personal pronouns. The following figure shows some examples.

English	Amharic
I	እኔ
She	እሷ
He	እሱ
They	እነሱ

Figure 2.2. Independent Personal Pronouns

2.4.2. Subject-Verb Agreement

All Amharic verbs agree with their subjects; that is, the person, number, and (second- and third-person singular) gender of the subject of the verb are marked by suffixes or prefixes on the verb.

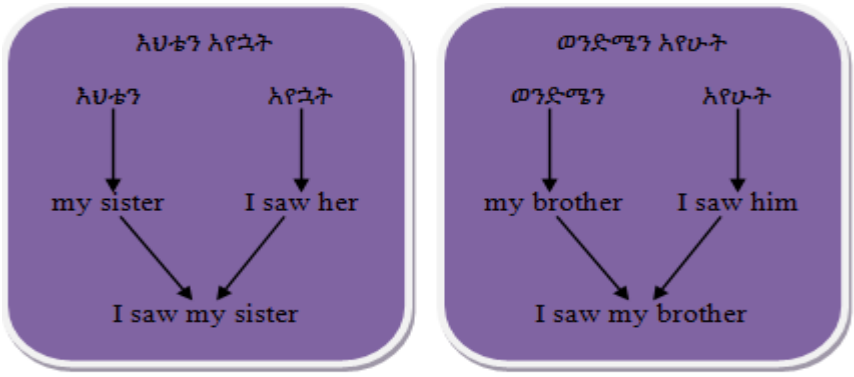


Figure 2.3. Subject-verb agreement

In the above figure, the letter "ኋ (ኡዋ)" and the letter "ሁ (ኡ)" in the words "አየኋት" and "አየሁት" respectively, explain the gender of the person. The words "እህት" and "ወንድም" being the root words, "ኤ" explains the person saying the sentence being possessive ("my").

2.4.3. Object Pronoun Suffixes

Amharic verbs often have additional morphology that indicates the person, number, and (second- and third-person singular) gender of the object of the verb.

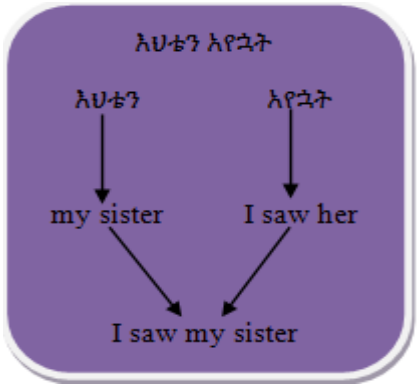


Figure 2.4. Amharic Morphological Structure

While morphemes such as (-ኣት) in the above figure are sometimes described as signaling object agreement, analogous to subject agreement, they are more often thought of as object pronoun suffixes because, unlike the markers of subject agreement, they do not vary significantly with the tense/aspect/mood of the verb. For arguments of the verb other than the subject or the object, there are two separate sets of related suffixes, one with a benefactive meaning (to, for), the other with an adversative or locative meaning (against', to the detriment of, on', at).

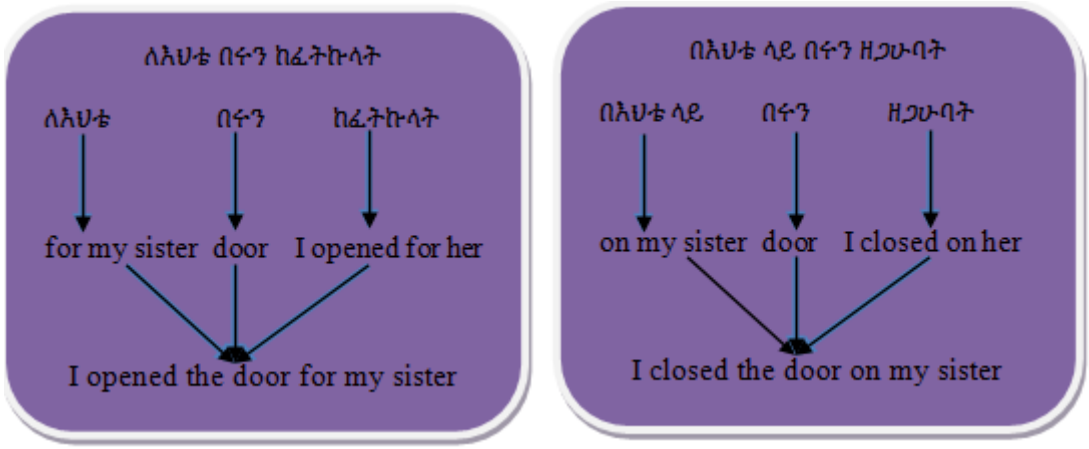


Figure 2.5. Benefactive and adversative meanings of a sentence

The above figure clearly shows the benefactive (to, for) and adversative (on, at) meanings of an Amharic sentence.

2.4.4. Possessive Suffixes

Amharic has a further set of morphemes that are suffixed to nouns, signaling possession:

ቤት *house*

ቤቴ *my house*

ቤቷ *her house*

In each of these four aspects of the grammar, independent pronouns, subject–verb agreement, object pronoun suffixes, and possessive suffixes, Amharic distinguishes eight combinations of person, number, and gender. For first person, there is a two-way distinction between singular (I) and plural (we), whereas for second and third persons, there is a distinction between singular and plural and within the singular a further distinction between masculine and feminine (you m. sg., you f. sg., you pl., he, she, they).

Amharic is a pro-drop language. That is, neutral sentences in which no element is emphasized normally do not have independent pronouns:

ኢትዮጵያዊ ነው	*he's Ethiopian*
ጋበዝኳት	*I invited her*

The Amharic words that translate he, I, and her do not appear in these sentences as independent words. However, in such cases, the person, number, and (second- or third-person singular) gender of the subject and object are marked on the verb. When the subject or object in such sentences is emphasized, an independent pronoun is used:

እሱ ኢትዮጵያዊ ነው	*he's Ethiopian*
እኔ ጋበዝኳት	*I invited her*
እሷን ጋበዝኳት	*I invited her*

Amharic has a complex morphology. Word formation involves pre-fixation, suffixation, in-fixation, reduplication and Semitic stem inter-digitations, among others. Like other Semitic languages, Amharic verbs and their derivations constitute a significant part of lexicon. In Semitic languages, words, especially verbs, are best viewed as consisting of discontinuous morphemes that are combined in a non-concatenative manner [6].

2.5. Amharic Phrasal Categories

A phrase is a structure in a language that is constructed from one or more words in the language. In Amharic, phrases are categorized into five categories, namely noun phrase, verb phrase, adjectival phrase, adverbial phrase and prepositional phrase [21]. Each phrase type can be categorized into 'simple' (where only one word class is represented) and 'complex' (where more than one word classes are represented).

2.5.1. Noun Phrases

A noun phrase is a syntactic unit in which the head (H) is a noun or a pronoun. It can be simple or complex. The simplest NP consists of a single noun (e.g. Abebe) or pronoun such as አሱ (he), አሷ (she), አነሱ (they), etc. A complex NP can consists of a noun (called head) and other constituents (like complements, specifiers, adverbial and adjectival modifiers) that modify the head from different aspects [21]. For example, in the NP ያ ያለፈው ሳምንት ቆንጆ መኪና 'That last week's beautiful car', ያ 'that' is a specifier, ያለፈው ሳምንት 'last week's' is an adverbial modifier, ቆንጆ 'beautiful' is an adjectival modifier specifying the material from which the object named by the መኪና 'car' (head) is made of. The grammar rule for the above example NP can be formulated as:

NP => Spec AdvP Adj NP

2.5.2. Verb Phrases

A verb phrase is composed of a verb as a head and other constituents such as complements, modifiers and specifiers. For example, in the VP ወደስራ ሄደች 'she went to work', ወደስራ 'to work' is prepositional phrase (PP) modifying the verb ሄደች 'went' from the place point of view. Therefore, structure rule for this example VP is:

VP => PP V

2.5.3. Adjectival Phrases

The construction of Amharic adjectival phrase is similar to that of a NP and a VP. It can be composed of an adjective (head), and other constituents such as complements, modifiers and specifiers. For example, in the AdjP ያ በጣም የሚያምር 'That very handsome (boy)', ያ 'that' is a specifier, በጣም 'very' is a modifier modifying the head of the AdjP, የሚያምር 'handsome'. The structural rule governing this phrase is:

AdjP => Spec Adv Adj

2.5.4. Prepositional Phrases

Prepositional phrase is constructed from a preposition (P) head and other constituents such as nouns, noun phrases, verbs, verb phrases, etc. In the PP እንደ እንሰሳ በዱር 'like an animal on the forest', for instance, እንደ 'like' and በ 'on' are prepositions which are combined with the nouns እንሰሳ 'an animal' and ዱር 'the forest', respectively to form their PPs. The two PPs, in turn, combine to result in the bigger PP that is provided in the example. This PP is formed by the structural rule:

PP => PP PP

And each of the PPs on the right hand side can further be analyzed as:

PP => P N

2.5.5. Adverbial Phrases

An Adverbial phrase is constructed from one or more adverbs in the language. In the example, ክፉኛ ታማለች 'she is severely ill', ክፉኛ 'severely' (head) is the only adverb that formed the AdvP and is governed by the rule:

AdvP => Adv

2.6. Amharic Sentence Structure

The sentence structure for Amharic language is a Subject-object-verb structure unlike English with a subject-verb-object combination. We can take the following as an Example.

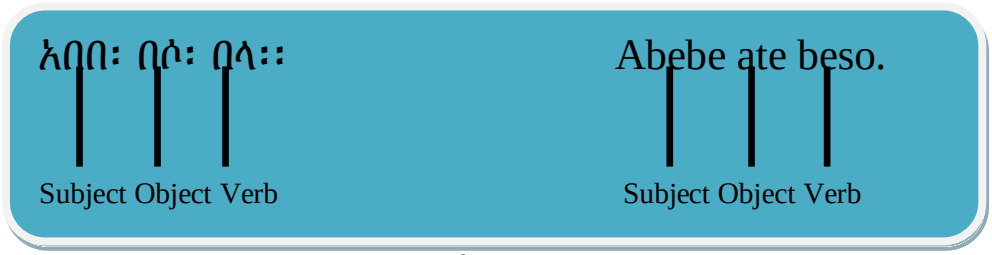


Figure 2.2. Structure of Amharic and English languages

Like English, Amharic nouns are words used to name or identify any class of things, people, places or ideas or a particular one of these. An important property of the Amharic is that any word that comes at the end of a complete grammatical Amharic sentence is a verb. As an outcome of this property, a word at the end of such a sentence is expected to be tagged as a verb by an Amharic tagger. Amharic verbs are also known for taking such subject markers as "ሁ፣ ህ፣ ክ፣ ሽ …" and so on. Adjectives in Amharic usually precede the nouns that they modify or describe.

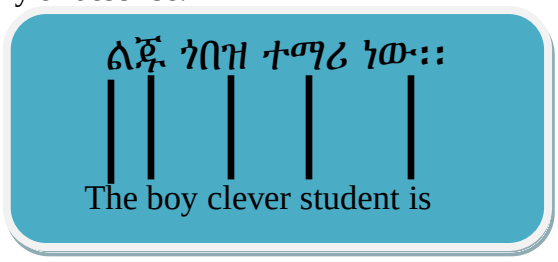


Figure 2.3. Sentence structure of Amharic and English Languages

Although the exact combination for the above Amharic sentence when converted to English is: "the boy is a clever student". In the above example, the adjective ጎ በ ዝ "clever" precedes the noun ተ ማ ሪ "student" which it modifies. But it does not mean that a word is an adjective just because it precedes a noun. For instance, in ይ ሄ ል ጅ "This boy", the word ይ ሄ "This" precedes the noun ል ጅ 'boy'. Although the word "ይ ሄ " functionally shares the feature of an adjective (modifier), it is a pronoun, a demonstrative pronoun.

Based on the number of verbs they contain, sentences can be classified into two: simple and complex sentences.

2.6.1. Simple Sentences

A simple Amharic sentence consists of an NP, which is the subject, followed by a VP that comprises the predicate.

በሩ ክፍት ነው

The door open is

The door is open

The morphemes like /-ኡ/ that is attached to a verb (e.g. መጥ-ኡ) refers to definiteness and number of the subject, and objects of the sentences. The morpheme that indicates the subject always precedes the morpheme that indicates the object. A given simple sentence may describe the state of being of the subject or an action that takes place in the sentence.

Example: *ራሄል ዶክተር ነች*

Rahel is a doctor; this sentence describes the present state of being of Rahel. Researchers classify simple sentences into four, namely: declarative sentences, interrogative sentences, negative sentences and imperative sentences. Declarative sentences are used to convey ideas and feelings that the speaker has about things, happenings, feelings, etc, that could be physical, mental, real or imaginary.

Example: *ራሄል ዶክተር ሆነች*

Rahel became a doctor

A sentence that questions about the subject, the complement, or the action the verb specifies, is called an interrogative sentence.

Example: *ራሄል መቼ ዶክተር ሆነች?*

When did Rahel become a doctor?

In order to construct interrogative sentences, Amharic sentences usually involve such interrogative pronouns as ማን 'who', ምን 'what', የት 'where', ስንት 'how many', and መቼ 'when'. These interrogatives can then be combined with prepositions to produce some more interrogative prepositional phrases like ከማን 'from whom', ለምን 'why', etc. One interesting feature of Amharic interrogative sentences is the fact that the interrogative

pronouns usually appear at the same position of the sentence where the thing being enquired would also appear.

Negative sentences simply negate a declarative statement made about something.

Example: *ራሄል ዶክተር አይደለችም*

Rahel is not a doctor

Simple imperative sentences convey instructions and mostly their subject is a second person pronoun that is usually but implied by the suffix on the verb.

Example: *ዝም በል!*

Shut up!

For the purpose of this study, the type of sentence used is the simple sentence. Complex sentences are also used but not exhaustively as the simple sentences.

2.6.2. Complex Sentences

Complex sentences in Amharic are those sentences that are composed of complex phrases such as NP, VP, or AdjP. The pattern of combination could take the form of a complex NP and a simple VP, a simple NP and a complex VP, or both complex NP and VP.

A complex NP is one that contains an embedded sentence with in it. The phrase, for instance, ራሔል የገዛችው የተበላሸ እንቁላል 'The rotten egg that Rahel bought' is a complex NP whose head is እንቁላል 'egg'. This head has been combined with the complement የተበላሸ 'rotten' in order to produce the simple NP የተበላሸ እንቁላል 'a rotten egg'. This simple NP, in turn, was combined with the dependent clause ራሔል የገዛችው 'that Rahel bought' to produce the above complex NP. The presence of the የ 'that' in it indicates that the clause is a subordinate clause and it cannot stand alone.

Similarly, a VP is complex if it contains more than one verb or a sentence within it. That is, like a complex NP, a complex VP also contains an embedded sentence that plays the role of a compliment or a modifier.

ራሔል ናሆም ወደ ኮሌጅ ስለሄደ ተደሰተች

Rahel was happy because Nahom went to college

The dependent clause here is ናሆም ወደ ኮሌጅ ስለሄደ 'because Nahom went to college' and ስለ is the part that made the clause dependent. As this clause indicates the reason for 'Rahel's happiness', it is used as an adverbial clause of reason.

Simple sentences are composed of simple noun phrase and simple verb phrases while complex sentences can consist of a complex NP and a simple VP, a simple NP and a complex VP, or a complex NP and a complex VP.

2.7. Articles

Articles are words used with a noun that specify whether the noun is definite or indefinite. English articles have three semantic choices for article selection:

- Definite article: the
- Indefinite article: a, an, some, any
- No article

Amharic language doesn't require articles that appear before nouns. Instead suffixes are added to show definiteness instead of using definite article.

E.g. *The boy:* ልጁ

In this example, "boy" refers to ልጅ and the definite article "the" is replaced by the suffix ኡ, in which it is pronounced as ልጅ-ኡ (lij-u).

2.8. Punctuation Marks

Punctuation marks are symbols that are used to organize and clarify the meaning of writing. Almost all punctuation marks used in English and Amharic are different. The following punctuation marks are used, in both languages, for the same purpose:

- Question mark **(?):** placed at the end of a sentence intended as direct question.
- Exclamation mark (**!**): placed at the end of a sentence to symbolize the anger, surprise or excitement of that particular sentence.
- Bracket (**( )**): to enclose an additional inserted word.
- Hyphen **(-**): to link the parts of a compound word or phrase.

The above punctuation marks are "t'iyake milkit", "kale agano", "kinif" and "serez", respectively in Amharic. The punctuation mark, apostrophe (**'**), is not used in Amharic. For example,

The boy's habit - የ ል ጁ ል ማድ

Here the word "the boy's" is translated as "የ ል ጁ". For the suffix ('s) in English also adds the suffix (የ) in Amharic. There are also other punctuation marks that are not used in Amharic, such as, bracket ([ ]), colon (:), swung dash (~) and so on, the colon is used in Amharic but for a different purpose.

Table 2.1. Punctuation marks used in both languages

English Symbol	Amharic	English Name	Purpose
White space character	:	space	To separate words
.	::	period	To show the end of a sentence
,	፣	Comma	To separate words or figures in a list
;	፤	semicolon	To indicate a pause longer than a comma
“ ” or ‘ ’	“ ”	Quotation mark	Used around direct speeches, quotations or to give emphasis to a word or phrase

The task of taking an input sentence and inserting legitimate word boundaries, called word segmentation, is performed using the white space characters in English. In Amharic, that uses Geez script for textual purpose and ‘:’ characters are used to separate words from each other. For the sake of simplicity, we will remove the punctuation marks in Amharic. That is, we’ll be using ‘space’ instead of ‘hulet net’ib’.

2.9. Conjunction

Conjunctions are used to connect words, phrases or clauses. Conjunctions in Amharic are coordinating or subordinating. They coordinate words, phrases, clause and sentences. The following are a list of English coordinating and subordinating conjunctions and their Amharic equivalent.

Table 2.2. Coordinating and Subordinating Conjunctions

English	Amharic Equivalent
Coordinating conjunctions	
As a result, Consequently, therefore, so	ስለዚህ፣ በዚህም ምክንያት፣ ስለሆነ ም
And	እና
But	ግን
For	ለ
Or	ወይም
Yet	ገና
Subordinating conjunctions	
As, as if	እንደ
Because	ምክንያቱም
Even if	ቢሆን ም
However, Although, though, even though	ሆኖ ም
Whenever	መቼ ም
Wherever	የ ት ም

CHAPTER THREE
LITERATURE REVIEW

3.1. Introduction

As mentioned earlier, the objective of this study is to design and develop a bidirectional English-Amharic machine translation system using constrained corpus. The human translation process may be described as decoding, the meaning of the source text, and re-encoding, the meaning in the target language. Behind this ostensibly simple procedure lies a complex cognitive operation. To decode the meaning of the source text in its entirety, the translator must interpret and analyze all the features of the text, a process that requires in-depth knowledge of the grammar, semantics, syntax, idioms, etc., of the source language, as well as the culture of its speakers. The translator needs the same in-depth knowledge to re-encode the meaning in the target language. Therein lays the challenge in machine translation: how to program a computer that will understand a text as a person does, and that will create a new text in the target language that sounds as if it has been written by a person [2].

Machine translation performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. Solving this problem with corpus and statistical techniques is a rapidly growing field that is leading to better translations, handling differences in linguistic typology, translation of idioms and the isolation of anomalies. In this study, statistical machine translation, the chiefly used approach of machine translation, was used [2]. Though it is difficult to find large corpora, a small parallel corpus was prepared and a linguist was contacted to verify the translation so as to have an effective result.

This chapter briefly identifies the details on machine translation, the several approaches and types of machine translation and how those types will be applicable in this wide spreading technological world are also discussed in this portion of the research work. It also discusses other studies that have been performed which are related to this study.

3.2. Machine Translation

Machine translation is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Amharic). To process any translation, human or automated, the meaning of a text in the source language must be fully restored in the target language, i.e. the translation. While on the surface this seems straightforward, it is far more complex. Translation is not a mere word-for-word substitution. A translator must interpret and analyze all of the elements in the text and know how each word may influence another. This requires extensive expertise in grammar, syntax, semantics, etc., in the source and target languages, as well as familiarity with each local region in which syntax and semantic mean sentence structure and meanings respectively [2].

Human and machine translation each have their share of challenges. For example, no two individual translators can produce identical translations of the same text in the same language pair, and it may take several rounds of revisions to meet customer satisfaction. But the greater challenge lies in how machine translation can produce publishable quality translations. As companies are faced with higher volumes of content that require translation, and as the time allotted for these projects shrinks, more and more organizations are weighing the pros and cons of machine translation as a viable solution to tackle these time-critical projects.

Combining machine translation with complementary technologies and human translators, quick, reliable and usable translations for review of large volumes of text can be produced, all completed in a fraction of the time that would be required for a standard translation process. While machine translation quality falls far short of human translation, if used properly in conjunction with human reviewers, the utility of machine translation can be stretched to include both non-distribution applications including document review, legal discovery, and internal correspondence, as well as distribution-level materials for which extremely fast turnaround times are a requirement.

Machine translation has some advantages and disadvantages. The merits of machine translation are:

- **Quick Translation**: using the machine translation system enables you to save your time while translating large texts.
- **Low price**: if a professional translator is hired to translate a text, enough money should be paid for each page but very often we need just a point of matter, general idea. In this case machine translation system is reliable and effective.
- **Confidentiality:** Many people use machine translation systems to translate their private emails, because no one would agree to give a private correspondence to translator who is not known, or no one would entrust documents to other people.
- **Universality**: Usually a professional translator becomes specialized in a definite field, but machine translation system can translate any text about any area.
- **Online translation and translation of web page content:** The advantage of online translation services is obvious. Online translation services are at hand and you can translate information quickly with this service. Furthermore you can translate any web page content and query of search engine by the use of machine translation systems.

As every concept has an advantage, it also has a disadvantage. So, the demerits of machine translation are:

- **Lack of superior exactness**: entrusting machine translation system is hard if superior exact translation of the official documents, agreements and so on is needed.
- **Inferior translation quality of the texts with ambiguous words and sentences:** machine translation is based on formal and systematic rules so sometimes it can't solve ambiguity by concentrating on a context and using experience or mental outlook as a human translator.

3.2.1. History of Machine Translation

Machine translation got off the ground in the late forties right after the World War II and there were several events which led the development of machine translation. MT was constrained by several factors: limitation of hardware particularly, inadequacy of memories and slow access and unavailability of high level programming language. The linguistic study was not correlated with machine translation research, so researchers relied on the dictionary- based approach and the application of statistical method [13].

Military intelligence needs were the main concern of the translation task in this period. Translating large volumes of technological research gave a boost to the American Industrialization.

Faced with technical constraints, early researchers knew that there could be no perfect high quality translation, and suggested human involvement in the translation process. They also proposed the "development of controlled languages and restriction of systems to specific domains."

A criterion was set concerning the success and failure of machine translation in its first 50 years of research and development [13]. These criteria are the conceptual, engineering, operational, commercial and communicative criteria.

1. The conceptual level concerns primarily the researchers in processing new interesting concepts and demonstrating their feasibility and advantages in laboratory prototypes.
2. The engineering level primarily engages the developers in implementing innovative architects in using better programming technique to build prototypes or system.
3. The operational level primarily concerns the users in running prototypes or systems in a cost efficient and satisfactory way under operational conditions.
4. The commercial level concerns vendors and should be judged in terms of financial returns not the number of installations or client.
5. The communicative level concerns the image that decision-makers and the public will form about the field in general.

3.2.2. Machine Translation Approaches

MT has taken long strides in research and development, but still it has a long way to go. With present computer developments and its viability for integration to Machine Translation, there is a wide horizon for huge developments in all aspects of MT [2]. The following are the different approaches that could be pursued in the machine translation system and in which ways they are applicable to be used as an effective translation approach.

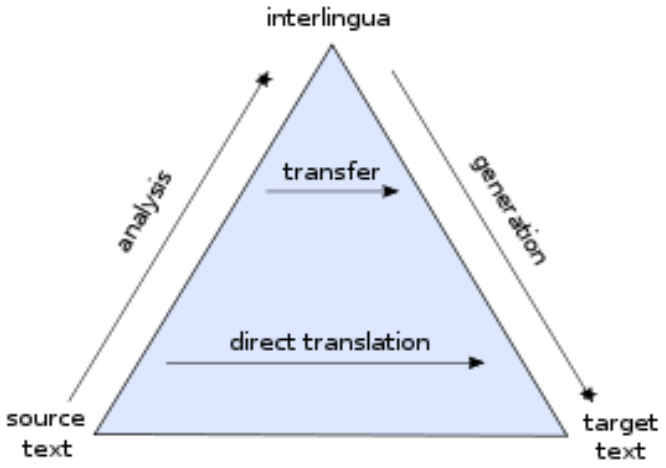


Figure 3.1. Depths of Interlingua machine translation

Rule-based Machine Translation

A rule based machine translation system, also known as "Knowledge-based Machine Translation", is a general term that denotes machine translation systems based on linguistic information about source and target languages basically retrieved from (bilingual) dictionaries and grammars covering the main semantic, morphological and syntactic regularities of each language respectively. It consists of collection of rules called grammar rules, lexicon and software programs to process the rules.

Rule based approach is the first strategy ever developed in the field of machine translation. It is extensible and maintainable. Rules are written with linguistic knowledge gathered from linguists. Rules play major role in various stages of translation: syntactic processing, semantic interpretation and contextual processing of language [2].

Translations are built on gigantic dictionaries for each language pair and sophisticated built-in linguistic rules. Users can improve the out-of-the-box translation quality by adding their terminology into the translation process. They create user-defined dictionaries which override the system's default settings. In most cases, there are two steps: an initial investment that significantly increases the quality at a limited cost, and an ongoing investment to increase quality incrementally. While rule-based machine translation brings companies to the quality threshold and beyond, the quality improvement process may be long and expensive.

The software parses text and creates a transitional representation from which the text in the target language is generated. This process requires extensive lexicons with morphological, syntactic, and semantic information, and large sets of rules. The software uses these complex rule sets and then transfers the grammatical structure of the source language into the target language.

A typical English sentence consists of two major parts: noun phrase (NP) and verb phrase (VP). These two parts can be further divided as per the structure of the sentence. ‘Rewrite rules’ are used to describe what tree structures are allowable for a given sentence. Only the sentence with right structure can lead to correct translation. Following are the rules to represent a simple grammar.

S => NP VP
VP => V NP
NP => Name
NP => ART N

Where, S stands for sentence, V for verb, N for noun and ART for article. A grammar can derive a sentence if there is a sequence of rules to rewrite the start symbol, S, into a sentence [7].

Translation, in rule based machine translation system, is done by pattern matching of the rules. The success lies in avoiding the pattern matching of unfruitful rules. Knowledge and reasoning are used for language understanding. General world knowledge is required for solving interpretation problems such as disambiguation. Context specific knowledge can be used to determine the referent of noun phrases and disambiguating word senses based on what makes sense in the current situation. A knowledge representation consists of knowledge-base and inference techniques. Inference techniques apply inference rules to derive new sentences from the knowledge-base.

Anaphora is the linguistic phenomenon of pointing back to a previously mentioned item in the text. The pointing back word or phrase is called anaphor and the entity to which it refers or for which it stands is its antecedent. For example,

Mary is very sick and she is going to the hospital.

Here the 'Mary' is anaphor and 'she' is antecedent. When the anaphor refers to an antecedent and when both have the same referent in the real world, they are termed co-referential. Co-reference is the act of referring to the same referent in the real world. The process of determining the antecedent of an anaphor is called anaphora resolution. Rules that are used for resolution are called resolution rules. These rules are based on different sources of knowledge. Needless to say that interpretation of anaphora is crucial for the successful operation of a machine translation system.

The advantage of rule based machine translation approach is that it can deeply analyze at syntax and semantic levels. There are drawbacks such as requirement of huge linguistic knowledge and very large number of rules to cover all the features of a language but it can produce an efficient result as compared to the other types of machine translation approaches. Having input sentences (in some source language), an RBMT system generates them 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.

Rule-based MT provides good out-of-domain quality and is by nature predictable. Dictionary-based customization guarantees improved quality and compliance with corporate terminology. But translation results may lack the fluency readers expect. In terms of investment, the customization cycle needed to reach the quality threshold can be long and costly. The performance is high even on standard hardware.

Example-based approach

Example-Based Machine Translation (EBMT) relies on previous translations performed by human to create new translations without the need for human translators. The previous translations are called the training corpus. For the best translation quality, the training corpus should be as large as possible, and as similar to the text to be translated as possible [8].

When the exact sentence to be translated occurs in the training material, the translation quality is human-level, because the previous translation is re-used. As the sentence to be translated differ more and more from the training material, quality decreases because smaller and smaller fragments must be combined to produce the translation, increasing the chances of an incorrect translation.

As the amount of training material decreases, so does the translation quality; in this case, there are fewer long matches between the training texts and the input to be translated. Conversely, more training data can be added at any time, improving the system's performance by allowing more and longer matches. EBMT usually finds only partial matches, which generate lower-quality translations.

Highly agglutinative languages, like Amharic language, pose a challenge for Example Based MT. Because there are so many inflected versions of each stem, most inflected words are rare. If the rare words do not occur in the corpus at all, they will not be translatable by EBMT. If they occur only a few times, it will also be hard for EBMT to have accurate statistics about how they are used.

Connectionist Approach

Connectionism, parallel to computation, is significant development in the computational modeling of cognition. A distinctive feature is the computation of the strengths of links between nodes of networks and the adjustment of the weightings as a result of actual analyses, i.e. the network learns about the links and their strengths for later use [13].

The approach has demonstrated empirical success in tackling Language Understanding tasks, which can be considered as a particular case of translation. The connectionist system separately approaches the syntactic and semantic features associated to a language, resulting in a translation model which is quite complex.

The development of machine translation is geared towards commercialization of a practical MT system. In the advent of computer revolution, there are possibilities for integration of various types of approaches and from there; the MT community will be able to carve another unique paradigm of MT for the future.

Statistical Machine Translation

The goal of statistical machine translation is to translate a source language sequence in to a target language sequence by maximizing the posterior probability of the target sequence given the source sequence. Probabilities that describe correspondences between the words in the source language and words in the target language are learned from a bilingual parallel corpus and language models are learned from a monolingual text in the target language. As the available training corpus becomes large, the performance of the system increases. Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora. Where such corpora are available, impressive results can be achieved translating texts of a similar kind, but such large corpora are still very rare.

Statistical machine translation utilizes statistical translation models whose parameters stem from the analysis of monolingual and bilingual corpora. Building statistical translation models is a relatively quick process, but the technology relies heavily on existing multilingual corpora. A minimum of 2 million words for a specific domain and even more for general language are required [9]. Theoretically it is possible to reach the quality threshold but most companies do not have such large amounts of existing multilingual corpora to build the necessary translation models. Additionally, statistical machine translation is CPU intensive and requires an extensive hardware configuration to run translation models for average performance levels.

Statistical MT provides good quality when large and qualified corpora are available. The translation is fluent, meaning it reads well and therefore meets user expectations. However, the translation is neither predictable nor consistent. Training from good corpora is automated and cheaper. But training on general language corpora, meaning text other than the specified domain, is poor. Furthermore, statistical MT requires significant hardware to build and manage large translation models.

Table 3.1. Rule-based MT Vs. Statistical MT

Rule-Based MT	Statistical MT
+ Consistent and predictable quality	- Unpredictable translation quality
+ Out-of-domain translation quality	- Poor out-of-domain quality
+ Knows grammatical rules	- Does not know grammar
+ High performance and robustness	- High CPU and disk space requirements
+ Consistency between versions	- Inconsistency between versions
- Lack of fluency	**+ Good fluency**
- Hard to handle exceptions to rules	**+ Good for catching exceptions to rules**
- High development and customization costs	**+ Rapid and cost-effective development provided the required corpus exists**

Given the overall requirements, there is a clear need for another approach through which users would reach better translation quality and high performance (similar to rule-based MT), with less investment (similar to statistical MT).

In this study, statistical machine translation was implemented as a way to translate the source language to the target language. For this activity to be accomplished, a set of processes were performed, different tools are applied and a small corpus were prepared in order to build the system.

3.2.3. Types of Machine Translation

The development of machine translation had employed different types of methodology and approaches. In more than five decades of research and development, machine translation has evolved into a truly practical system which is now beginning to build its foundation in the world market. The analysis of machine translation's 50 years of history reveals several factors in its success and failures which are determined at several levels such as Conceptual level, engineering level, Operational level, Commercial level and Communicative level. The following are the four types of Machine Translation [13].

1. **MT for Watchers:** intended for readers who wanted to gain access to some information written in foreign language who are also prepared to accept possible bad 'rough' translation rather than nothing. This was the type of MT envisaged by the pioneers. It came in with the need to translate military technological documents. It was like dictionary- based translation far away from linguistic based machine translation.
2. **MT for Revisers:** aims at producing raw translation automatically with a quality comparable to that of the first drafts produced by human. The translation output can be considered only as brush-up so that the professional translator freed from that very boring and time consuming task can be promoted to revisers.
3. **MT for Translators:** aims at helping human translators do their job by providing on-line dictionaries, thesaurus and translation memory. This type of machine translation system is usually incorporated into the translation work stations and the PC based translation tools. And those systems running on standard platforms and integrated with several text processors are the ones that attained operational and commercial success.
4. **MT for Authors:** aims at authors wanting to have their texts translated into one or several languages and accepting to write under control of the system or to help the system disambiguate the utterance so that satisfactory translation can be obtained without any revision.

3.2.4. Machine Translation Processes

In machine translation, for decoding the meaning of the source text and re-encoding the meaning in the target language, a series of processes will be taken. The following are machine translation processes provided a particular query.

- Query input: the text to be translated will be inserted as a query.
- Query translation: the text will be processed and translated from one natural language to the other.
- Displaying result: the translated text will be returned so as to get the desirable output.

3.2.5. Machine Translation in a World of Information

Undoubtedly, the role of MT in the world of information can no longer be ignored. This is represented by the system's ability to convey sufficient information so that one can gain necessary information from it. With the emergence of the world wide web of information channels, millions of users all over the world can gain access to the information super high-way, thereby speakers of different languages will avail themselves of the automatic translation service.

MT can be an effective tool for the facilitation of multi-lingual on line discussions allowing users who speak different languages to communicate with one another and promote the globalization of information highway [19]. As the computer-based translation activities are expanding, they embrace any process which will result in the production of texts, documents in bilingual and multilingual contexts. There is a possibility that MT will be seen as the most significant component in the facilitation of international communication and understanding in the future "information age". And the integration of automatic machine translation technology in the on-line environment is the potent force that initialized its importance in the world of information.

3.3. Morphology

Morphology is the structure of words, including pattern of inflections and derivation. It makes translation easier and minimizes the size of the corpus since the words will be separated into different forms. Followed are the two approaches of morphology.

3.3.1. Morphological Analyzer

Morphological Analysis was developed by Fritz Zwicky in the 1940's and 50's as a method for systematically structuring and investigating the total set of relationships contained in multi-dimensional, usually non-quantifiable, problem complexes [2]. Morphological Analysis is an extension of Attribute listing. It can be extended to virtually any problem area that can be structured dimensionally.

This method can be used in the area of machine translation by integrating it to the system. Morphological analysis, in machine translation, is the identification of a stem-form from a full word-form. The morphological analyzer performs a recursive and exhaustive search on all possible segmentations of a given word. It takes a particular word as an input and it produces all possible segmentations of the word as an output. Every segmentation should specify:

- A single stem in that word
- Each suffix in that word
- A syntactic analysis for the stem and each identified suffix

Once all the possible and correct segmentations of a word has been found, the morphological analyzer combines the feature information from the stem and the suffixes encountered in the analyzed word to create a syntactic analysis that is returned [8].

3.3.2. Morphological Synthesizer

Morphological synthesis or generation is a process of returning one or more surface forms from a sequence of underlying (lexical) forms. The morphological generator delivers a target language surface form for each target-language lexical form, by suitably inflecting it. Morphological synthesis systems are used as components in many applications, including machine translation, spell-check, speech recognition, dictionary (lexicon) compilation, POS tagging, morphological analysis, conversational systems, automatic sentence construction and many others.

Morphological synthesizers have vital role in NLP systems. They are used to generate surface word forms, which are the ones that are found in everyday communication, from lexical components that could be stored separately in different databases (lexicons). The combination of morphs to give meaningful words is the concern of morphological synthesis or generation. The morphological generator will synthesize the inflected word in its right form based on the morphological features.

3.4. Alignment

Alignment is the arrangement of something in an orderly manner in relation to something else. It can be performed at different levels, from paragraphs, sentences, segments, words and characters. Word alignment is relevant for this research since it will mainly be dealing with simple sentences.

3.4.1. Word Alignment

Word alignment is an inference problem of word correspondences between different languages given parallel sentence pairs. The task of word alignment is to link the correspondences between words in a source language and their translations in a target language, in such a way that the aligned words supply the same contents.

3.4.2. Challenges of Automatic Word Alignment

Word alignment plays a critical role in statistical machine translation by mapping source sentence words to target sentence words. However, automatic word alignment of parallel sentence pair is not a simple task. For most parallel texts, choosing the sentences in one natural language to be the translation of another language is challenging. The following figure shows an example of aligned word pairs and sentence translated from English to Amharic.

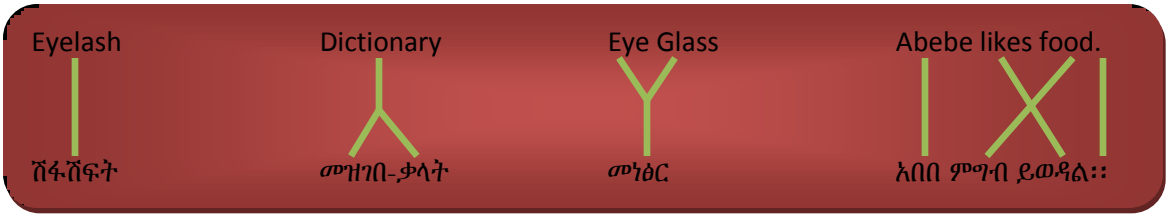


Figure 3.2. Aligned sentence and word pairs

As it can be seen from the above example, there are three possible structures that could mess up the alignment of a particular word, that is, the correspondence could be:

- One-to-one
- One-to-two
- Two-to-one

And if we look at the sentence, the subject-verb-object word-order in English is changed to the subject-object-verb combination. As a result, there must be a way to solve this alignment problem so as to get an effective output. Expectation maximization algorithm will be used to align the words in a particular sentence.

3.5. Measuring Retrieval Effectiveness

The effectiveness of retrieval systems can be evaluated using several measures. The basic and most widely used measures are precision (measures how efficiently the system provides only the relevant items) and recall (measures the ability of the system to retrieve the available relevant documents) of search output. Given a set of relevant judgments one can determine how best a systems performance is.

In this research work, the accuracy will be calculated based on two methodologies. The first one was calculated using the BLEU Score which calculates the effectiveness by taking the reference of a particular text and matching it with the queried translation.

The second methodology for the testing was recorded by preparing a questionnaire and giving it for thirteen candidates so as to assess the translation produced by taking an average manually.

3.6. Related Works

Considerable research efforts have been expended in developing and designing machine translation system. Researches on machine translation approaches and strategies, techniques and implementation have been documented in many old and recent publications. In Ethiopia, some machine translation systems has been tried to be developed and documented as a research work. The following subtopics are some of the related works that has been carried out in and outside of Ethiopia.

3.6.1. English-Amharic Statistical Machine Translation (EASMT)

This research work is being done on English-Amharic Statistical Machine Translation in which it uses two million bilingual corpora. Papers and essays have been published during the process of doing the study. The following topics briefly discusses on those papers that are totally found from the published papers.

Bilingual Data Mining for the English Amharic Statistical Machine Translation

SMT system is data driven that rely on bilingual parallel aligned corpus, the larger, the better. To develop the system, a size of two Million word pairs of parallel data was collected on English Amharic sentence pairs collected from News, Parliamentary and constitutional documents. And out of the two million pairs, 40 thousand sentence pairs were taken for experiment purposes.

The study used five steps to process a bilingual text corpus which are: raw data collection, document alignment, tokenization, sentence splitting and sentence alignment. Due to some difficulties occurred, the study put some solution as to how to make it better, that are:

- increasing the number of English-Amharic proclamation corpus as much as possible,
- further analyzing the experiment conducted ,
- increasing the translation quality by using linguistic knowledge [25].

The study is being developed to come up with a better translation so as those who need it could use it and others who want to do further research relying on this study could refer to it.

Preliminary Experiments on EASMT

The challenge to develop MT using rule-based approach to Amharic, which is considered as one of the NLP scarce resource language, is enormous. The same might not be true for well developed NLP resourced languages. What makes it challenging for under resourced languages is that the rule-based MT heavily employs integrated linguistic knowledge, rules and resources of both the source and target languages. However, it is almost impossible to develop a MT system using the rule-base method for Amharic at least in the near future as it is under resourced language with respect to the different linguistic knowledge, rules and resources.

Thus, this study, English-Amharic Statistical Machine Translation, followed the statistical approach which relies heavily on bilingual parallel aligned corpora of the source and target languages. The challenge is minimized since the statistics based approach requires very limited computational linguistic resources compared to the rule-based approach that might take so many years to develop some or all of the mentioned linguistic resources. The corpus collected consisted of raw English-Amharic corpus from the Parliament of the Federal Democratic Republic of Ethiopia.

The study followed the approach of phrase-based system. It has been trained using the English-Amharic parallel Training Set as translation examples and tested using the English Source Text as new sentences that gives an output Target Text of translated Amharic sentences. The baseline phrase-based BLEU score result indicates that 35.32% translation has been achieved [23].

3.6.2. English-Oromo Machine Translation

The research work, English-Oromo Machine Translation, has two main goals: the first is to test how far one can go with the available limited parallel corpus for the English-Oromo language pair and the applicability of existing Statistical Machine Translation (SMT) systems on this language pair. The second goal is to analyze the output of the system with the objective of identifying the challenges that need to be tackled.

The architecture includes four basic components of SMT: language modeling, translation modeling, decoding and evaluation. The language modeling component takes the monolingual corpus and produces the language model for the target language. The translation modeling component takes the part of the bilingual corpus (English and Oromo) as input and produces the translation model for the given language pairs. The decoding component takes the language model, translation model and the source text to search and produce the best translation of the given text. The Evaluation component of the system takes the system output and the reference translation and compares them according to some metric of textual similarity.

The parallel documents that were used are: Oromo versions of some chapters of the Bible that are available in English and Oromo as well as the United Nation's Declaration of Human Rights, the Ethiopian constitution and so on. Since the documents were not prepared for the translation purpose, the result was not as pure as it is needed for the system and one sentence in one language may be equivalent to more than one sentence in another. The research work concluded that the system performed 43.96% BLEU score which makes the performance not to low as compared to the systems built on a relatively sufficient amount of resource [10].

3.6.3. Dictionary-based Amharic-English Information Retrieval

The approach followed in this study is a dictionary-based machine translation to translate the Amharic queries in to English bag-of-words. At a general level, two approaches are used; both consisting of a first step that transforms the Amharic topics in to English queries, followed by a second step that takes the English queries as input to a retrieval system. In both approaches the translation was done through a simple dictionary lookup that takes each stemmed Amharic word in the topic set and tries to get a match and the corresponding translation from an MRD.

One of the experiments reported removes non-content bearing words from the Amharic queries, while the other uses a list of English stop words to perform the same task. Two runs were performed on the data set using two sets of queries. In the first run, stop word removal was done before the translation of terms, in the second one, the stop word removal was done only after the terms were translated in to English. The resulting translated (English) terms are then submitted to a retrieval engine that supports the Boolean and vector space models.

Although non-content bearing words were removed from the Amharic queries in the first approach, a lot of stop words were introduced while performing the dictionary lookup, hence introducing noise. The study concluded that the combination of the two approaches may result in a better performance in terms of precision; while means of query expansion in order to increase the recall remains open for investigation [11].

3.6.4. Apertium: Free/Open Source Rule-Based Machine Translation

Apertium is a free/open-source, rule-based machine translation platform that provides a modular shallow-transfer machine translation engine with text format management, finite-state lexical processing, statistical lexical disambiguation and shallow transfer based on finite-state pattern matching.

In Apertium, language-pair development has also motivated changes in the platform such as three-stage and multi-stage (>3) structural transfer was introduced to deal with the languages. To generate translations, which are reasonably understandable and easy to correct, between related languages, one can just augment word for word translation. It has more than one hundred developers in order for the codes to be updated very frequently.

Apertium was selected to participate as a mentoring organization in the 2009 Google Summer of Code. Successful projects like two new language pairs, an improved part-of-speech tagger, a web-service infrastructure, porting of the lexical component to Java and hybridizing Apertium with other systems has been performed and lots of works are required. Since Apertium is a transfer system, generating a new pair involves the creation of explicit bilingual resources which is considered as a limitation and it is expected to be worked-on [12].

CHAPTER FOUR

DESIGN AND DEVELOPMENT OF THE SYSTEM

4.1. Introduction

As mentioned earlier, this study plans to develop bidirectional English-Amharic machine translation system. In order for this study to be accomplished well, some measures were taken. Approaches were followed, a corpus was collected according to the approach that has been pursued and different tools were used to develop a working system. This chapter briefly entails how this study was exhaustively done.

4.2. Approach Followed for the Design

As indicated earlier, machine translation comprises different approaches. For this study, Statistical machine translation was used. Statistical machine translation is an approach that tries to generate translations using statistical methods based on bilingual text corpora. Researchers discovered that as the size of the corpora increases, the accuracy of the translation improves. The first statistical machine translation software was CANDIDE from IBM [2]. Statistical machine translation has three components: translation model, language model and decoder. The following depicts the architecture of this approach.

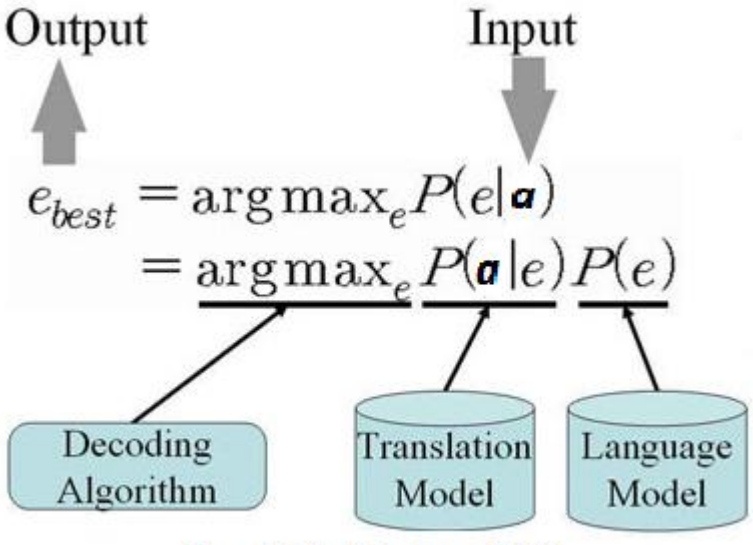


Figure 4.1. Architecture of SMT

If we want to translate a sentence **a** in the source language **A** to a sentence **e** in the target language **E**, the noisy channel model describes the situation in the following way [14]:

Suppose that the sentence **a** to be translated was initially conceived in language **E** as some sentence **e**. During communication e was corrupted by the channel to **a**. Now, assuming that each sentence in **E** is a translation of **a** with some probability, and the sentence that we choose as the translation (**ê**) is the one that has the highest probability. In mathematical terms,

$$\hat{e} = \underset{e}{\text{argmax}}\ P(e|a) \tag{4.1}$$

Intuitively, **P (e|a)** should depend on two factors:

1. The kind of sentences that is likely in the language **E** which is known as the language model, **P (e).**
2. The way sentences in **E** get converted to sentences in **A** which is called the translation model, **P(a|e)**.

Baye's rule states the following:

$$P(e|a) = P(a|e) * P(e) / P(a) \tag{4.2}$$

Where **a** is the source text and **e** is the target.

$$\underset{e}{\text{argmax}}\ P(e|a) = \underset{e}{\text{argmax}}\ P(a|e) * P(e) / P(a) \tag{4.3}$$

Since a is fixed, we omit it from the From equation 4.1 and 4.2, we get the noisy channel equation, which is:

$$\hat{e} = \underset{e}{\text{argmax}}\ P(a|e) * P(e) \tag{4.4}$$

Where **P(a|e)** is the translation model and **P(e)** is the language model for a given text **a.**

Given two different languages, another way of describing the approach is [13]:

The essence of the method is the alignment of sentences in the two languages and the calculation of the probabilities that any one word is a sentence of one language corresponds to two, one or zero words in the translated sentence in the other language. The probabilities are estimated by matching bigrams (two consecutive words) in each source sentence against bigrams in equivalent target sentence.

For the purpose of this study, statistical machine translation was taken as an approach. The tools and other necessities that are essential to accomplish using this approach were implemented and will be explained as follows.

4.3. Architecture of the System

The following figure entails the main architecture of the system and the succeeding subsections explain the exact meaning of the architecture in detail.

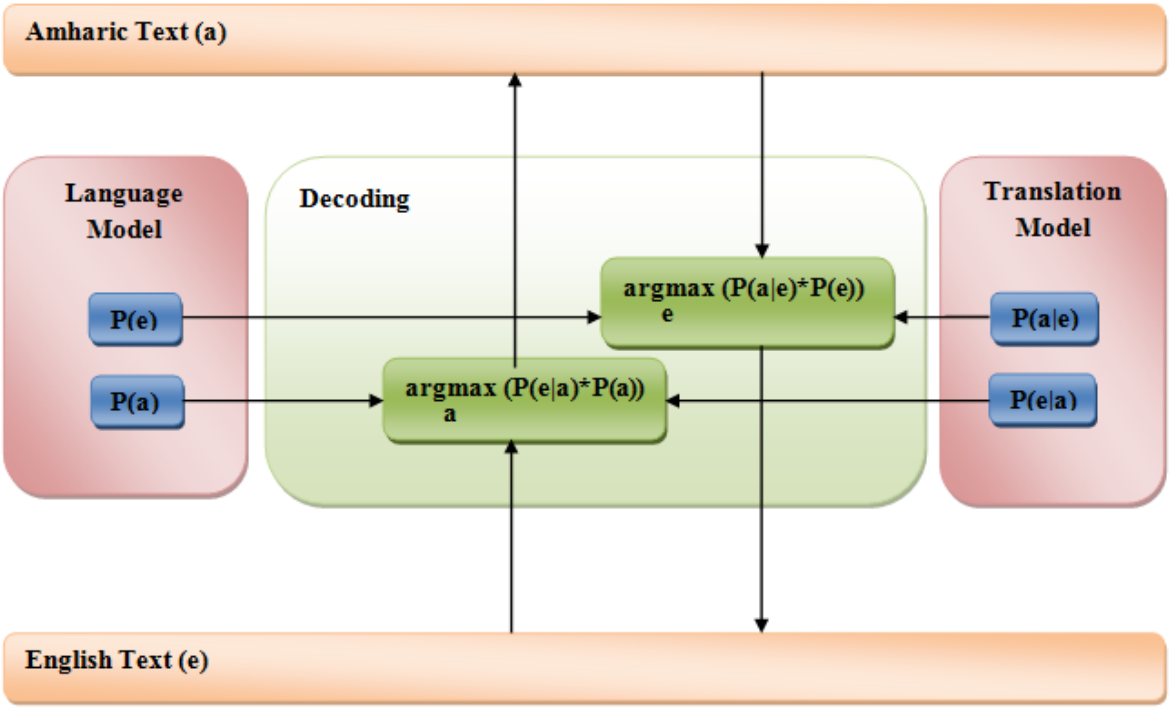


Figure 4.3. Architecture of the system

4.3.1. Source and Target Sentences

The source and target sentences in this study are both the English and Amharic languages. As it has been depicted in figure 4.3, the decoding algorithm takes both the languages as an input and produces both the languages as an output, meaning, if the English text is taken as an input, the Amharic text will be the output and vice versa. This is so because the translation is bidirectional.

4.3.2. Language Model

Language model tries to ensure that words come in the right order including some concept of grammar. Given an Amharic string **a**, the language model assigns **P(a):**

- Good **a** implies high **P(a)**
- Bad **a** implies poor **P(a)**

P(e) and **P(a)** in figure 4.3 represent the language model of the English text and the Amharic text, respectively. The language model can be calculated with a statistical grammar or an n-gram language model. N-gram model was used for the purpose of the study. N-gram corpus is computed from monolingual corpus. The probabilities obtained from the n-gram model could be unigram, bigram, trigram or higher order n-grams. Given the following Amharic sentences፡

አበበ በሶ በላ

አበበ በሶ ጠጣ

አበበ ሱቅ ሄደ

ማርታ ሻይ ጠጣች

አስቴር ቡና ገዛች

The bigram probability can be computed by:

$$P(\mathbf{a_1}) = \frac{\text{count}(\mathbf{a_1})}{\text{Total words observed}} \Rightarrow \quad P(\text{አበበ}) = \frac{3}{15} = \mathbf{0.2}$$

where 3 is the number of times the word 'አበበ' was used and 15 is the total words in the above example.

The bigram probability can be computed by:

$$P(\mathbf{a_2}|\mathbf{a_1}) = \frac{count(\mathbf{a_1}\mathbf{a_2})}{count(\mathbf{a_1})} \Rightarrow \quad P(በሶ|አበበ) = \frac{count(አበበ\ በሶ)}{count(አበበ)} = \frac{2}{3} = \mathbf{0.667}$$

where 2 is the number of times the words 'አበበ' and 'በሶ' have been used together and 3 is the number of times the word 'አበበ' is used.

And the trigram probability becomes:

$$P(\mathbf{a_3}|\mathbf{a_1}\mathbf{a_2}) = \frac{count(\mathbf{a_1}\mathbf{a_2}\mathbf{a_3})}{count(\mathbf{a_1}\mathbf{a_2})} \Rightarrow \quad P(በላ|አበበ\ በሶ) = \frac{count(አበበ\ በሶ\ በላ)}{count(አበበ\ በሶ)}$$

$$P(በላ|አበበ\ በሶ) = ½ = \mathbf{0.5}$$

where 1 is the number of times 'አበበ', 'በሶ' and 'በላ' have been used together, 2 is the number of times 'አበበ' and 'በሶ' were used and $\mathbf{a_1}$, $\mathbf{a_2}$ and $\mathbf{a_3}$ represent the words 'አበበ', 'በሶ' and 'በላ' respectively.

The language model models the target language, that is, if the input is English and the output to be produced is Amharic, it models the Amharic language. For this study, four language models were developed using the whole target corpus. Meaning, two language models for the simple sentences and two for the complex since both language serve as a target language.

For the corpus with simple sentences, the n-gram model performs well with the unigram, bigram and trigram models since the words in the sentence are not that long. But a problem exists if the sentences are too long and the solution would be smoothing which is avoiding zero probability. What we mean by avoiding zero probability is no matter how long the decimal gets, it shouldn't be approximated to zero. Giza++ performs various techniques of smoothing.

4.3.3. Decoding

Decoding is a searching problem that can be reformulated to search for the shortest path in an implicit graph. A decoder searches for the best sequence of transformations that translates source sentence to the corresponding target sentence. It looks up all translations of every source word or phrase, using word or phrase translation table and recombine the target language phrases that maximizes the translation model probability multiplied by the language model probability, which is,

$\underset{\mathbf{a}}{\mathbf{argmax}}$ **(P(e|a)*P(a)),** taking English as an input and displaying Amharic as an output

$\underset{\mathbf{e}}{\mathbf{argmax}}$ **(P(a|e)*P(e)),** taking Amharic as an input and displaying English as an output

as illustrated in figure 4.3.

4.3.4. Translation Model

The translation model assigns a probability that a given source language sentence generates target language sentence. As mentioned above, for a given source and target sentences **E** and A, it is the way sentences in **E** get converted to sentences in **A** which is denoted by **P(a|e)**. It is calculated as:

$$\mathbf{P(a|e)} = \frac{\text{count } (\mathbf{a,e})}{\text{count } (\mathbf{e})} \qquad \mathbf{(4.5.)}$$

The above equation is impossible to achieve because sentences are novel, so we could never have enough data to find values for all sentences. A solution is to decompose the sentences into smaller chunks, as in language modeling.

$$\mathbf{P(e|a)} = \sum_{\mathbf{g}} \mathbf{P(g, e|a)} \qquad \mathbf{(4.6.)}$$

The variable **g** represents alignments between the individual chunks in the sentence pair where the chunks in the sentence pair can be words or phrases. In word-based translation, the fundamental unit of translation is a word. Phrase-based translations, most commonly used, translates whole sequences of words (called blocks or phrases), where the lengths may differ in which blocks are not linguistic phrases but phrases found using statistical methods from corpora. The alignment probability **P(*a*,a|e)** is defined as follows:

$$\mathbf{P}(a,\mathbf{a}|\mathbf{e}) = \prod_{j=1}^{m} \mathbf{t}(\mathbf{a}_j|\mathbf{e}_i) \qquad \textbf{(4.7.)}$$

Where $\mathbf{t}(\mathbf{a}_j|\mathbf{e}_i)$ is the translation probability and it is calculated by counting:

$$\mathbf{t}(\mathbf{a}_j|\mathbf{e}_i) = \frac{\text{count } (\mathbf{a}_j,\mathbf{e}_i)}{\text{count } (\mathbf{e}_i)} \qquad \textbf{(4.8.)}$$

The following table is an example of a word translation table.

table 4.1. word translation table

	Abebe	ate	beso	.
አበበ	■			
በሶ			■	
በላ		■		
።				■

As we can see from the above table, the words and the punctuation mark are aligned perfectly which is difficult to get word aligned data to compute word translation probability. In order to perform this translation probability, we need to have an indirect solution which is expectation maximization which will be discussed under subsection 4.5.2.

4.4. The Corpus

Two different corpora have been developed for the purpose of this study. The first corpus was made of simple sentences and the other corpus was taken from the Bible and directives since both the sources contain both the English language and its corresponding Amharic.

4.4.1. Corpus Collection

The corpus, English-Amharic corpus based on simple sentences, was prepared manually. It entails two different files, one for the English and the other for the corresponding Amharic. It consists of around 1020 simple sentences for each language. Sample text from the corpus prepared for this study is illustrated on Appendix V.

The second corpus was made from two different notions. The first one was from the Bible, Iota version 2.3.3. It entails two main parts, Old Testament and New Testament in which these parts contain different subparts. For this study, one subsection was taken from Old Testament, which is Genesis which also contains fifty other subcategories. It contains around 1537 relatively long sentences for each language. The other selected data was from Public Procurement directive of Ministry of Finance and Economic Development. The English directive as well as the equivalent Amharic was used for the purpose of this study which entails 1951 complex sentences for each language.

4.4.2. Corpus Verification

Since the corpus consisting of the simple sentences was made from scratch by the researcher, it needed to be checked by a certified linguist to identify that the sentences were correct. The other corpus though, is done by professionals since it is made for the use of the public. Therefore, only the first corpus was verified by a linguist.

4.5. Designing Methodology

As mentioned earlier, the approach used for this study is statistical machine translation. Since the research work follows this approach, it needs a Statistical machine translation system, a language modeling toolkit as well as a word alignment tool which will be discussed in the subsections below.

4.5.1. SMT System

Statistical Machine Translation as a research area started in the late 1980s with the Candide project at IBM [15]. IBM's original approach maps individual words to words and allows for deletion and insertion of words.

Moses is an SMT system that was used for this study. It is a system that automatically trains translation models for any language pair. An efficient search algorithm finds quickly the highest probability translation among the exponential number of choices. Moses provides the following features [15]:

- It offers two types of translation models: phrase-based and tree-based.
- It features factored translation models, which enable the integration linguistic and other information at the word level.
- It allows the decoding of confusion networks and word lattices, enabling easy integration with ambiguous upstream tools, such as automatic speech recognizers or morphological analyzers.

Moses is an open source project that is at home in the academic research community. The decoder was originally developed for the phrase model proposed by Marcu and Wong. At that time, only a greedy decoder was available which is a hill-climbing algorithm that does not work one word at a time. Now, it is being developed as a reference implementation of state-of-the-art methods in statistical machine translation [15].

Moses is a cutting-edge machine translation program that reflects the latest developments in the area of statistical machine translation research. It can be trained to translate between any two languages, and yields high quality results.

4.5.2. Word Alignment

Word alignment is the natural language processing task of identifying translation relationships among the words in a text, resulting in a bipartite graph between the two sides of the text, with an arc between two words if and only if they are translations of one another. Word alignment is typically done after sentence alignment has already identified pairs of sentences that are translations of one another.

Word alignment is an important supporting task for most methods of statistical machine translation; the parameters of statistical machine translation models are typically estimated by observing word-aligned texts, and conversely automatic word alignment is typically done by choosing that alignment which best fits a statistical machine translation model [15]. For reasons of making hardware simpler, words are often stored at word aligned addresses. Word-aligned means the address is stored at an address that's divisible by 4 [20].

First, the parallel corpus is aligned bidirectional. This generates two word alignments that have to be reconciled. If the two alignments are intersected, a high-precision alignment of high-confidence alignment points will be acquired. If the union of the two alignments is taken, a high-recall alignment with additional alignment points will be obtained.

The most common algorithm to establish a word alignment is Expectation Maximization Algorithm which will be briefly discussed in the next subtopic.

Expectation Maximization

An expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. In short,

> Expectation step: applies model to the data and assigns probability to possible values using the model.
>
> Maximization step: estimates model from the data by taking assigned values as fact and collecting counts.
>
> Iterating these steps until convergence is reached.

let's take the following phrase as an example:

> *White house*

Step 1. set parameter values uniformly,

> *t(ነጭ|house)=1/2*
>
> *t(ቤት|house)=1/2*
>
> *t(ነጭ|white)=1/2*
>
> *t(ቤት|white)=1/2*

Step 2. compute **P(*a*,a|e)** for all alignments,

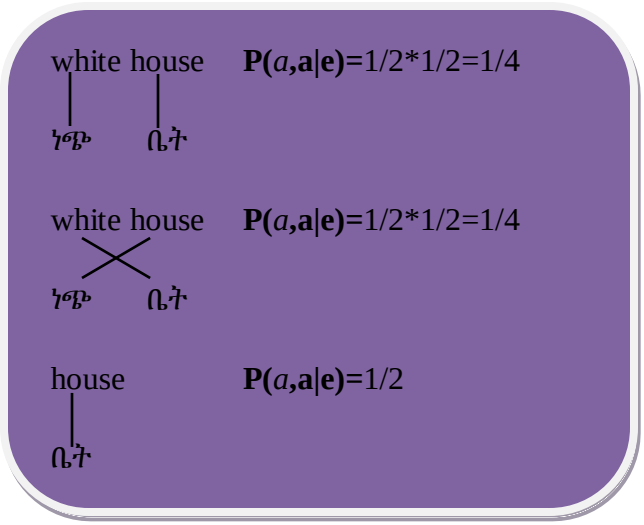


Figure 4.4. computing **P(*a*,a|e)**

Step 3. normalize **P(*a*,a|e)** values to yield **P(*a*|e,a)** values

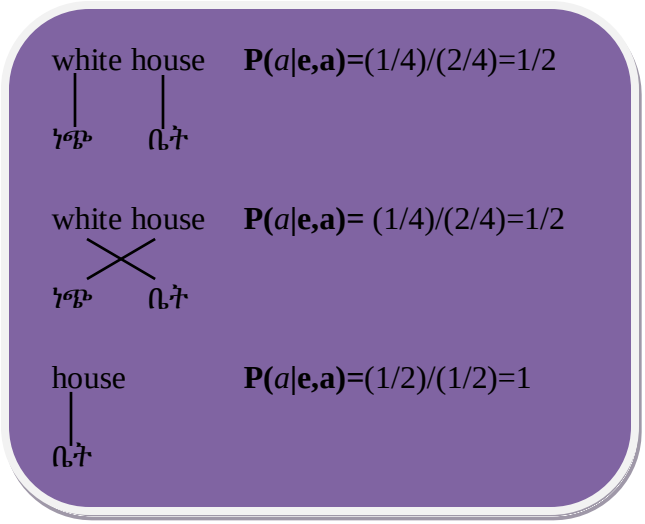


Figure 4.5. normalizing **P(*a*|e,a)**

In the third alignment, since there is only one alignment, **P(*a*|e,a)** will always be 1.

Step 4. collect fractional counts.

t(ነጭ|house)=1/2

t(ቤት|house)=1/2 + 1 = 3/2

t(ነጭ|white)=1/2

t(ቤት|white)=1/2

Step 5. normalize fractional counts to get revised parameter values

t(ነጭ|house)=(1/2)/(4/2)=1/4

t(ቤት|house)=(3/2)/(4/2)=3/4

t(ነጭ|white)=(1/2)/1=1/2

t(ቤት|white)=(1/2)/1=1/2

Repeat step 2. compute **P(*a*,a|e)** for all alignments,

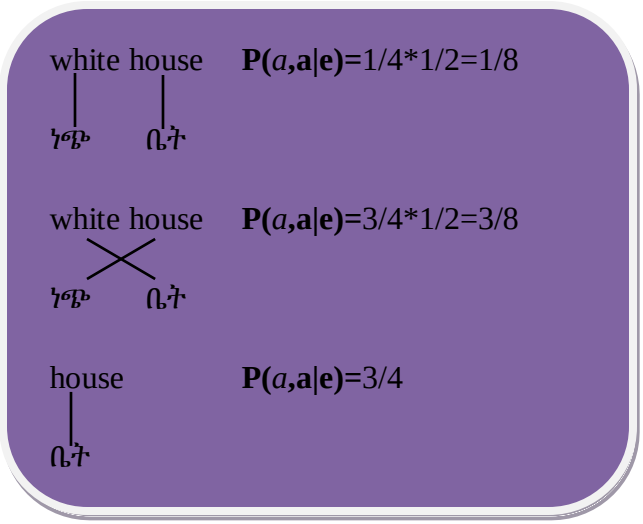


Figure 4.6. computing **P(*a*,a|e)**

Repeat step 3. normalize **P(*a*,a|e)** values to yield **P(*a*|e,a)** values

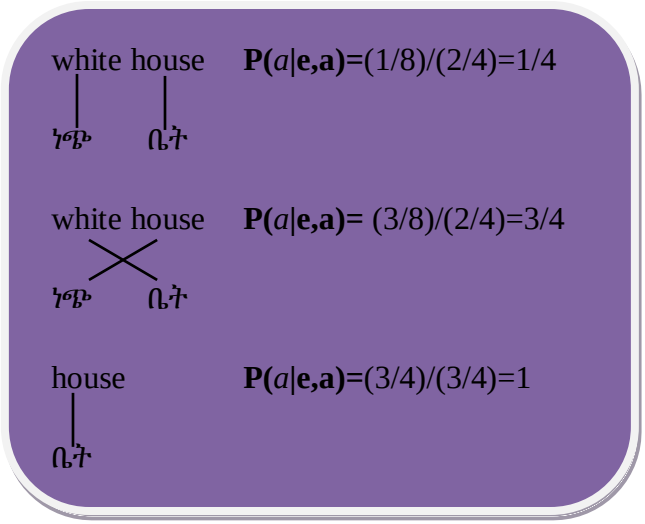


Figure 4.7. normalizing **P(*a*|e,a)**

Repeat step 4. collect fractional counts.

t(ነጭ|house)=1/4

t(ቤት|house)=3/4 + 1 = 7/4

t(ነጭ|white)=3/4

t(ቤት|white)=1/4

Repeat step 5. normalize fractional counts to get revised parameter values

t(ነጭ|house)=(1/4)/(4/2)=1/8

t(ቤት|house)=(7/4)/(4/2)=7/8

t(ነጭ|white)=(3/4)/1=3/4

t(ቤት|white)=(3/4)/1=1/4

Repeating step 2-5 yields:

t(ነጭ|house)=0.0001

t(ቤት|house)=0.9999

t(ነጭ|white)=0.9999

t(ቤት|white)=0.0001, which means the probability of 'ቤት' being 'house' and 'ነጭ' being 'white' increased.

4.5.3. Language Model

There are different Language Modeling Toolkits that are open-source. SRILM, IRSTLM, CMU SLM or KenLM are developed to be integrated with the SMT system, Moses. From the above, IRSLTM will be used as a language modeling toolkit for the purpose of this study.

The IRST Language Modeling Toolkit features algorithms and data structures suitable to estimate, store, and access very large LMs. The software has been integrated into a popular open source SMT decoder called Moses.

LM estimation starts with the collection of n-grams and their frequency counters. Then, smoothing parameters are estimated for each n-gram level; infrequent n-grams are possibly pruned and, finally, a LM file is created containing n-grams with probabilities and back-off weights. This procedure can be very demanding in terms of memory and time if it is applied on huge corpora. But a way has been provided to split LM training into smaller and independent steps, which can be easily distributed among independent processes. The procedure relies on a training script that makes little use of computer RAM and implements the Witten-Bell smoothing method in an exact way [16].

This toolkit supports three output formats of LMs. These formats have the purpose of permitting the use of LMs by external programs. External programs could in principle estimate the LM from an n-gram table before using it, but this would take much more time and memory. So the best thing to do is to first estimate the LM, and then compile it into a binary format that is more compact and that can be quickly loaded and queried by the external program. The three output formats are [16]:

1. **ARPA format**: was introduced in DARPA ASR evaluations to exchange LMs. ARPA format is also supported by the SRI LM Toolkit. It is a text format which is rather costly in terms of memory. There is no limit to the size n of n-grams.

2. **qARPA format**: extends the ARPA format by including codebooks that quantize probabilities and back-off weights of each n-gram level.
3. **iARPA format**: an intermediate ARPA format in the sense that each entry of the file does not contain in the first position the full n-gram probability.

4.6. Steps Undertaken

As mentioned earlier, this research planned to use Moses decoder, GIZA++, IRST Language Modeling Toolkit for building the bidirectional English Amharic Machine Translation. They are all open source tools and the latest version available was used. All the tools have been downloaded, installed and implemented as follows.

4.6.1. Installation

The basic tool, as could be easily identified, is Moses. In order to install it in Windows platform, a software called "CYGWIN" needs to be installed. Cygwin uses packages to manage installing various software. After installing it, it helps download other software necessary and installs it. After some steps were performed, the Moses developers recommended using the Linux Platform rather Windows Platform because it was not tested very well.

A VMware workstation, version 7.1.4, has been chosen to be installed in the windows platform rather than formatting the computer with Ubuntu. VMware is a software which installs different workstations and takes up 8GB of the computer's memory for each workstation installed. After installing the VMware, Debian 5 was installed. Debian is a computer operating system composed of software packages released as free and open source software primarily under the GNU General Public License along with other free software licenses and it was one of the earlier Linux distributions to compose it from packages. But after performing almost all steps necessary, the system could not proceed because the Moses system was not successfully installed. The Moses developers again recommended using Ubuntu because it has been tested on Ubuntu perfectly. Then, the process was started all over again installing Ubuntu 11.04.

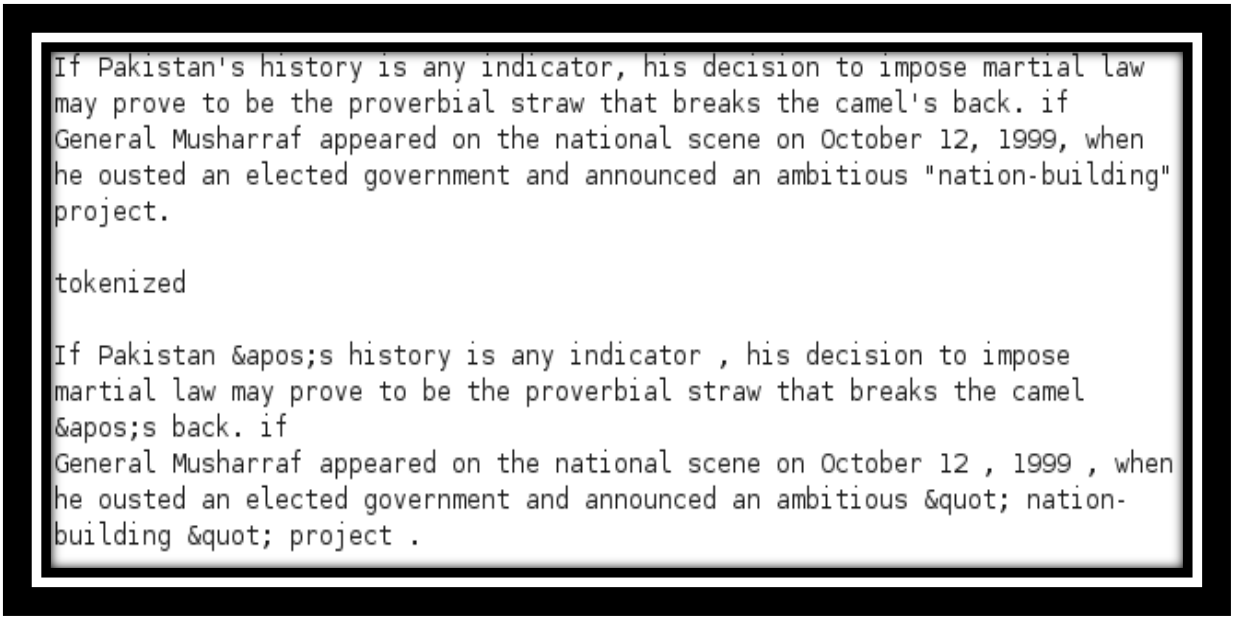
Once the Linux Platform has been installed, installing the tools was the only thing left to do so as to get to the real task. In order to install the tools, the 'terminal' was used. The Terminal in Linux is like CMD in Windows where commands were used to execute each and every operation. First, Moses decoder was installed. To install Moses, boost must be installed first. Then GIZA++ was installed. GIZA++ incorporates two tools, one GIZA++ itself and the other MKCLS which is a tool used to train word classes. At last, IRST Language Modeling Toolkit was installed. The last step done was integrating GIZA++ and IRSTLM in to Moses which makes the installation task completed.

4.6.2. Corpus Preparation

The corpus was prepared in a format that needs to be applicable for the translation. To prepare the data for training the translation system, the following steps need to be performed [17]:

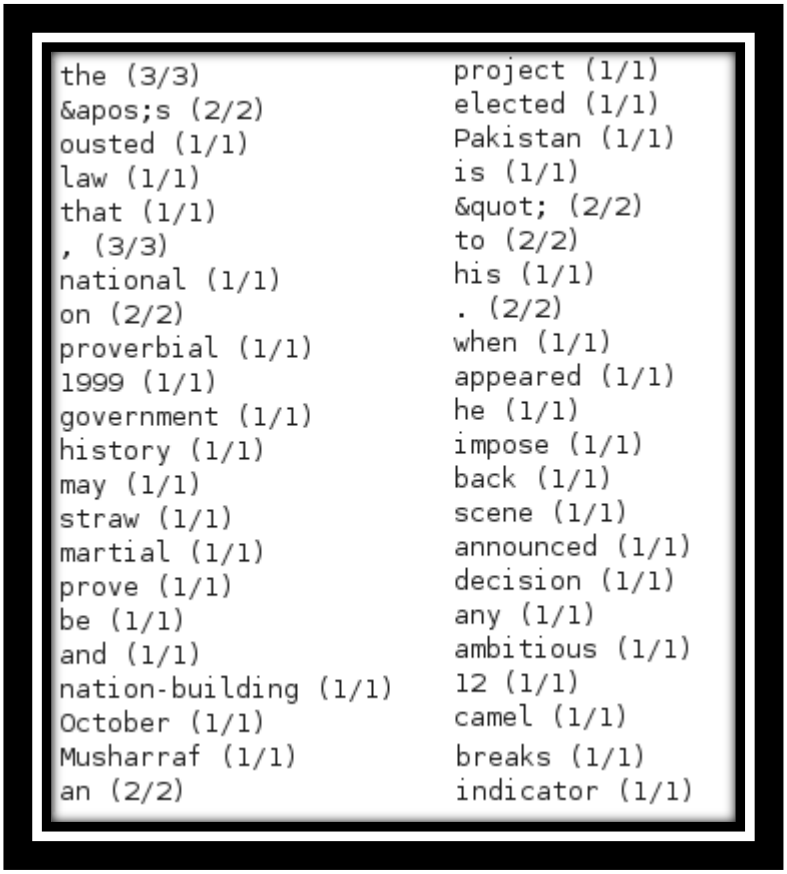
1. Tokenization: This means that spaces have to be inserted between words and punctuation.
2. True-casing: The initial words in each sentence are converted to their most probable casing. This helps reduce data being sparse.
3. Cleaning: Long sentences and empty sentences are removed as they can cause problems with the training pipeline, and obviously misaligned sentences are removed.

For example, let's look at the following paragraph in English.

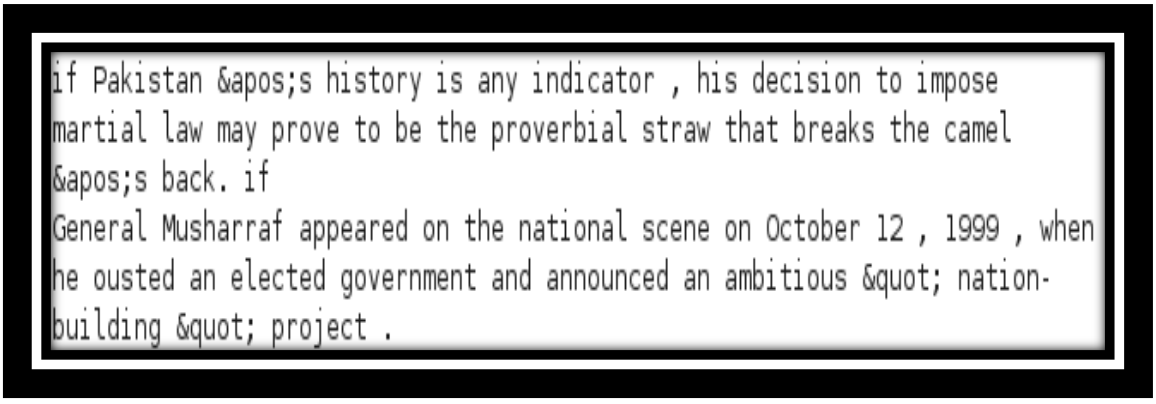


As it can be clearly seen, the tokenized paragraph inserts space between words and punctuation marks. The apostrophe and the quotation marks are marked in a different way so it can be easy to identify.

The true casing first requires training, in order to extract some statistics about the text. In the first step, it tries to identify how many times a particular word has been used in the corpus and the output is as follows:



Another script from the Moses distribution was used for the true-casing. And the result obtained was:



As it can be observed, the true-cased paragraph is the almost same as the tokenized paragraph. This is so because, most of the words are used only once and those that are used more than once have the same case. The only word that has a difference is ‘if’. In the tokenized paragraph, the first word ‘If’ has a capital letter ‘I’ but in the true-cased paragraph, the letter ‘I’ became a small letter. The cleaning step needs a longer sentence to remove, usually a sentence that has more than 80 words. So, it can’t work for the above paragraph.

4.6.3. Language Model Training

The language model is used to ensure fluent output, so it is built with the target language, that is, for Amharic as well as English separately since it is bidirectional translation both the languages become a target language at some point. As it has been stated, IRST Language modeling toolkit was used. An appropriate 3-gram language model, removing singletons, smoothing with improved Kneser-Ney and adding sentence boundary symbols were built.

4.6.4. Training the Translation System

At this step, word-alignment, phrase extraction and scoring were used and lexicalized reordering tables and Moses configuration file were created with. Mainly this step creates a ‘moses.ini’ file, which is used for decoding and the phrase table is also created which basically contains the probabilities of a word following another.

4.6.5. Tuning

As mentioned in the above subsection, while training the translation system, a 'moses.ini' file was produced which is used for decoding. The querying process could be started right away but the weights used by Moses to weight the different models against each other are not optimized. Therefore, to find better weights, the translation system needs to be tuned. This process produces another ‘.ini’ file that is used for decoding. The step that is followed after this will be the testing process, the final stage, that will be discussed in the next chapter.

4.7. Prototype of the System

This section will try to demonstrate the prototype of the system when it is queried to translate from English to Amharic.

The following figure displays the output when it was queried to translate the simple sentence:

i am sick

```
Defined parameters (per moses.ini or switch):
        config: /home/eleni/forlinux/boost_1_45_0/mosesdecoder/sample-models/phr
ase-model/moses.ini
        input-factors: 0
        lmodel-file: 8 0 3 lm/europarl.srilm.gz
        mapping: T 0
        n-best-list: nbest.txt 100
        ttable-file: 0 0 0 1 phrase-table
        ttable-limit: 10
        weight-d: 1
        weight-l: 1
        weight-t: 1
        weight-w: 0
Loading lexical distortion models...have 0 models
Start loading LanguageModel lm/europarl.srilm.gz : [0.000] seconds
Loading Internal LM: lm/europarl.srilm.gz
lm/europarl.srilm.gz
Finished loading LanguageModels : [0.000] seconds
Start loading PhraseTable phrase-table : [0.000] seconds
filePath: phrase-table
Finished loading phrase tables : [0.000] seconds
IO from STDOUT/STDIN
Created input-output object : [0.000] seconds
Translating: i am sick

Collecting options took 0.000 seconds
Search took 0.000 seconds
አማርኛ
BEST TRANSLATION: አማርኛ [111]  [total=-201.204] <<0.000, -1.000, 0.000, -200.000,
 -1.204>>
Translation took 0.000 seconds
Finished translating
```

Figure 4.3. The course of action to translate the sentence ‘I am sick’

As it can be seen from figure 4.3., it took 0.000 seconds to display the translation of the English sentence to the Amharic "አሞኛል". This is so, because the sample corpus was so small and the sentence that has been queried was already in that sample corpus. Therefore, as the size of the corpus becomes larger, the time it takes to acquire the translation increases. The time it takes to obtain the translation also increases as the query becomes sparse within the corpus.

The next figure, that is figure 4.4., tries to translate the query:

እነሱ እየተጫወቱ ነው

```
Defined parameters (per moses.ini or switch):
        config: /home/eleni/working-ss/am-en/mert-work/moses.ini
        distortion-file: 0-0 wbe-msd-bidirectional-fe-allff6 /home/eleni/working-ss/am-
en/train/model/reordering-table.wbe-msd-bidirectional-fe.gz
        distortion-limit: 6
        input-factors: 0
        lmodel-file: 8 0 3 /home/eleni/lm/enamt-ss.en-am.blm.en
        mapping: 0 T 0
        ttable-file: 0 0 0 5 /home/eleni/working-ss/am-en/train/model/phrase-table.gz
        ttable-limit: 20
        weight-d: 0.00693393 0.123204 0.0783404 0.0997544 0.077088 0.171961 0.128379
        weight-l: 0.0182902
        weight-t: 0.0843785 -0.0311639 0.0767146 -0.0360695 -0.0213065
        weight-w: 0.0464166
/home/eleni/mosesdecoder/bin
Loading lexical distortion models...have 1 models
Creating lexical reordering...
weights: 0.123 0.078 0.100 0.077 0.172 0.128
Loading table into memory...done.
Start loading LanguageModel /home/eleni/lm/enamt-ss.en-am.blm.en : [0.170] seconds
Finished loading LanguageModels : [0.173] seconds
Start loading PhraseTable /home/eleni/working-ss/am-en/train/model/phrase-table.gz : [0.173] seconds
filePath: /home/eleni/working-ss/am-en/train/model/phrase-table.gz
Finished loading phrase tables : [0.173] seconds
Start loading phrase table from /home/eleni/working-ss/am-en/train/model/phrase-table.gz : [0.173]
seconds
Reading /home/eleni/working-ss/am-en/train/model/phrase-table.gz
----5----10----15----20----25----30----35----40----45----50----55----60----65----70----75----80----85----90----95----100
**********************************************************************************************
**********
Finished loading phrase tables : [0.253] seconds
IO from STDOUT/STDIN
Created input-output object : [0.254] seconds
Translating line 0  in thread id 139653580052224
Translating:  እነሱ እየተጫወቱ ነው

Line 0: Collecting options took 0.001 seconds
Line 0: Search took 0.000 seconds
they are playing
BEST TRANSLATION: they are playing [111]  [total=-0.442] <<0.000, -3.000, 0.000, -1.143, 0.000, 0.000, -
1.379, 0.000, 0.000, -8.350, 0.000, -1.853, -0.057, -2.404, 2.000>>
Line 0: Translation took 0.002 seconds total
```

Figure 4.4. The course of action to translate the sentence 'እነሱ እየተጫወቱ ነው'

As it can be seen, the translation produced is 'they are playing'. The sentence 'እየተጫወቱ ነው' could also be translated as 'they are playing'. The word 'እነሱ' was used so as the system could be able to identify which Amharic word stands for which English word.

CHAPTER FIVE
EXPERIMENT

5.1. Introduction

This study is based on Bidirectional English Amharic Machine Translation. Therefore, in order to translate English to Amharic, a series of steps were followed as mentioned in the last chapter. After training the language model and the translation system, the next step is to produce a translation using the decoder. This chapter explains the experiments conducted, and the analysis and discussions made based on the findings of the experiments.

5.2. Methodologies for Testing

Two methodologies were used to test the system. The first one is BLEU Score and the second methodology used is preparing a questionnaire manually. The following subtopics discusses on each of the methods briefly.

5.2.1. BLEU Score

BLEU (Bilingual Evaluation Understudy) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" - this is the central idea behind BLEU [23]. BLEU was one of the first metrics to achieve a high correlation with human judgments of quality, and remains one of the most popular automated and inexpensive metrics.

Scores are calculated for individual translated segments, generally sentences, by comparing them with a set of good quality reference translations. Those scores are then averaged over the whole corpus to reach an estimate of the translation's overall quality. This was the first methodology used for the testing process.

5.2.2. Questionnaire

The second methodology followed was using a manual technique. The questionnaire consisted of all the testing dataset and the translation acquired. It has an evaluation method on the scale of 1 to 5 in which if a candidate gives 5, it means that the translation was perfect and if 1 is given, it means it has a very poor translation. Thirteen candidates were chosen to fill the questionnaire in which all are degree holders or above from a well-known university or college.

Two questionnaires were prepared for the test set i.e. from English to Amharic and from Amharic to English. The questionnaires developed were different because the result obtained from Amharic to English and from English to Amharic were not similar.

5.3. Corpus

As it was discussed in Chapter One, the main objective of this study was to design and develop a statistical machine translation system using an English-Amharic corpus. Hence, the experimentation began with checking whether it could produce an accurate translation provided the corpus.

The expectation maximization was going to be developed for accurate word alignment but GIZA++, which implements this algorithm for the purpose of aligning words, was already provided as an open source tool. Therefore, this tool was used for the word alignment procedure.

As mentioned in Chapter Four, two different corpora were prepared for the purpose of this research work, one with simple sentence and the other with complex sentence. Consequently, different experimentation was taken for both the corpora. For the purpose of the following topics, the corpus with the simple sentences will be called Corpus I and the corpus with the complex sentences will be known as Corpus II.

5.3.1. Corpus I

As it has been tried to be illustrated earlier, Corpus I was made of about 1020 simple sentences that had been prepared manually. And again those sentences were verified by a certified linguist in order for them to be accurate. The experimentation process for Corpus I was classified in to two, one called training set and the other called test set.

For Corpus I, the translation system was trained both ways, that is, from English to Amharic and from Amharic to English. The same steps were followed for the translation process. And the same training and test sets will be used to test which way is more effective than the other.

Out of the 1020 simple sentences, all the sentences were used for the training set. For the test set, the sample text was prepared manually. The sentence in the sample is not going to be the same as the sentence in the corpus but the phrases will be taken from the sample text that has been formulated from the corpus for the experiment on the training set. For example, let's take the following two sentences from the corpus:

Aster watches football, and

Almaz plays football

The sentence that is going to be produced for the sample data will be something like:

Almaz watches football, and the same for the Amharic to English as well.

This is so, because the words in a sentence still need to be from the corpus. Otherwise, the decoder couldn't translate a word that is not accessible in the corpus. The sample data for experimenting contains 102 simple sentences which is 10% of the corpus. The questionnaire prepared for the test set is demonstrated on Appendix I and II.

5.3.2. Corpus II

Corpus II was composed of two different sources, one from the Bible and the other from the public procurement directive of Ministry of Finance and Economic Development. Like the experiment on Corpus I, the experimentation process consisted of two methodologies. 1951 complex sentences were found on Corpus II and 2% were taken for the testing process which is around 40sentences. This is so because the complex sentence is very large and complicated which makes it hard for the candidates to assess the translation. The first part was taken from the main corpus itself. Some portion of test set was prepared manually as well. Some phrase from a particular sentence and another phrase from another sentence were taken to formulate the sample texts. The questionnaire that includes the complex sentences can be seen on Appendix III and IV, for the bidirectional translation.

5.4. Result

The result was seen from two perspectives, one from the accuracy point of view and the other from the time it takes to translate a particular sentence. From the experiments taken, the following findings were presented.

5.4.1. Result on Corpus I

Similar to the experiment, the corpora was named corpus I and corpus II. For each experiment taken, the result was recorded. For corpus I, the following result was obtained for the training set as well as the test set.

Result on the Test Set

All the sample texts prepared from the corpus and the test set was queried in to the system and results were produced. There is no doubt that it is going to be less accurate because the sentence cannot be fetched directly from the corpus. The result recorded from the first methodology (BLEU Score) was 82.22% for the English-Amharic translation and 90.59% for the Amharic-English translation. The result recorded from English to Amharic was low mostly because it was hard for the system to identify the feminine and masculine representation. Let’s see the following sentence as an example:

Haile is beautiful

Which is translated as,

ሀይሌ ቆንጆ ነች

Although the name "Haile" is a name for a man, the translation considered it as if it was for a woman. This is because the phrase "*ቆንጆ ነች*" has been used for the feminine perception. But for the Amharic, it doesn't have any problem translating it correctly.

Another example is:

People talk a lot

The corresponding Amharic sentence retrieved was:

ሰዎች በጣም ያወራሉ

Which is a right translation. And when the Amharic sentence was queried, it gave the output:

People talk too much

Which is an alternative correct sentence. From this, we can see that the system has learned other ways of translating a given text to display a target text.

And for the second methodology, the result recorded was 91% for the English-Amharic translation and 97% for the Amharic to English. The result for English to Amharic translation was higher than the earlier result because the candidates gave a normal grade for the simple errors made from the translation and the simple errors from the previous testing method were taken as wrong.

And the time it took for each translation to take place was recorded and for the English-Amharic translation, the highest time it took was 17 microseconds. For the Amharic-English translation as well, the maximum time recorded was 0.009. From these, it could be concluded that the Amharic-English translation produces faster and more accurate results as compared to English-Amharic translation. This is so because Amharic is a language that is morphologically rich and it could produce more accurate results if it is the source language rather than being the target language.

The result retrieved has a high percentage because the test sentences are manually prepared from the corpus itself and the sentences taken for the language modeling were the whole corpus.

5.4.2. Result on Corpus II

Like the findings on Corpus I, the following are the results acquired from corpus II for the merged training and test sets.

Results were obtained from the complex sentence that was taken as a sample text from Corpus II. As mentioned above, forty sentences were taken as a sample, that is, both from the directive and the Bible. For the first methodology, the accuracy of the translation from English to Amharic was 73.38%. And the translation from Amharic to English was 84.12% effective. From this, we can see that the Amharic to English translation is easier for the decoder to interpret for reasons mentioned in subtopic 5.4.1.

While trying to translate the sentences, some errors in the corpus were observed. Let's look at the following sentence:

> *And blessed be the most high God, which hath delivered thine enemies into thy hand. And he gave him tithes of all.*

The translation presented in the corpus was:

> *ጠላቶችህን በእጅህ የጣለልህ ልዑል እግዚአብሔርም የተባረከ ነው አለውም። አብራምም ከሁሉ ዐሥራትን ሰጠው።*

First of all, in English language, it is impossible to consider “most high” as a phrase. The exact word to be used is “highest”. This is an error that was observed which makes the Bible's English very complicated. Let’s look at the last sentence:

And he gave him tithes of all.

That was translated as:

አብራምም ከሁሉ አሥራትን ሰጠው።

The word “he” was represented as “አብራም” which is not the exact translation. “he” in Amharic is “እሱ” and “አብራም” is a name called “Abraham” in English. Although the word “he” represents “Abraham”, it cannot be used this way because the decoder trains “he” as “Abraham” and the next time “he” is used, it might also translated as “Abraham” which is not right. This sentence was translated correctly since the decoder trained the source sentence to be translated to the target but when some sentence which is not directly taken from the corpus is queried, it might not be translated correctly. The word “አብራም” is not correct as well, it was misspelled. The exact word is “አብርሃም” and this also causes another problem because the decoder refers to both words as if they were different.

Some of the errors found were mostly because of the corpus. It doesn’t use consistent translation. Some problems were shown in the Amharic sentence. Since Amharic is a complex language, the same words used in this language could seem different when observed. We might take the following word as an example,

ገፅታ

This word is the same as:

ገጽታ

But when this word is used in the corpus, it does not give the same meaning. If "ገፅታ" is only used in the corpus and if "ገጽታ" was queried, it is definitely obvious that the system is going to label it unknown. But the others which are trained well, perform well. For example, the following was generated from two different sentences in the corpus:

> *if the contract value is below Birr 5,000, the head of the procuring entity may approve the recommendations of the Tender Committee;*

And the exact translation produced was:

> *የግዥው መጠን ከብር 5000.00 በታች ሲሆን፣ የግዥ ፈፃሚው አካል የበላይ ሀላፊ በጨረታ ኮሚቴ የቀረበለትን የውሳኔ ሀሳብ ሲያፀድቅ፣*

The accuracy from the second methodology was 87% for the English to Amharic translation and 89% for the Amharic to English. The performance of the second methodology is higher than the first methodology because the candidates noticed that the sentences with slight error were understandable.

The time it took for each result to be produced is an average of 4.987 seconds. At this step, the time count jumped from milliseconds to seconds because the result acquired is from complex sentences which consist of very long sentences.

5.5. Discussion

This subsection discusses the problems created during the testing process and provides solutions to the established problems. Some of the errors encountered were due to the loss of large corpus. Large corpus could be organized but the English-Amharic documents that are prepared by a linguist are complicated and could not be used for the day-to-day activities and communication of human beings.

The other problems were mostly covered on the result subsection. The solution to be proposed for all these problems is finding or preparing a very large corpus and checking the words and sentences used in each and every line so as to produce a better result. This is very time consuming, but it is a necessity if the translation to be acquired is somewhat error-free and satisfactory. This corpus to be produced must include all the domains of study. And the words used should also be used repetitively so it could be easier to identify which source represents the corresponding target word.

CHAPTER SIX
CONCLUSION AND REOCOMMENDATION

6.1. Conclusion

The purpose of this study was to design and develop a bidirectional English-Amharic machine translation using constrained corpus that is mainly on simple sentences although its applicability was also tested on complex sentences. In this research work, an attempt was made to describe how to develop a bidirectional machine translation using the statistical machine translation approach. A corpus was prepared and collected so as it could be used for the machine translation process.

The study began with a brief discussion on the language Amharic and how it differs from English. It described the phrasal categories as well as the sentence structure of Amharic and how the sentence structure affects the translation process with English. It also explained the articles, punctuation marks and conjunctions that are used in both languages.

The research work continued on elaborating the role that NLP plays in enhancing computers' capability to process natural language. In this discussion, it is indicated that all applications of NLP have the common objective of understanding and extracting meaning from a natural language input. This process involves transforming the natural language into a form where the meaning is explicit and is easily usable by the application program. As a way to this end, machine translation, a process which converts texts from one natural language to another, was discussed as one task in the step towards achieving the stated objective.

Also discussed, is the main part of the study, designing and development of the system. This part entails the process and procedure followed to accomplish the main task of the study. Moses was used as a statistical machine translation system. And the corpus was taken and put through all the necessary steps to be pursued for the training of the language model and translation system. As a language modeling toolkit, IRSTLM was used. And in

order for the words to be aligned perfectly, Expectation maximization algorithm was used. All the steps were followed and four models were formulated which translate simple English sentences into Amharic, simple Amharic sentences into English, complex English sentences into Amharic and complex Amharic sentences into English.

Experiments were taken and results were recorded for all translation. And the results obtained were, all in all, accurate using BLEU Score methodology and preparing a questionnaire. The result obtained for the simple sentence using BLEU Score had an average of 82.22% accuracy for the English to Amharic, 90.59% for the Amharic to English and for the complex sentences, the result acquired was approximately 73.38% for the English to Amharic, 84.12% for the Amharic to English. From the questionnaire method, the accuracy from English to Amharic was 91% and from Amharic to English was 97% for the simple sentences and from English to Amharic was 87% and from Amharic to English was 89% for the complex sentences. And the maximum time taken for each translation to be carried out is 17 microseconds and 4.987 seconds, for the simple sentences and complex sentences respectively. The result recorded was somehow high because the test set taken was from the corpus itself and the whole corpus was used for language modeling.

From this, we can see that, given a large corpus, a proper result will be produced with a suitable time limit. And since the BLEU Score is not that much different as compared to the manual questionnaire preparation method, we can also use both the methods for the evaluation.

6.2. Recommendation

The corpus taken for this study cannot be enough and a representative of the language, and future researches should be conducted using a larger set of corpus. A large, well written and appropriately review corpus should be conducted so as to generate a flawless translation.

The following areas could be explored further as a continuation of this study.

- Further researches in machine translation on Amharic to other languages, even using languages in Ethiopia such as Tigrigna, Oromifa or so could be performed while preparing a large corpus.
- It can be enhanced to handle larger set of complex sentences in the language, and to develop a full-fledged bidirectional English Amharic Machine Translation.
- Morphological analyzers and synthesizers should be developed for Amharic and used for the translation purpose. This method decreases the size of the corpora to be used which is a magnificent idea since the language is very complex; it breaks it into pieces and makes it easier to be translated.
- Since this system could be enhanced easily with a larger corpus, speech to text translation could be developed. It is easier to develop this system because the text to text translation is already available and at hand.

Reference

[1] *Ethnologue, languages of the world*

[2] *Wikipedia, the free encyclopedia*

[3] W.John Hutchins, *Machine translation: A brief history,* 1995

[4] *Ethiopian Treasures,* http://www.ethiopiantreasure.co.uk

[5] *Omniglot, the online encyclopedia of writing systems and languages*

[6] S.Amsalu and S. Fissaha Adafre, *Machine translation for Amharic: where we are,* 2006

[7] *Natural Language Processing, articles on Natural Language Processing*

[8] C. Monson, A. Fort Llitjos, R. Aranovich, L. Levin, R. Brown, E. Peterson, J. Carbonell, A. Lavie, *Building NLP systems for two resource-scarce indigenous languages: Mapudungun and Quechua,* 2006

[9] *SYSTRAN, the leading supplier of language translation software*

[10] S. Adugna and A. Eisele, *English-Oromo Machine Translation: An Experiment Using a Statistical Approach,* May 2010

[11] A. Alemu Argaw and L. Asker, R. Coster and J. Karlgren, *Dictionary-Based English-Amharic Information Retrieval,* September 2004

[12] M. Forcada, Machine Translation Marathon, Dublin, *Apertium: Free/Open Source Rule-Based Machine translation*, Jan. 29, 2010

[13] Prof. A. Homiedan, *Machine Translation*, 2010

[14] A. Ramanathan under the guidance of Prof. P. Bhattacharyya and Dr. M. Sasikumar, *Seminar Report, Statistical Machine Translation,* 2002

[15] P. Koehn, *Moses, Statistical Machine Translation System User Manual and Code Guide,* 2012

[16] M. Federico, N. Bertoldi, M. Cettolo. Trento, Italy, *IRST Language Modeling Toolkit, user manual,* September 11, 2008

[17] P. Koehn, *Statistical Machine Translation,* http://www.statmt.org, 2009

[18] C. Boitet, GETA, institute of IMAG, *Perspective of Machine Aided Translation,* 2010

[18] M. Flanagan, *MT in the Online Environment: Challenges and Cooperation,* 1995

[20] *Word Alignment,*
http://www.cs.umd.edu/class/sum2003/cmsc311/Notes/Data/aligned.html

[21] D. Gochel Agonafer, *An integrated approach to automatic complex sentence parsing for Amharic text,* June 2003

[22] J. Hutchins, *Latest Development in Machine Translation Technology: Beginning a New Era in MT Research,* 1999

[23] M. Gebreegziabher & L. Besacier (Prof.), *Preliminary Experiments on English-Amharic Statistical Machine Translation,* 2012

[24] K. Papineni, S. Roukos, T. Ward and W. Zhu, *BLEU: a Method for Automatic Evaluation of Machine Translation,* July 2002

[25] M. Gebreegziabher & L. Besacier (Prof.), *Bilingual Data Mining for the English-Amharic Statistical Machine Translation*, December 2011

Appendices

Appendix I: Questionnaire for the Simple Sentences (English-Amharic)

Grade 1 to 5,
1 - Very poor translation (not related at all),
2- Poor,
3 – Good,
4 - Almost perfect,
5- Perfect translation,

English	Amharic	Point
Haile is beautiful	ሀይሌ ቆንጆ ነች	
Abebe is an addict	አበበ ሱሰኛ ነው	
kalkidan is a student	ቃልኪዳን ተማሪ ነው	
the children are taking an exam	ህፃናቶቹ ፈተና እየወሰደች ናቸው	
he loves enjera	እሱ እንጀራ ትወዳለች	
Abebe is not crazy	አበበ እብድ አይደለም	
Aster quit	ለቀቀች አስቴር	
Hana is getting married	ሀና ልታገባ ነው	
he is a runner	እሱ ሯጭ ነው	
Haile loves competition	ሀይሌ ውድድር ይወዳል	
he has a car	እሱ መኪና አለው	
addis wants to be a president	አዲስ ፕሬዘዳንት መሆን ይፈልጋል	
Meaza wants to be a doctor	መዓዛ ዶክተር መሆን ትፈልጋለች	
she is a doctor	እሷዶክተር ነው	
Mary plays volleyball	ሜሪ መረብ ኳስ ትጫወታለች	
Almaz likes playing volleyball	አልማዝ መረብ ኳስ መጫወት ትወዳለች	
the students are playing	ተማሪዎች ተማሪዎቹ እየተጫወቱ ነው	
Getnet has a doll	ጌትነት አሻንጉሊት አላት	
he is a child	እሱ ህፃን ነች	
she is her child	እሷ የእሷ ልጅ ነው	
Alice is his mother	አሊስ የእሱ እናት ነች	
he loves his mother	እሱ እናቱን ይወዳታል	
John is drinking tea	ጆን ሻይ እየጠጣ ነው	
Nigist is his wife	ንግስት የእሱ ሚስት ነች	
Addisu has been married for two years	አዲሱ ካገባ ሁለት አመት ሆኖታል	
Addisu is an employee	አዲሱ ተቀጣሪ ነው	
Almaz is fast	አልማዝ ፈጣን ነች	
Mary is his wife	ሜሪ የእሱ ሚስት ነች	
Haile is her husband	ሀይሌ የእሷ ባል ነው	
addis has a book	አዲስ መፅሀፍ አላት	
Abebe has a pen	አበበ እስክሪብቶ አለው	
he is divorced	ፈቷል እሱ	

he loves watching movies	እሱ ፊልም ማየት ይወዳል	
she is listening to music	እሷ ዘፈን እያዳመጠች ነው	
Getnet loves listening to music	ጌትነት ዘፈን ማዳመጥ ይወዳል	
Meaza is a student	መዓዛ ተማሪ ነው	
Ayele won the race	አየለ ውድድሩን አሸነፈ	
Ayele won the game	አየለ ጨዋታውን አሸነፈ	
Aster is playing a game	አስቴር ጨዋታ እየተጫወተች ነው	
Aster wants to be a president	አስቴር ፕሬዘዳንት መሆን ትፈልጋለች	
Nigist is a graduate	ንግስት ተመራቂ ነች	
Mary loves my car	ሜሪ የእኔን መኪና ትወደዋለች	
Bob is a teacher	ቦብ አስተማሪ ነው	
Mary is a teacher	ሜሪ አስተማሪ ነው	
she died	እሷ ሞተ	
The teacher is tall	አስተማሪው ረዥም ነው	
Almaz is short	አልማዝ አጭር ነች	
Bob bought a pen	ቦብ እስክሪብቶ ገዛች	
Nigist is pregnant	ንግስት ነፍሰ-ጡር ነች	
Addisu is not crazy	አዲሱ እብድ አይደለም	
Almaz was not happy	አልማዝ ደስተኛ አልነበረችም	
Bob is a good guy	ቦብ ጥሩ ሰው ነው	
Addisu is crying	አዲሱ እያለቀሰ ነው	
Mary went to school	ሜሪ ወደ ትምህርት ቤት ሄደ	
Bob is sweating	ቦብ እያላበው ነው	
kalkidan is fat	ቃልኪዳን ወፍራም ነች	
Haile is thin	ሀይሌ ቀጭን ነው	
the students are worried	ተማሪዎቹ ተጨንቀዋል ተማሪዎች	
people talk a lot	ሰዎች በጣም ያወራሉ	
kalkidan fetched water	ውሀውን ቀዳችው ቃልኪዳን	
Abebe is crazy	አበበ እብድ ነው	
Abebe killed himself	አበበ እራሱን ገደለ	
Haile committed suicide	እራሱን አጠፋ ሀይሌ	
Mary is begging him	ሜሪ እሱን እየለመነችው ነው	
the children are running	ህፃናቶቹ እየሮጡ ነው	
Haile is wounded	ቆስሏል ሀይሌ	
Abebe is angry	አበበ የተናደደ ነው	
Bob is happy	ቦብ ደስተኛ ነው	
kalkidan is a student	ቃልኪዳን ተማሪ ነው	
Addisu is a doctor	አዲሱ ዶክተር ነው	
Addis is an engineer	አዲስ መሀንዲስ ነች	
Bob is an architect	ቦብ አርክቴክት ነው	
Mary loves to clean	ሜሪ ማፅዳት ትወዳለች	
Chuck is a spy	ቸክን ሰላይ ነው	

people are spies	ሰዎች ሰላዮች ናቸው	
Belay is not a spy	በላይ ሰላይ አይደለችም	
Belay is not a doctor	በላይ ዶክተር አይደለም	
the dog died	ውሻው ሞተ	
the cat is eating	ድመቱ እየበላ ነው	
Helen left him	ሄለን ጥላው ሄደች	
Eyob is alone	እዮብ ብቻውን ነው	
Helen is right	ሄለን ልክ ነች	
Eyob is talking	እዮብ እያወራ ነው	
He is walking with Helen	እሱ ከሄለን ጋር እየተራመደ ነው	
Selam run away from him	ሰላም የሆነ ከእሱ ርቃ ሄደች	
he plays guitar	እሱ ጊታር ይጫወታል	
she loves musicians	እሷ ዘፋኞችን ትወዳለች	
he is a musician	እሱ ዘፋኝ ነው	
he is sick	እሱ አሞታል	
he loves Mary	እሱ ሜሪን ያፈቅራታል	
Chuck is a father	ቸክን አባት ነው	
Chuck is stupid	ቸክን ደደብ ነው	
he killed his brother	እሱ ወንድሙን ገደለው	
Eyob's wife is pregnant	እዮብ ሚስት ነፍሰ-ጡር ነች	
Selam is sick	አሞታል ሰላም የሆነ	
he gets drunk everyday	በየቀኑ ይሰክራል እሱ	
Mesfin is a football player	መስፍን እግር ኳስ ተጫዋች ነው	
Marta murdered a guy	ማርታ ሰው ነፍስ አጠፋች	
Brook told Fernado to stop walking	ብሩክ ፈርናንዶን መራመድ እንዲያቆም ነገረው	
Morgan is funny	ሞርጋን የሚያስቅ ልጅ ነው	
James bought a new phone	ጄምስ አዲስ ስልክ ገዛች	
Morgan is scared	ሞርጋን አስፈራራው ነው	

Appendix II: Questionnaire for the Simple Sentences (Amharic-English)

Grade 1 to 5,
1 - Very poor translation (not related at all),
2- Poor,
3 – Good,
4 - Almost perfect,
5- Perfect translation,

Amharic	English	Point
ሀይሌ ቆንጆ ነው	Haile is beautiful	
አበበ ሱሰኛ ነው	Abebe is an addict	
ቃልኪዳን ተማሪ ነው	kalkidan is a student	
ህፃናቶቹ ፈተና እየተፈተኑ ነው	the children students are taking an exam	
እሱ እንጀራ የወዳል	he loves enjera	
አበበ እብድ አይደለም	Abebe is not crazy	
አስቴር ለቀቀች	Aster quit	
ሀና ልታገባ ነው	Hana is getting married	
እሱ ሯጭ ነው	He is a runner	
ሀይሌ ውድድር ይወዳል	Haile loves competition	
እሱ መኪና አለው	he has a car	
አዲስ ፕሬዘዳንት መሆን ትፈልጋለች	addis wants to be a president	
መዓዛ ዶክተር መሆን ትፈልጋለች	Meaza wants to be a doctor	
እሷዶክተር ነች	she is a doctor	
ሜሪ መረብ ኳስ ትጫወታለች	Mary plays volleyball	
አልማዝ መረብ ኳስ መጫወት ትወዳለች	Almaz likes playing volleyball	
ተማሪዎቹ እየተጫወቱ ነው	the are playing	
ጌትነት አሻንጉሊት አለው	Getnet has a doll	
እሱ ህፃን ነው	he is a child	
እሷ የእሷ ልጅ ነች	she is her child	
አሊስ የእሱ እናት ነች	Alice is his mother	
እሱ እናቱን ይወዳታል	he loves his mother	
ጆን ሻይ እየጠጣ ነው	John is drinking tea	
ንግስት የእሱ ሚስት ነች	Nigist is his wife	
አዲሱ ካገባ ሁለት አመት ሆኖታል	Addisu has been married for two years	
አዲሱ ተቀጣሪ ነው	Addisu is an employee	
አልማዝ ፈጣን ነች	Almaz is fast	
ሜሪ የእሱ ሚስት ነች	Mary is his wife	
ሀይሌ የእሷ ባል ነው	Haile is her husband	
አዲስ መፅሀፍ አላት	addis has a book	
አበበ እስክሪብቶ አለው	Abebe has a pen	
እሱ ፈቷል	he is divorced	
እሱ ፊልም ማየት ይወዳል	he loves watching movies	
እሷ ዘፈን እያዳመጠች ነው	she is listening to music	

ጌትነት ዘፈን ማዳመጥ ይወዳል	Getnet loves listening to music	
መዓዛ ተማሪ ነች	Meaza is a student	
አየለ ውድድሩን አሸነፈ	Ayele he won the competition	
አየለ ጨዋታውን አሸነፈ	Ayele won the game	
አስቴር ጨዋታ እየተጫወተች ነው	Aster is playing a game	
አስቴር ፕሬዘዳንት መሆን ትፈልጋለች	Aster wants to be a president	
ንግስት ተመራቂ ነች	Nigist is a graduate	
ሜሪ የእኔን መኪና ትወደዋለች	Mary loves my car	
ቦብ አስተማሪ ነው	Bob is a teacher	
ሜሪ አስተማሪ ናት	Mary is a teacher	
እሷ ሞተች	She is dead	
አስተማሪው ረዥም ነው	the teacher is fat	
አልማዝ አጭር ነች	Almaz is short	
ቦብ እስክሪብቶ ገዛ	Bob bought a pen	
ንግስት ነፍሰ-ጡር ነች	Nigist is pregnant	
አዲሱ እብድ አይደለም	Addisu is not crazy	
አልማዝ ደስተኛ አልነበረችም	Almaz was not happy	
ቦብ ጥሩ ሰው ነው	Bob is a good guy	
አዲሱ እያለቀሰ ነው	Addisu is crying	
ሜሪ ወደ ትምህርት ቤት ሄደች	Mary went to school	
ቦብ እያላበው ነው	Bob is sweating	
ቃልኪዳን ወፍራም ነች	kalkidan is fat	
ሀይሌ ቀጭን ነው	Haile is thin	
ተማሪዎቹ ተጨንቀዋል	the are worried	
ሰዎች በጣም ያወራሉ	people talk too much	
ቃልኪዳን ውሀውን ቀዳችው	kalkidan fetched water	
አበበ እብድ ነው	Abebe is crazy	
አበበ እራሱን ገደለ	Abebe killed himself	
ሀይሌ እራሱን አጠፋ	Haile committed suicide	
ሜሪ እሱን እየለመነችው ነው	Mary is begging him	
ህፃናቶቹ እየሮጡ ነው	the children are running	
ሀይሌ ቆስሏል	Haile is wounded	
አበበ ተናዷል	Abebe is upset	
ቦብ ተደስቷል	Bob is happy	
ቃልኪዳን ተማሪ ነች	kalkidan is a student	
አዲሱ ዶክተር ነው	Addisu is a doctor	
አዲስ መሀንዲስ ነች	Addis is an engineer	
ቦብ አርክቴክት ነው	Bob is an architect	
ሜሪ ማፅዳት ትወዳለች	Mary loves to clean	
ቸክ ሰላይ ነው	Chuck is a spy	
ሰዎች ሰላዮች ናቸው	people are spies	
በላይ ሰላይ አይደለችም	Belay is not a spy	

በላይ ዶክተር አይደለም	Belay is not a doctor	
ውሻው ሞተች	the is dead	
ድመቷ እየበላ ነው	is eating cat	
ሄለን ጥላው ሄደች	Helen left him	
እዮብ ብቻውን ነው	Eyob is alone	
ሄለን ልክ ነች	Helen is right	
እዮብ እያወራ ነው	Eyob is talking	
እሱ ከሄለን ጋር እየተራመደ ነው	he is walking with Helen	
ሰላም ከእሱ ርቃ ሄደች	Selam run away from him	
እሱ ጊታር ይጫወታል	he plays guitar	
እሷ ዘፋኞችን ትወዳለች	she loves musicians	
እሱ ዘፋኝ ነው	he is a musician	
እሱ አሞታል	he is sick	
እሱ ሜሪን ያፈቅራታል	he loves Mary	
ቸክ አባት ነው	Chuck is a father	
ቸክ ደደብ ነው	Chuck is stupid	
እሱ ወንድሙን ገደለው	he killed his brother	
እዮብ ሚስት ነፍሰ-ጡር ነች	Eyob wife is pregnant	
ሰላም አሟታል	Selam is sick	
እሱ በየቀኑ ይሰክራል	he gets drunk everyday	
መስፍን እግር ኳስ ተጫዋች ነው	Mesfin is a football player	
ማርታን የሆነ ሰው ነፍስ አጠፋች	Marta is murdered someone	
ብሩክ ፈርናንዶን መራመድ እንዲያቆም ነገረው	Brook told Fernado to stop walking	
ሞርጋን የሚያስቅ ልጅ ነው	Morgan is funny	
ጀምስ አዲስ ስልክ ገዛች	James bought a new phone	
ሞርጋን ፈርቷል	Morgan is scared	

Appendix III: Questionnaire for the Complex Sentences (English-Amharic)

Grade 1 to 5,
1 - Very poor translation (not related at all),
2- Poor,
3 – Good,
4 - Almost perfect,
5- Perfect translation,

English	Amharic	Point
Public Procurement Directive Short Title	የግዥ መመሪያ ርዕስ	
Unless the context shall otherwise require the amount of such payment	የቃሉ እግባብ ሌላ ትርጉም የሚያሰጠው ካልሆነ በስተቀር የቅድሚያ ክፍያውን	
And all the days of Enoch were not ashamed	እግዚአብሔር ሆነ ዘመን ሁሉ አይተፋፈሩም ነበር	
And all the days of Noah were three hundred sixty and five years :	እግዚአብሔር ኖኅ ዘመን ሁሉ ሦስት መቶ ስድሳ አምስት ዓመት ሆነ ።	
Tax Clearance Certificate	ታክስ መክፈሉን የሚያረጋግጥ የምስክር ወረቀት	
the waters prevailed, and were increased greatly upon the earth; and the ark went upon the face of the waters.	ውኃውም አሸነፈ ፥ በምድር ላይም እጅግ በዛ መርከቢቱም በውኃ ላይ ሄደች ።	
Public Procurement Directive	የግዥ መመሪያ	
Short Title	አጭር ርዕስ	
Unless the context shall otherwise require	የቃሉ እግባብ ሌላ ትርጉም የሚያሰጠው ካልሆነ በስተቀር	
the amount of such payment	የቅድሚያ ክፍያውን	
the following shall be observed	የሚከተሉት ሁኔታዎች መጠበቅ አለባቸው	
And they were both naked , the man and his wife , and were not ashamed .	እነርሱም ፥ አይተፋፈሩም ነበር ።	
And all the days of Enoch were three hundred sixty and five years :	እግዚአብሔር ሆነ ዘመን ሁሉ ሦስት መቶ ስድሳ አምስት ዓመት ነው	
Thus did Noah ; according to all that God commanded him , so did he .	እርሱ እግዚአብሔር እንዳዘዘው ሁሉ እንዲሁ አደረገ ።	
And they said one to another , Go to , let us make brick , and burn them throughly	እነርሱም እርስ በርሳቸው ፥ ጡብ እንሥራ ፥ በእሳትም እንተኵሰው ተባባሉ	
And Pharaoh called Abram , and said , What is this that thou hast done unto me ?	ፈርዖንም ፥ እንዲህም አለ ፦ ይህ ያደረግህብኝ ምንድር ነው ? አለው ።	
But Abram said unto Sarai , Behold , thy maid is in thy hand ;	አብራምም ሦራን ፦ እነሆ ባሪያሽ ናት	
and he gave up the ghost and died ; and was gathered unto his people .	ነፍሱን ሰጠ ሞተም ወደ ወገኖቹም ተከማቸ ።	
And Isaac was forty years old	ይስሐቅም አርባ ዓመት ሰው ነበረ	
These are the sons of Ishmael	የእስማኤል ልጆች እነዚህ ናቸው	
And the Lord said unto her ,	እግዚአብሔርም ፦	
What is this thou hast done unto us ?	ምን ይህ ያደረግህብን ምንድር ነው ?	
And Abimelech charged all his people , saying , He that toucheth this man or his wife shall surely be put to death .	አቢሜሌክም ወገኖቹ ሕዝብ ሁሉ አዘዘ ፦ ይህን ሰው ሚስቱንም እንደ ሞትን ይሙት ።	
Now therefore , my son , obey my voice according to that which I command thee .	አሁንም ልጄ ሆይ ፥ ልጄ ሆይ ፥ እኔ በማዝዝህ ነገር ።	

Therefore God give thee of the dew of heaven , and the fatness of the earth , and plenty of corn and wine :	እግዚአብሔር ስለዚህ ብዛት ይስጥህ ፥ ከሰማይም ጠል የእህልንም	
And Isaac called Jacob , and blessed him , and charged him , and said unto him , Thou shalt not take a wife of the daughters of Canaan .	ይስሐቅም ፥ ባረከውም ፥ እንዲህም ብሎ አዘዘው ፦ ከከነዓናውያን ሴቶች ልጆች ሚስትን አታግባ ሴቶች ልጆች ።	
And Leah said , A troop cometh : and she called his name Gad .	ልያም ። ጉድ አለች ስሙንም ጋድ ብላ ጠራችው ።	
Procuring entities may engage in Direct Procurement when the conditions laid down under Article 27 are fulfilled.	ግዥ ፈፃሚ አካላት ከአንድ አቅራቢ ግዥ ለመፈፀም የሚችሉት በአዋጁ አንቀጽ 27 የተዘረዘሩት ሁኔታዎች ሲሟሉ ይሆናል ።	
For all bids, procuring entities shall prepare bid documents that include the following	ግዥ ፈፃሚ አካላት ለማናቸውም ጨረታ ከዚህ በታች የተዘረዘሩትን ያካተተ የጨረታ ሰነድ ማዘጋጀት አለባቸው	
All Procurement Authorities and the execution thereof must achieve the following objectives .	ግዥን የተሰጠ ስልጣን እንዲሁም ለማስፈፀም የግዥ አፈፃፀም የሚከተሉትን ዓላማዎች ግብ ማድረስ ይኖርበታል ፡ ፡	
30 days for National Competitive bid ; and	ለአገር ውስጥ ግዥ 30 ቀናት ፣	
Rules Applicable to Review Complaints	አቤቱታ የሚጣራበት ስርዓት ጠሚመራባቸው ደንቦች ፣	
But unto Cain and to his offering he had not respect . and Cain was very wroth , and his countenance fell .	አለው እቀብራለሁ ወደ መሥዋዕቱ ግን አልተመለከተም ። ቃየንም እጅግ ተናደደ ፊቱም ጠቆረ ።	
And the Lord said unto him , Therefore whosoever slayeth Cain , vengeance shall be taken on him sevenfold . and the Lord set a mark upon Cain , lest any finding him should kill him .	እግዚአብሔርም እርሱን አለው ፦ አለው ፦ እንግዲህ ቃየንን የገደለ ሰባት እጥፍ ይበቀልበታል ። እግዚአብሔርም ቃየንን ያገኘው ሁሉ እንዳይገድለው ምልክት አደረገለት ።	
And Pathrusim , and Casluhim , (out of whom came Philistim ,) and Caphtorim .	ከእነርሱ የፍልስጥኤም ሰዎች የወጡባቸውን ከስሉሂምን ፥ ቀፍቶሪምንም ወለደ ።	
And you , be ye fruitful , and multiply ; bring forth abundantly in the earth , and multiply therein .	እናንተም ብዙ ፥ ተባዙ በምድር ላይ ተዋለዱ ፥ እርቡባትም ።	
Therefore is the name of it called Babel ; because the Lord did there confound the language of all the earth : and from thence did the Lord scatter them abroad upon the face of all the earth .	ስለዚህም ስምዋ ባቢሎን ተባለ ፥ እግዚአብሔር በዚያ የምድርን ቋንቋ ሁሉ ደባልቋልና ከዚያም እግዚአብሔር ሁሉ ላይ እነርሱን በትኖአቸዋል ።	
And blessed be the most high God , which hath delivered thine enemies into thy hand . and he gave him tithes of all .	ጠላቶችህን በእጅህ የጣለልህ ልዑል እግዚአብሔርም የተባረከ ነው ። አብራምም ከሁሉ አሥራትን ሰጠው ።	
Therefore Abimelech rose early in the morning , and called all his servants , and told all these things in their ears : and the men were sore afraid .	አቢሜሌክም በነጋታው ፥ ባሪያዎቹንም ሁሉ ጠራ ፥ ይህንንም ነገር ሁሉ በጆሮአቸው ተናገረ ሰዎቹም እጅግ ፈሩ	

Appendix IV: Questionnaire for the Complex Sentences (Amharic-English)

Grade 1 to 5,
1 - Very poor translation (not related at all),
2- Poor,
3 – Good,
4 - Almost perfect,
5- Perfect translation,

Amharic	English	Point
የቃሉ እግባብ ሌላ ትርጉም የሚያሰጠው ካልሆነ በስተቀር የቅድሚያ ክፍያውን	Unless the context shall otherwise require the amount of such payment	
ሄኖክም የኖረበት ዘመን ሁሉ አይተፋፈሩም ነበር	And Enoch all the days of his wife , and were not ashamed	
ኖኅ የኖረበት ዘመን ሁሉ ሦስት መቶ ስድሳ አምስት ዓመት ሆነ ።	And all the days of Noah were three hundred sixty and five years :	
ታክስ መክፈሉን የሚያረጋግጥ የምስክር ወረቀት	Tax Clearance Certificate	
ውኃውም አሸነፈ ፥ በምድር ላይም እጅግ በዛ መርከቢቱም በውኃ ላይ ሄደች ።	waters prevailed , and the waters increased greatly upon the earth ; and the ark went upon the .	
የመንግስት ግዥ መመሪያ	Public Procurement Directive	
አጭር ርዕስ	Short Title	
የቃሉ እግባብ ሌላ ትርጉም የሚያሰጠው ካልሆነ በስተቀር	Unless the context shall otherwise require	
የቅድሚያ ክፍያውን	advance payment	
የሚከተሉት ሁኔታዎች መጠበቅ አለባቸው	the following shall be observed	
እነርሱም ፥ ኢይተፋፈሩም ነበር ።	And they , and wife , and were not ashamed .	
እነርሱም እርስ በርሳቸው ፥ ጡብ እንሥራ ፥ በእሳትም እንተኵሰው ተባባሉ	And they said one to another , Go to , let us make brick , and burn them throughly	
ነፍሱን ሰጠ ሞተም ወደ ወገኖቹም ተከማቸ ።	and he gave up the ghost and died ; and was gathered unto his people .	
ይስሐቅም አርባ ዓመት ሰው ነበረ	And Isaac was forty years old	
የእስማኤል ልጆች እነዚህ ናቸው	These are the sons of Ishmael	
እግዚአብሔርም ከሰማይ ጠል ከምድርም ስብ የእህልንም የወይንንም ብዛት ይስጥህ	And God the dew of heaven , and the fatness of the earth , and plenty of corn and wine : give thee of the	
ይስሐቅም ፥ ባረከውም ፥ እንዲህም ብሎ አዘዘው ፦ ከከነዓናውያን ሴቶች ልጆች ሚስትን አታግባ ሴቶች ልጆች ።	And Isaac , and Jacob blessed him , and charged him , and said Thou shalt not take a wife of	
ልያም ። ጉድ አለች ስሙንም ጋድ ብላ ጠራችው ።	And Leah said , A troop cometh : and she called his name Gad .	
ግዥ ፈፃሚ አካላት ከአንድ አቅራቢ ግዥ ለመፈፀም የሚችሉት በአዋጁ አንቀጽ 27 የተዘረዘሩት ሁኔታዎች ሲሟሉ ይሆናል ።	Procuring entities may engage in Direct Procurement when the conditions laid down under Article 27 are fulfilled.	
ግዥ ፈፃሚ አካላት ለማናቸውም ጨረታ ከዚህ በታች የተዘረዘሩትን ያካተተ የጨረታ ሰነድ ማዘጋጀት አለባቸው	For all bids, procuring entities shall prepare bid documents that include the following	
ግዥን የተሰጠ ስልጣን እንዲሁም ለማስፈፀም የግዥ አፈፃፀም የሚከተሉትን ዓላማዎች ግብ ማድረስ ይኖርበታል ፡ ፡	All Procurement Authorities and the execution thereof must achieve the following objectives .	
ለአገር ውስጥ ግዥ 30 ቀናት ፣	30 days for National Competitive bid ; and	
አቤቱታ የሚጣራበት ስርዓት ጠሚመራባቸው ደንቦች ፣	Rules Applicable to Review Complaints	

አለው እቀብራለሁ ወደ መሥዋዕቱ ግን አልተመለከተም ። ቃየንም እጅግ ተናደደ ፊቱም ጠቆረ ።	unto Cain and to his offering : he had not respect . and Cain was very wroth , and his countenance fell .	
እግዚአብሔርም እርሱን አለው ፦ አለው ፦ እንግዲህ ቃየንን የገደለ ሰባት እጥፍ ይበቀልበታል ። እግዚአብሔርም ቃየንን ያገኘው ሁሉ እንዳይገድለው ምልክት አደረገለት ።	And the Lord said unto him , Therefore whosoever slayeth Cain , vengeance shall be taken on him sevenfold . and the Lord set a mark upon Cain , lest any finding him should kill him .	
ከእነርሱ የፍልስጥኤም ሰዎች የወጡባቸውን ከስሉሂምን ፥ ቀፍቶሪምንም ወለደ ።	is filled with violence through them , and unto his father 's house , and Casluhim , (out of Abraham his father , the Philistines had የወጡባቸውን came Philistim ,) and Caphtorim .	
እናንተም ብዙ ፥ ተባዙ በምድር ላይ ተዋለዱ ፥ እርቡባትም ።	ye , Be forth abundantly in the earth , and multiply therein .	
ስለዚህም ስምዋ ባቢሎን ተባለ ፥ እግዚአብሔር በዚያ የምድርን ቋንቋ ሁሉ ደባልቋልና ከዚያም እግዚአብሔር ሁሉ ላይ እነርሱን በትኖአቸዋል ።	I praise the Lord : therefore she name of it called Babel ; because all Lord did there confound the language of and from thence did the Lord scatter them abroad upon the face of all .	
ጠላቶችህን በእጅህ የጣለልህ ልዑል እግዚአብሔርም የተባረከ ነው ። አብራምም ከሁሉ አሥራትን ሰጠው ።	And blessed be the most high God , which hath delivered thine enemies into thy hand . and he gave him tithes of all .	
የግዥ መመሪያ	public Government Procurement Directive	
ፈርዖንም ፥ እንዲህም አለ ፦ ይህ ያደረግህብኝ ምንድር ነው ? አለው ።	And Pharaoh said unto his servants him yet again , and said , What is this that thou hast done thou camest ?	
ኖኅም እንዲሁ አደረገ እግዚአብሔር እንዳዘዘው ሁሉ እንዲሁ አደረገ ።	and Noah , so did he all that God commanded him , so did he .	
አብራምም ሦራን ፦ እነሆ ባሪያሽ ናት	Abram said unto Sarai , Behold , thy maid is	
እግዚአብሔርም አላት ፦	And the Lord said unto her ,	
ሄኖክም የኖረበት ዘመን ሁሉ ሦስት መቶ ስድሳ አምስት ዓመት ሆነ ።	And all the days of Enoch were three hundred sixty and five years :	
ይህ ያደረግህብን ምንድር ነው ?	hast thou done unto us ?	
አቢሜሌክም ሕዝቡን ሁሉ ፦ ይህን ሰው ሚስቱንም የሚነካ ሞትን ይሙት ብሎ አዘዘ ።	And Abimelech charged all his people , saying , He that toucheth this man or his wife shall surely be put to death .	
አሁንም ፥ ልጄ ሆይ ፥ እኔ በማዝዝህ ነገር ስማኝ	Now therefore , my son , obey my voice according to that which I command thee .	
ይስሐቅም ያዕቆብን ጠራው ፥ ባረከውም ፥ እንዲህም ብሎ አዘዘው ፦ ከከነዓናውያን ሴቶች ልጆች ሚስትን አታግባ	And Isaac called Jacob , and blessed him , and charged him , and said unto him , Thou shalt not take a wife of the daughters of Canaan .	

Appendix V: Sample Corpus on Simple Sentences

Abebe loves enjera	አበበ እንጀራ የወዳል
Aster likes Abebe	አስቴር አበበን ትወደዋለች
Almaz is getting married	አልማዝ ልታገባ ነው
Abebe is married	አበበ አግብቷል
John is divorced	ጆን ፈቷል
Aster is a student	አስቴር ተማሪ ነች
John plays basketball	ጆን ቅርጫት ኳስ ይጫወታል
Aster watches football	አስቴር ኳስ ታያለች
Almaz plays football	አልማዝ ኳስ ትጫወታለች
Haile is a runner	ሀይሌ ሯጭ ነው
He loves competition	እሱ ውድድር ይወዳል
Haymanot enjoys watching football	ሀይማኖት ኳስ ማየት ያዝናናታል
he loves Haymanot	ሀይማኖትን ይወዳታል
Addisu has a car	አዲሱ መኪና አለው
his car is comfy	መኪናው ምቹ ነው
Meaza wants to be a president	መዓዛ ፕሬዘዳንት መሆን ትፈልጋለች
addis wants to be a doctor	አዲስ ዶክተር መሆን ትፈልጋለች
kalkidan is a doctor	ቃልኪዳን ዶክተር ነች
Getnet is her child	ጌትነት የእሷ ልጅ ነው
Nigist is his mother	ንግስት የእሱ እናት ነች
he loves his mother	እሱ እናቱን ይወዳታል
my mother is rich	እናቴ ሀብታም ነች
my father loves my mother	አባቴ እናቴን ይወዳታል
my mother loves my father	እናቴ አባቴን ትወደዋለች
he has a son	እሱ ወንድ ልጅ አለው
she has a daughter	እሷ ሴት ልጅ አላት
her daughter is clever	የእሷ ልጅ ጎበዝ ነች
Ayele loves watching movies	አየለ ፊልም ማየት ይወዳል
Aster is listening to music	አስቴር ዘፈን እያዳመጠች ነው

he loves listening to music	እሱ ዘፈን ማዳመጥ ይወዳል
the movie is new	ፊልሙ አዲስ ነው
The news was good	ዜናው ጥሩ ነበር
the movie was long	ፊልሙ ረዥም ነበር
the movie was short	ፊልሙ አጭር ነበር
the movie was boring	ፊልሙ አሰልቺ ነበር
I needed a tissue	እኔ ሶፍት ፊልጌ ነበር
I need a tissue	እኔ ሶፍት እፈልጋለሁ
she is an optimist	እሷ ጥሩ አሳቢ ነች
she is a pessimist	እሷ መጥፎ አሳቢ ነች
Abel killed someone	አቤል የሆነ ሰው ገደለ
Mary was shocked	ሜሪ ደንግጣ ነበር
Eyob is walking	እዮብ እየተራመደ ነው
Eyob is walking with Helen	እዮብ ከሄለን ጋር እየተራመደ ነው
Helen run away from him	ሄለን ከእሱ ርቃ ሄደች
Eyob plays guitar	እዮብ ጊታር ይጫወታል
Helen loves musicians	ሄለን ዘፋኞችን ትወዳለች
Eyob is a musician	እዮብ ዘፋኝ ነው
Chuck is sick	ቸክን አሞታል
Chuck loves Mary	ቸክ ሜሪን ያፈቅራታል
Mary is calling her husband	ሜሪ ለባሏ እየደወለችለት ነው
Fikadu is in trouble	ፍቃዱ ችግር ውስጥ ነው
Belay killed his brother	በላይ ወንድሙን ገደለው
Moges's wife is pregnant	የሞገስ ሚስት ነፍሰ-ጡር ነች
Marta is sick	ማርታን አሟታል
Mesfin gets drunk everyday	መስፍን በየቀኑ ይሰክራል
Naod is a spy	ናዖድ ሰላይ ነው
Samuel works for the government	ሳሙዔል ለመንግስት ይሰራል
Bethlehem graduated in computer science	ቤተልሄም በኮምፒውተር ሳይንስ ተመርቃለች