ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

Geez to Amharic Automatic Machine Translation: A Statistical Approach

A thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Master of Science in Information Science

BY

DAWIT MULUGETA

MAY, 2015

AAU

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURAL SCIENCES
SCHOOL OF INFORMATION SCIENCE

GEEZ TO AMHARIC AUTOMATIC MACHINE TRANSLATION: A STATISTICAL APPROACH

BY

DAWIT MULUGETA

Approved by the Examining Board

_____ _____

Chairman, Examining Committee

_____ _____

Advisor Signature

_____ _____

Examiner

Declaration

I, the under signed, declare that this thesis is my original work, has not been submitted as a partial requirement for a degree in any university and that all sources of materials used for the thesis have been duly acknowledged.

_____

Dawit Mulugeta

May, 2015

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

Name: _____

Signature: _____

Date: _____

ACKNOWLEDGMENTS

First and for most, I would like to express my heartfelt thanks to God, who gave me the strength, determination, endurance and wisdom to bring this thesis to completion.

I also wish to express my sincere gratitude to all whom through their supports contributed to the successful completion of this work. First, I am highly indebted to my research advisor, Dr. Martha Yifiru, for her expertise, generous time, and patience in helping me complete this thesis.

I am especially thankful to my friend Ato Solomon Mekonnen for initiating the research idea and giving me invaluable assistance in finishing this research. I am also grateful to my friends Abreham Shewarega, Ezana Girma, Meseret Ayano, Tariku Tenkir, Gebeyehu Kebede, Hirut Timerga, Henok Kebede, and Eyasu Mekete for their support and encouragement.

Last and most importantly, I would like to extend my heartfelt gratitude to all my family members for their love, encouragements and moral support.

Table of Contents

List of Tables

List of Figures

List of Appendces

List of Acronyms

AAU – Addis Ababa University

AI - Artificial Intelligence

BLUE - Bilingual Evaluation Understudy

CBMT - Corpus Based Machine Transfer Approach

EBMT – Example Based Machine Translation

EM – Expectation Maximization

EOTC – Ethiopian Orthodox Tewahdo Church

MT – Machine Translation

NLP – Natural Languge Processing

RBDTA - Rule Based Direct Transfer Approach

RBIA – Rule Based Interlingua Approach

RBMT – Rule Based Machine Translation

RBTA – Rule Based Transfer Approach

SL – Source Language

SMT – Statistical Machine Translation

TL – Target Language

CV – Cross Validation

ABSTRACT

Machine Translation (MT) is the task of automatically translating a text from one natural language to another. MT is essential for many applications including multilingual information retrieval, speech to speech and others. The theme of this thesis is Geez to Amharic MT based on statistical approach which addresses the problem of automatically translating Geez text to Amharic text. Geez is classical South Semitic language which is attested in many inscriptions including historic, medical, religious and other since the early 4th century. Today Geez remains only as a spoken language and the liturgy language of the Ethiopian Orthodox Tewahedo Church. Whereas, Amharic is among the most spoken language in Ethiopia and the official working language of the Federal Government of Ethiopia, where it has about 30 million native and non-native speakers. The machine translation of Geez document to Amharic will be of paramount importance in order to enable Amharic user to easily access the invaluable indigenous knowledge decoded in Geez language.

Therefore, the thesis is focused on investigating the application of corpus based machine translation approach in order to translate Geez documents to Amharic. The method that is employed to conduct the experimentation is a Statistical Machine Translation (SMT) approach. This approach requires availability of a large volume of parallel documents prepared in Geez and Amharic. The experiment was conducted using Moses (statistical Machine Translation tool), GIZA++ word alignment toolkit and IRSTLM language modeling tools on 12, 840 parallel bilingual sentences and an average translation accuracy of BLUE score 8.26 was achieved on 10-fold cross validation experimentation. With the use sufficiently large parallel Geez-Amharic corpus collection and language synthesizing tool, it is possible to develop a better translation system for the language pairs.

CHAPTER ONE

INTRODUCTION

1.1 Background

Machine translation (MT) is the automatic translation from one natural language into other using computers. It uses of computers to automate some or all of the process of translating from one language to another. It is an area of applied research that draws ideas and techniques from linguistics, computer science, Artificial Intelligence (AI), translation theory, and statistics. Machine translation has many application areas that could use the result of the translation (Arnold et.al, 1994).

MT, in recent years, has become a great concern in relation to natural language processing. The advances in technology, increasing digital data collections, the technical facility and the continuing interest of Interlingua resources sharing have necessitated the development of MT. Especially languages with rare digital collection are to benefit from the translation of the other digital corpus rich languages.

Various methodologies have been devised to automate translation process. The major approaches can be divided in to two: the older Rules-Based Machine Translation (RBMT) and the Corpus based Machine Translation (CBMT). RBMT relies on manually built-in large collections linguistic rules and bilingual dictionaries for each language pair. Corpus based machine translation uses a large amount of raw data in the form of parallel corpora and is able to overcome the many of the challenges in rule based machine translation. Corpus based approach is further classified into two sub approaches: Statistical Machine Translation (SMT) and Example-based Machine Translation Approach (EBMT). The SMT

seems more dominantly preferred approach of many industrial and academic research laboratories (Schmidt, 2007). As the SMT basis on statistical models whose parameters are derived from the analysis of bilingual text corpora, the size of the bilingual corpus size will matter on the performance of the system. However, the acquisition of large amount of high-quality bilingual parallel text is difficult. Researches shows that an acceptable translation quality can be achieved with the available small amount of parallel corpus, especially if specific domain parallel corpus, phrasal corpus, text processing techniques, as well as some morpho-syntactic knowledge are used (Denkowski et al, 2014) and (Popovic et.al, 2006). Pre- and post-editing technologies are also one of the most recent research focuses in MT (Gerlach at al, 2013) and (Seretan et al, 2014).

Amharic is one of the languages in the Semitic family which is widely spoken in Ethiopia (Bender, 1976). Amharic, being the official working language of the Democratic Republic of Ethiopia, has a large number of speakers either as mother tongue or as their second language. It is also estimated that Amharic is spoken by about 30 million people as a first or second language (Rubin, 2010), making it the second most spoken Semitic language in the world (after Arabic), the second largest language in Ethiopia (after Oromo), and one of the five largest languages on the African continent (Adejumobi, 2007).

Geez, also called Ethiopic, has been serving as big source of the resources literary for a long period of time around the introduction of Christianity in Abyssinia and Axumite period. Among the oldest of the Semitic languages, Geez is now confined to ecclesiastical use (Adejumobi, 2007). Geez literature is ordinarily divided into two periods; the first dates back from the establishment of Christianity in the 5th century and ends on the 7th century – basically religious books translation. The second period starts from the

reestablishment of the Solomonic dynasty in 1268 counting to the present time which is dealing with religious, story, culture and philosophy in the country. In addition, it has also been serving as a language of instruction in the country's traditional schools (Harden, 1926). The huge amount of indigenous knowledge that has been accumulated in the country is found in Geez language.

Nevertheless, manuscripts in Amharic are known from the 14th century and the language has been used as a general medium for literatures, journalism, education, and communication since 19th C. The statistical technique to machine translation has emerged as a highly promosing approach. The approach is expected to be more promising for translations between two related languages like Geez and Amharic (Ferreira, 2007). This research presents the translation of Geez to Amharic using the SMT approach. It is also initiated not only considering the benefits to Amharic users but also other language user when Amharic to other language translation is done.

1.2 Statement of the Problem

Geez is an ancient language and many manuscripts are already documented by Ethiopian Orthodox Church as well as by the National Archival agency (Tadese, 1972) and (Ullendorff, 1955). Geez has been known to be used in Ethiopian since the fourth century and probably died out as a spoken language close to a thousand years but have been serving as official written language practically up to the end of the nineteenth century (Baye, 1992) and (Hetzron, 1997). Since currently Geez is not a widely spoken language, there is a need to translate the manuscripts to Amharic and other Ethiopian Languages in order to make the decoded knowledge accessible to everyone especially

Amharic users. Some attempts are done by the EOTC and individuals to translate some of the religious manuscripts, law and some philosophical works manually (Harden, 1992). However still there are many literatures in medicine, astronomy, history, religious manuscripts and other materials that are not translated to Amharic and other widely used language (Tadese, 1972), (Leslau, 1995) and (Ilana, 2013).

In addition, the manual translations are relatively slow, monotonous, and resource intensive as it requires professionals to specialize in a specific field (Osborne, 2004). The other alternative is to develop machine translation software which is relatively less costly and does not require Geez linguistic experts once the system is developed. From the reviews made in the area of Geez Language, there are a very huge amount of resources available in Geez Language that range from religious to the philosophical, medical and other disciplines. Hence, practice of the Machine Translation of Geez to Amharic becomes paramount.

Machine translation, although it has its own challenges, can improve performance, reduce cost and exposure to error. This work will investigate the application of machine learning approaches in order to translate Geez documents to Amharic.

As discussed earlier, however, there are few researches made on MT in Ethiopian languages in general and in Amharic and Geez in particular. Mulu et al.(2012) and Adugna (2009) attempted to do a preliminary experiment to translate English-Amharic and English-Oromo using the SMT respectively. Gasser (2012) made efforts toward a Rule-Based System for English-Amharic Translation. He tried to implement RBMT using the L^3 framework which relies on a powerful and flexible grammatical theory. Similarly, Dagnachew (2011) has made an attempt on Machine Translation System for Amharic

Text to Ethiopian Sign Language. Saba et al (2006) Present a short summery of some works done in areas of Amharic language processing with a special focus to the development of machine translation. All these attempts are at an experimental stage and the outputs are prototypes which are limited in scope.

However, as to the researcher knowledge, there are no researches outputs reported in related to machine translation in Geez. To the knowledge of the researcher the approaches has not been experimented for SMT for Amharic to Geez language. To this end, the purpose of this study is to explore the possibilities of translating Geez documents to Amharic using corpus based approach especially SMT. The study particularly aims to answer the following questions:

- Does variation in the language model corpus size brings a change in the performance of the system?
- Does normalization of the target language corpus lead to better performance of the translation result?

1.3 Objectives of the Research

1.3.1 General Objective

The general objective of this research is to investigate the application of Statistical Machine learning technique to Machine Translation from Geez to Amharic.

1.3.2 Specific Objective

To achieve the general objective, the study attempts to address the following specific objectives:

- Review the basic writing system, punctuation marks and syntactic structure of Geez and Amharic Language as well as approaches that are used for machine translation for other languages;
- Prepare training and test data;
- Train a machine translation system using the selected Machine Learning algorithm;
- Test the performance of the system; and
- Forward conclusion and recommendations.

1.4 Significance of the Study

The results of the study are expected to produce experimental evidences that demonstrate the applicability of machine learning approach to machine translation from Geez to Amharic. Moreover the results of the study can be used to develop machine translation software for Geez to Amharic, which will be used to translate enormous literatures in Geez to Amharic. In addition, it will also contribute to future researches and developments in other application areas like cross lingual Information Retrieval from Amharic to Geez and vice versa as such applications need machine translation as a compliment.

1.5 Methodology

This paper use Quantitative Experimental as research methodology. It has been reported in literature that this methodology is best for obtaining information about causal relationships (Robson, 1993), allowing researchers to assess the correlation (relationship) between one variable and another. In the experiment, the paper used different variables and investigated their effect such us normalization, corpus size, and test split options.

1.5.1 Literature Review

Since there are different approaches used in machine translation, review of literatures in the area of machine translation with special focus on SMT approach and algorithms used have been done.

Translation of Geez to Amharic using machine translation approach requires review of synthetic structure of the two languages in order to understand the Interlingua structures, morphological characteristics the two languages and foresee their impact on the translation. In addition, discussions with the expertise of Geez and Amharic language have been done.

1.5.2 Data Collection and Preprocessing

In order to perform corpus based machine translation, a large amount of bilingual and monolingual data is required. In order to obtain the required amount of parallel data, a Holy Bible Geez-Amharic translation and some other religious books (Wedase Mariam and Arganon) are used. 12860 parallel sentences are used for the training and testing. The collected data were divided in to training and testing set in such a way that more than 90% of the collected data was used as a training set. The proportion is selected to make the training data set relatively large.

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

document (Widase Mariam and part of Arganon), which are not aligned at sentence level were aligned manually.

1.5.3 Tools and Techniques

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

The Stanford Phrasal phrasal[1], Pharaoh[2] and Moses are among phrase-based machine translation toolkit used for SMT (Philipp, 2007) and (Galley, 2009). The common statistical MT platform, namely Moses, is used for the translation. Moses is selected due to the familiarity of the researcher to the tool and because of its accessibility, processing capability and language independent features. Moses consists of all the components needed to preprocess data, train the language models and the translation models (decoding) (Och, 2003). Although Moses integrates both the IRSTLM[3] and SRILM language modeling toolkits, the IRSTLM, which requires about half memory than SRILM for storing an equivalent LM during decoding (Federico et.al, 2007), is used in this research.

---

1 A Phrase-Based Translation System - http://nlp.stanford.edu/software/phrasal/

2 A decoder for phrase-based SMT - http://www.isi.edu/licensed-sw/pharaoh/

3 http://sourceforge.net/projects/irstlm/

In building the word alignment, GIZA++[4], word alignment toolkit is used. GIZA++ is the most widely applied package in SMT word alignment that uses to train IBM Model 1 to Model 5 (Brown et al., 1993) and the Hidden Markov Model (HMM) (Och et al., 2003). The BLUE (Bilingual Evaluation Understudy), which is one of the famous evaluation methods of a comparison among different Machine Translation systems (Zhu, 2001), is used for evaluation.

1.5.4 Experiments

In the preprocessing, The parallel bilingual corpus, both Amharic and Geez data, are aligned sentences level and then normalized, tokenized and cleaned from noise character before training and testing (see section 5.2). In the exponent 90% of the dataset was used for training and the remaining 10% of the dataset used for testing using 10 fold cross Validation test split (see section 5.3).

The Moses decoder toolkit is given 90% of the sentence level Geez - Amharic parallel corpus and Amharic monolingual corpus to build the translation model and the language model. And finally the remaining dataset are used to test the experiment. Ten experiments are done using a 10-fold Cross Validation.

1.6 Scope and Limitations of the Study

Though there are word based, phrased based and tree based SMT approaches, due to time constraint to train, test and analyze the results, only phrased based SMT is used for this thesis. There are different limitations faced during the process of conducting this research. The first and the most challenge was the lack of bilingual corpus for the training

[4] http://www.statmt.org/moses/giza/GIZA++.html

and testing. The limitation comes from the absence of sufficient amount digitally available documents in Geez. Due to lack of digitized data other than the religious one, we were not able to test the performance of the system using different data other than the religious domain. In addition, the lack of educational materials including books and journals in the two languages are the other limitation.

1.7 Organization of the Thesis

The thesis is organized into six chapters comprising Introduction, review of Geez and Amharic Languages, Machine Translation, Statistical Machine Translation, Geez – Amharic SMT, and Conclusion and Recommendations. This chapter gives the general overview of the whole thesis. It describes the background of the research, statement of the problem, the objectives of the research, the methods used and limitation of the study. The second chapter briefly discusses the synthetic structure of the two languages in order to understand the Interlingua structures, morphological characteristics and analyze the semantics between the two languages in order to foresee their impact on the translation.

The third chapter reviews different literatures regarding Machine Translation together with its different approaches with a special focus on Statistical Machine Translation. The chapter covers the components in the SMT in detail. The fourth chapter discusses the experimental setup, software tools used, the hardware environment, architecture of the system, the data used for the experimentation of the research. The fifth chapter discusses the experimentation, analysis, and the performance level of the system that has been achieved together with discussions of the reasons for the result. Finally, chapter Six presents the conclusion and the recommendations drawn from the findings of the study.

CHAPTER TWO

GEEZ AND AMHARIC LANGUAGES

In machine translation of documents from one natural language to another language, ambiguities that could arise from lexical, structural, semantic and other forms of ambiguities are inevitable (Getahun, 2001). This chapter is intended to cover the overview of the linguistic relationship between the Geez and Amharic Languages so as to understand the ambiguities and sources of errors that could arise in the process of translation.

2.1 Geez Language

Geez, sometimes called Ethiopic, is an ancient South Semitic language of Ethiopia and Eritrea in the Horn of Africa later became the official language of the Kingdom of Aksum (Rubin, 2010). Geez is still the liturgical language of the Ethiopian Orthodox Tewahido Church (EOTC) which is attested in inscription since the early 4th century. Geez has probably died out as a spoken language close to 13thC, but remained the primary written language of Ethiopia up to the 20th century. The literature includes religious texts, as well as secular writings (Dillmann, 1907).

Today Geez language remains only as the main language used in the liturgy of the EOTC, the Eritrean Orthodox Tewahedo Church, the Ethiopian Catholic Church, and also the Beta Israel Jewish community of Ethiopia.

2.2 Amharic Language

Amharic is the second most spoken Semitic language in the world (after Arabic) and the second largest language in Ethiopia (after Affan Oromo) (Rubin, 2010). It is the official working language of the Federal government of Ethiopia, where it has about 30 million native and non-native speakers. Manuscripts in Amharic are known from the 14th century and the language has been used as a general medium for literatures, journalism, education, and communication since 19th C.

2.3 Linguistic Relationships of Geez and Amharic

2.3.1 The Writing Systems

Geez script is an alphasyllabary script, also called an abugida, in which a character represents a consonant and a vowel combination. This is different from alphabetic script, where each character denotes one sound -- either a consonant or a vowel. The alphabets of Amharic are unique scripts acquired from the Geez and use an alphasyllabary writing system where the consonant and vowel are combined to form a single symbol. Thus, once a person knows all the alphabets, she/he can easily read and write both Geez and Amharic (Thomas, 1978).

Script in Geez includes thirty-three basic alphabets (called 'Fidel') , each having seven various forms created by fusing a consonant for an alphabet with vowels yielding 231 distinct symbols (Gambäck, 2005) and other non-basic forms derived from the basic alphabets like ኳ(kwa) from ከ (ke) and ቈ(qwa) from ቀ (qe) etc . The non-basic forms are derived from the basic ones by somewhat regular modifications for the first four orders and for the last two words it is irregular. Among the thirty-three consonants, only twenty-

seven have unique sounds. The remaining six consonants have twin sound with other alphabets. For example, each of the alphabets ሀ, ሐ, and ኀ has the same sound which is pronounced as ‘ha’ (Dillmann, 2005).

2.3.2 Syntactic Language Structure (Word Order)

The syntactic structure is formed by combining different words in sequence. The syntactic structure of Amharic is generally SOV (Subject-Object-Verb) whereas Geez follows Subject-Verb-Object (SVO) word order for declarative sentences. The Amharic equivalent for the Geez sentence “ውእቱ መጻአ እምቤቱ [weetu metsa embet]” is “እሱ እቤት መጣ [esu ebet meta]” meaning “He came home” where “እሱ[esu]” is the subject of the Amharic sentence equivalent to “ውእቱ [weetu] in the Geez , “እቤት[ebet]” is the object of the Amharic sentence equivalent to እምቤት [embet] in the Geez and “መጣ [meta]” is the verb of the Amharic sentence which is equivalent to መጻ[metsa] in Geez . But usually pronouns are omitted in both Geez and Amharic sentences and become part of the verb when they used as a subject “መጽአ እምቤቱ [metsa embet]” equivalent to “እቤት መጣ [ebet meta]”.

Question formation in both Geez and Amharic is the same as a declarative sentence except the usage of question mark at the end. To ask the question “Did he go home?” in Amharic, the sentence ends with question mark instead of the Amharic full stop (Arat netib - ::) and become “እሱ ወደ ቤት ሄደ ?”. The Geez equivalent is “ውእቱ ሖረኑ እምቤት ?”. Sometimes, in Amharic, question indicator words are added at the end of the sentence. In such cases the above question becomes “እሱ ወደ ቤት ሄደ እንዴ ?”. Here, the word “እንዴ” is added to indicate that the sentence is a question. Whereas the Geez has no such indicative words

Both Amharic and Geez have a complex morphology. The word formation, for instance, involves different formations including prefixation, infixation, suffixation, reduplication, and others forms. Most function words in Amharic and Geez, such as Conjunction, Preposition, Article, Pronominal affixes, Negation markers, are bound morphemes, which are attached to content words, resulting in complex words composed of several morphemes (Sisay, 2007). Morphologically complex languages also tend to display a rich system of agreements between the syntetic part of a sentence like nouns, verbs, person, number and gender and so on (Minkov, 2007). This will increase the complexity of word generation. In adition, morephologyicaly rich languages permits a flaxable word order, this make difficult to model words. When both the source and the traget languages are morphologucal rich, the difficulty in translation also gets complex (Ceausu, 2011).

2.3.3 Noun

Amharic nouns are either simplex (primary) (e.g. “ቤት[bet]” – house) or derived from verb roots, adjectives, other nouns and others (e.g. “ደግነት[degnet] meaning generosity is derived from ደግ[deg] - ’generous’) (Amsalu, 2004). Nouns in Amharic also inflect for Number (Plural and Singular), Gender (masculine and feminine), Case and Definiteness. Similarly Geez inflect the same morphosyntactic behavior and distinguish Number, Gender, Case and Definiteness by adding suffixes, prefixes and internal pattern modification (e.g “ጾም [tsome]” to “አጽዋማት [atsumat]“ – ‘Fasting’ , “ማይ[may] - ማያት[mayat]” – ‘Water’) (Berihu, 2011).

2.3.4 Verb

Generally, Amharic verbs are derived from roots and use a combination of prefixes and suffixes to indicate the person, number, voice (active/passive), tense and gender. Verbs in Amharic mostly are placed at the end of the sentence (Sisay, 2007) whereas in most Geez sentences the verbs are placed in the middle (Desie, 2003). The Geez Verbs are regularly inflected according to person, gender and number. Geez verbs exhibit the typical Semitic non-linear word formation with intercalation of roots with vocalic pattern. Verbs agree with their subjects and optionally with their objects in both Geez and Amharic (Berihunu, 2011). The main verbs in Geez are usually either perfect (past forms) or imperfect (present and future forms) (Desie, 2003).

2.3.5 Pronouns

Both Amharic and Geez are pro-drop languages where pronouns can be dropped without affecting the meaning. In addition to the first, second and third-person singular and plural pronouns, Amharic have polite independent pronouns (እርስዎ[Ersewo] and እሳቸው[Esachew]) to refer to a person and/or people the speaker wishes to show respect which is not available in Geez. Both Amharic and Geez are pro-drop languages where pronouns can be dropped without affecting the meaning. Geez has ten distinct personal pronouns that act as copulas whereas the Amharic has nine personal pronouns as indicated in the table below.

	SINGULAR					PLURAL				
	Geez		Amharic		English	Geez		Amharic		English
1st Person	አነ	Ane	እኔ	Ene	I	ንሕነ	nehne	እኛ	Egna	We
2nd Person male	አንተ	Ante	አንተ	Ante	You(m.)	አንትሙ	Antimu	እናንተ	Enante	You (m.)
2nd Person female	አንቲ	Anti	አንቺ	Anchi	You(f.)	አንትን	Antin	እናንተ	Enante	You (f.)
3rd Person male	ውእቱ	Weetu	እሱ	Esu	He/It	እሙንቱ	Emuntu	እነሱ	Enesu	They (m.)
3rd Person male	ይእቲ	Yeeti	እሷ	Esua	She/It	እማንቱ	Emantu	እነሱ	Enesu	They (f.)
2nd Person polite	አንተ/ አንቲ		እርስዎ	Ersewo	You (respectful)					
3rd Person polite	ውእቱ/ ይእቲ		እሳቸው	Esachew	He/She (respectful)					

Table 1.5.4-1 Pronouns in Geez and Amharic

The subjective, objective and reflexive pronouns follow the same patterns as the personal pronouns with the pre, post and internal modification as shown in the example below. ለራስዎ ያውቃሉ [Lerasewo yawkalu] - ለሊክ ትአምር [Lelike teamir] - 'you know for yourself' ለራስህ ታውቃለህ [Lerash tawkaleh] - ለሊክ ትአምር [Lelike teamir] - 'you know for yourself' "ንጉሥ አንተ [Negus Ante] - አንተ ንጉሥ ነህ [Ante Negus Neh] -'You are a king'".

2.3.6 Adjective

Adjectives are words or constructions used to qualify nouns. Adjectives in Amharic are either primary adjectives (e.g ጥቁር [Tikur] — 'black') or generally derived from nouns (e.g ኃይለኛ [Haylegna] — 'forceful' from ኃይል [Hayle] — 'force'), verbs (e.g ገላጭ [Gelach] — 'describer' from ገለጸ [Geletse] — 'discribe'), combination of verb and

verb, verb and noun, adjective and adjective (e.g ወጣ ገባ[weta geba] — ‘on and off’) and other parts of speech (Leslau, 1995). Adjectives are inflected for Number, Case, Gender and Definiteness (Saba and Gibbon, 2004). Adjectives are mostly placed before the noun in both Geez and Amharic sentences and agree with the noun in gender and number.

For example:

ጸአዳ መልበስ [Tseada Melbes] equal to ነጭ ልብስ [Nech Libse] - ‘Whilte Cloth’

ንስቲት ማየ[Nistite Maye] equal to ትንሽ ውኃ[Tinish weha] — ‘Some water’.

2.3.7 Adverbs

The notion of adverbs is to modify the verb’s place, time, degree etc. In most cases, Geez adverbs follow the verb they modify whereas the Amharic adverbs precede the verb they modify.

For example in the sentence, ሮፀ ኃይሌ ፍጡነ[Rotse Haile Fitune] - ኃይሌ በፍጥነት ሮጠ[Haile beftinet Rote] — ‘Haile ran fast’, the Adverb ፍጡነ[Fitune] follow the verb ሮፀ [Rotse] in the Geez sentence. However, the Amharic adverb በፍጥነት[beftinet] precede the verb ሮጠ[Rote] (Desie, 2003). Adverbial functions are often accomplished with noun phrases, prepositional phrases and subordinate clauses.

2.3.8 Conjunctions

Conjunctions are words that are used to connect clauses, words, and phrases together. Amharic conjunctions can either separable one that exist by themselves

as words in a sentence like “እና[ena] — ‘and’ ” and inseparable one that serve as conjunctions when joined with verbs and nouns like “ና[na] — ‘also and’ ”. Conjunctions and prepositions have similar behaviors, and are often placed in the same class (mestewadid). Geez conjunction are also separable including “አወ[Awe] — ‘or’ ” and “ወማሚመ[wemime] — ‘ether or’ ” and inseparable Conjunctions including “እም[Eme] — ‘from’ ” and “ወ[we] — ‘as well as’ ”

2.3.9 Punctuation Mark

Both Amharic and Geez apply similar punctuation marks (signs) for different purposes. However, only few of them are practically used, especially in computer-written text. The individual word-separator in the sentence (“hulet netib” - two dots arranged like colon (:)), and sentence-separator (“arat netib” - four dots arranged in a square pattern (: :)), Lists of text separator which is equivalent with comma (“netela serez” (፣)) and “derib sereze(፤)” equivalent to that of semicolon are the basic punctuation marks of writing system that are used consistently. Today, the use of Hulet Neteb is not seen in modern typesetting rather replaced by space. The symbol ‘?’ is used to represent questions.

CHAPTER THREE

STATISTICAL MACHINE TRANSLATION

In this chapter, review of literatures in the field of machine translation has been made. The chapter covers overview of machine translation, challenges and the major approaches of machine translation with special focus on corpus based approach (Statistical Machine Ttranslation). Furthermore, the different components, algorithms, recent development and tools used in Corpus based Machine Translation approaches are discussed in detail.

3.1 Machine Translation

The term machine translation (MT) refers to computer-based translation from one natural language (source language) into another language (target language) using computers with or without human assistance. Machine Translation was conceived as one of the first applications of the newly invented electronic computers back in 1940's. MT is an applied research that draws ideas and techniques from linguistics, computer science, artificial intelligence, translation theory and statistics (Clark et al, 2010). Machine Translation is important to minimize the language barrier in information access and promote multi-lingual real-time communications.

3.2 History of Machine Translation

Although there are some disputes about who first had the idea of translating automatically between human languages, the actual development of Machine Translation System can be traced back to an influential paper written in July 1949 by Warren Weaver - a director

at the Rockefeller Foundation. The letter introduced Americans to the idea of using the first non-military computers for translation purpose which marked machine translation as the first non-numerical application of computers. He outlined the prospects and suggested various methods: the use of statistical methods, Shannon's information theory, and the exploration of the underlying logic and universal features of language.

Since Andrew Booth and Warren Weaver's first attempt to use newly invented computers for machine translation appeared in 1946 and 1947, many machine translation approaches have been developed (Hutchins, 2007). The first conference on MT was organized in 1952 where the outlines of future research were made clear. Just two years later, there was the first demonstration of a translation system in January 1954. In 1966 the US sponsors of MT research committee called Automatic Language Processing Advisory Committee (ALPAC) published an influential report which concluded that MT was slower, less accurate and twice as expensive as human translation. However, in the following decade MT research took place largely outside the United States, in Canada and in Western Europe and work continued to some extent (Thurmair, 1991).

Research since the mid-1970s has three main strands: first, the development of advanced transfer systems building upon experience with earlier Interlingua systems; secondly, the development of new kinds of interlingua systems; and thirdly, the investigation of AI techniques and approaches. At the end of the 1980s, machine translation entered a period of innovation in methodology which has changed the framework of research. In 1981 came the first translation software for the newly introduced personal computers, and gradually MT came into more widespread use.

During the 1980s MT advanced rapidly on many fronts. The dominance of the rule-based approach waned in the late 1980s with the emergence of new methods and strategies loosely called `corpus-based' approaches, which did not require any syntactic or semantic rules in text analysis or selection of lexical equivalents. The major reason for this change has been a paradigm shift away from linguistic/rule-based methods towards empirical/data-driven methods in MT. This has been made possible by the availability of large amounts of training data and large computational resources (Hutchins, 1994).

3.3 Approaches to Machine Translation

Different researches efforts have been done to explore the possibility of automatic translation of one language to other language. The different method of machine translation have been explained in the next section

3.4 Rule-based Machine Translation

Rule-based machine translation relies on countless built-in linguistic rules and very large of number of bilingual dictionaries for each language pair. The approach essentially relied on linguistic rules such as rules for syntactic analysis, lexical transfer, syntactic generation, morphology, lexical rules, etc. (Kim, 2010). The assumption of rule-based MT is that translation is a process requiring the analysis and representation of the 'meaning' of source language texts and the generation of equivalent target language texts based on conversion of the source language structure to the target language structure (Sarkhel et al, 2010). Representations should be unambiguous lexically and structurally. There have been three basic approaches under rule based machine translation including transfer-based, Interlingua and dictionary-based machine translations approaches (source).

3.4.1 The Direct Approach

Direct translation approach is historically the earliest and known as the first generation of MT systems employed around from the 1950s to 60s when a need for machine translation was mounting. The Direct translation approaches are designed for translating one particular pair of language, called source language (SL) directly to another language, called target language (TL) without any intermediate representation, e.g. Geez as the language of the original texts, and Amharic as the language of the translated texts (source). This procedure involves taking a string of words from the source language, removing morphological infections from the words to obtain the lemmas, i.e. base forms and then looking up the lemmas in a bilingual dictionary between the source and target language. After a translation of each word is found, the positions of the words in the string are altered to best match the word order of the target language; these may include Subject-Verb-Object (SVO) rearrangements.

Since direct MT treats a sentence as a string of words and does not require syntactic or semantic analysis, it lacks taking structure, interdependencies, and semantic grouping among words can be lost and this will lead to the wrong interpretation for a given word. Generally the translation is performed through a direct translation of each individual Source Language word to the corresponding target Language words and followed by reorganizing the sentence structure. (Ramanathan et.al, 2002).

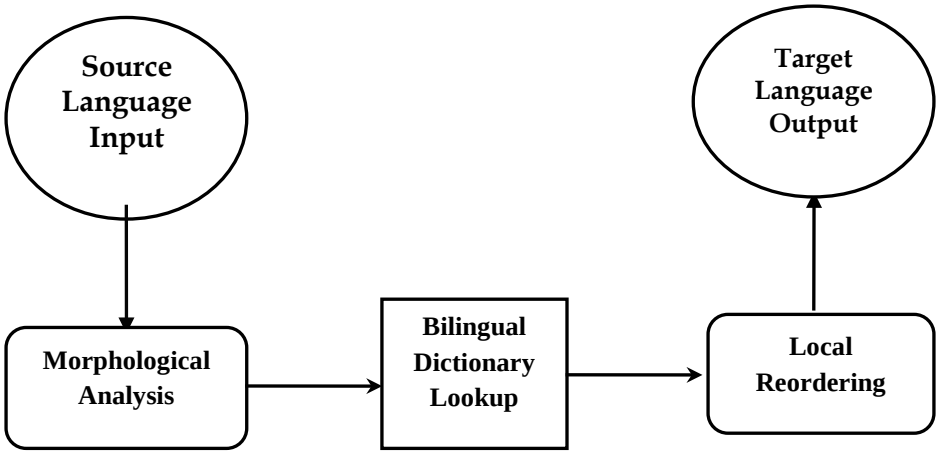

Figure 3.4.1.5.4 -1 Rule Based Machine Translation -Direct Approach

Local reordering is supposed to take some account of the grammar of the target language in putting the target words in the right order. The following examples of Geez-Amharic direct translation illustrate the process in the approach:

Original Geez Sentence	Ane Emetse Betike
Morphological Analysis	Ane Mesta + 1nd Person Singular + FUTURE Bet + 2nd Person Singular.
Identification of sentence parts	“Ane “, ” Mesta + 2nd Person Singular + FUTURE “, “Bet + 2nd Person Singular”
Reorder	“Ane “, “Bet + 2nd Person Single” , Mesta + 1nd Person Single.+FUTURE “
Dictionary Lookup	“Ene “, “Bet + 2nd Person Single ”, “Meta + 1nd Per. Sing”
Inflect	Ene Beteh Emetalehu
English Equivalent	I will come to your home

Table 1.5.4.1-1 Rule-based Machine Translation - Direct Approach

3.4.2 The Transfer Approach

The transfer approach is used on the basis of the known structural differences between the source and target language. A transfer system can be broken down into three stages: Analysis, Transfer and Generation. In the analysis stage, the source language sentence is parsed and converted to abstract Source Language oriented representations and structure. This is then input to a special component, called a transfer component, where transformations are applied to the source language oriented representations to convert the structure to that of equivalent Transfer Language (target language) oriented representations. The generation stage generates the final target language texts (Ramanathan, 2002). The rule Base Transfer approach also addresses the problem of language differences by adding structural and phrasal knowledge to the limitation of direct approach.

Transfer systems consist typically of three types of dictionaries: SL dictionaries containing detailed morphological, grammatical and semantic information, similar TL dictionaries, and a bilingual 'transfer' dictionary relating base SL forms and base TL forms and various grammars (for SL analysis, TL synthesis and for transformation of SL structures into TL forms) (Hutchins, 1994)

An advantage of transfer is that when you are using similar languages, which share the same syntax at times parts of the transfer system can be shared. In the direct approach, words are translated directly without passing through an additional representation. While, in the transfer approach the source language is transformed into an abstract and less language specific representation. Hence, for a system that handles the translation of

combination of n languages, n number of analysis, n number of generation components and n(n-1) transfer components are required.

Original Geez Sentence	nehene abasna mesele abawina
Analysis stage – close to Geez structure	abesna mesele abawina
Transfer Stage – converted to Amharic word ordered	mesele abawina abesna
Generation Stage – substitute Geez words with Amharic words	ende abatochachin bedelen
English Meaning	We sinned like our fathers

Table 1.5.4-1 Rule Base Machine Translation-Transfer Approach sample

3.4.3 The Interlingua Approach

Interlingua approach intends to translate source language texts to that of more than one language through an intermediate artificial language independent form called Interlingua. The Translation is from source language to Interlingua and then from Interlingua to target language.

Basically, the Interlingua approach consists of two stage processes: analysis and Synthesis. The analysis process is the extraction and complete representation of the meaning of the source language sentence in a language-independent form using a set of universal concepts and relations, and the synthesis phase generate a natural language sentence using a generation module between the representation language and the target language (Bonnie, 2004). The Interlingua allows a canonical representation of any

sentence, in the sense that all sentences that mean the same thing are represented in the same way, irrespective of the language.

The Interlingua approach is more economic approach in multilingual machine translations. Translation from and into n languages requires 2n interlingua programs where it requires n (n-1) bilingual translation systems in the direct and transfer translation systems. The Interlingua approach overcomes the problem of building thousands of transfer rules by using a central representation into which and from which all the languages are parsed and generated. Therefore, it is more efficient in exploiting the domain knowledge.

On the other hand, the complexity of the Interlingua itself is greatly increased. Finding a language independent representation which retains the precise meaning of a sentence in a particular language, which can then be used to generate a sentence of a different language, is a challenging task and maybe even impossible for a wider domain. (Alansary et al, 2006).

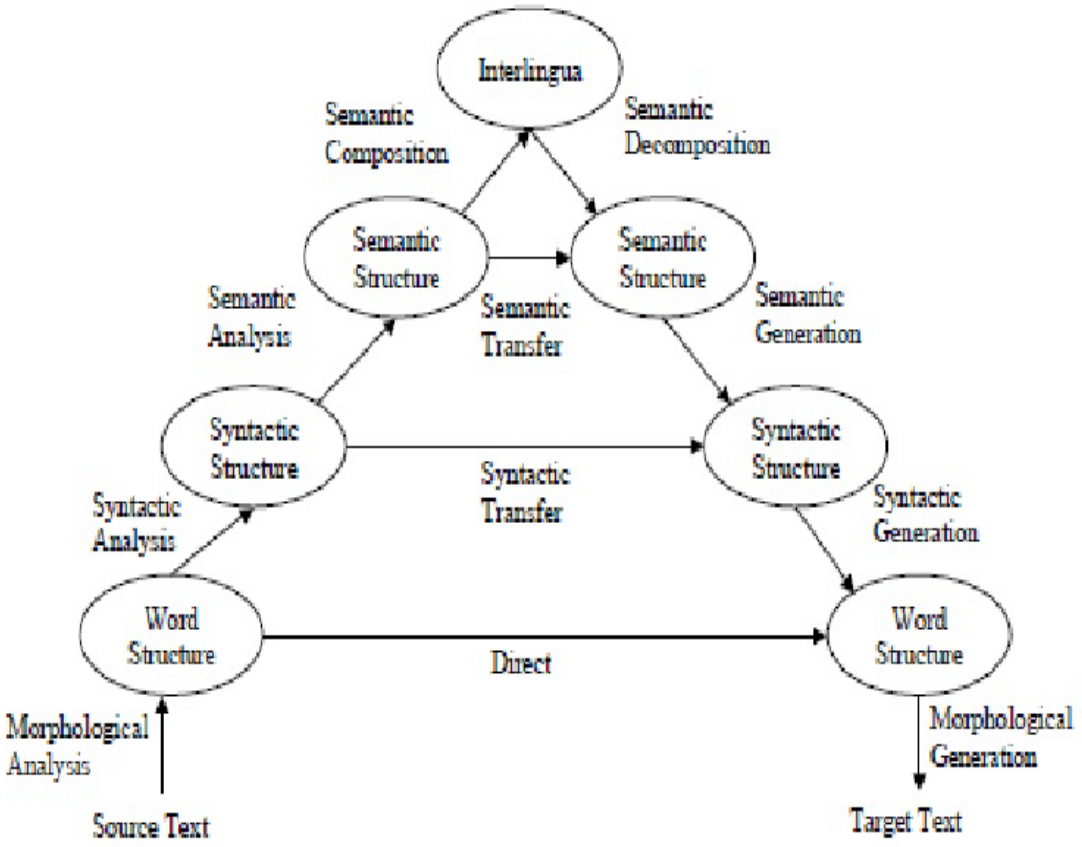

Figure 3.4.3-1 Vauquois Triangle for Rule-Based MT

The Vauquois Triangle in the Figure 0 above shows the pyramid of the rule based machine translation (Jurafsky et.al, 2006). It is evident that as we move up in the triangle towards an Interlingua, the burden on the analysis and generation components increases.

3.5 Corpus Based Machine Translation

The corpus based (Empirical) machine translation has been dominating the traditional rule based (Classical) ones since the late 1980s. The rule based has been requiring human encoded linguistic knowledge and intensive representation of the languages through different structural and language rules. The relative failure of rule-based approaches, the growing availability of machine readable parallel corpus (collection of

source language document with its counterpart target language documents) and the increase in capability of hardware (CPU, memory, disk space) with decreasing in cost are among the critical factors for the flourishing of corpus based machine translation systems. In addition, the freeing of corpus based approaches from any syntactic or semantic rules in text analysis or selection of lexical equivalents has contributed significantly. Corpus-base Machine translation includes Example Based Machine Translation (EBMT) and SMT. The following is the description of these two corpus based approaches.

3.6 Example-Based Machine Translation Approach

The idea for Example-based Machine Translation can be dated to a conference paper presented by Makoto Nagao in 1981 and later published in 1984. However, EBMT was only developed from about 1990 onwards (Hutchines, 2003). The underlying hypothesis of EBMT is that translation often involves the finding or matching of analogous examples, a pair (or couple) of texts in two languages that are a translation of each other, that have been translated before.

EBMT considers a bilingual corpus as a database and retrieves examples that are similar to an input sentence (texts). The input sentences can be of any size at any linguistic level: words, phrase, sentence, and even paragraph (Gros, 2007). The approach is founded on processes of extracting and selecting equivalent phrases or word groups from a databank of parallel bilingual texts, which have been aligned either by statistical methods or by more traditional rule-based methods - semantic network or a hierarchy (thesaurus) of domain terms. The essence EBMT is based on analogy principle from the previous examples. The analogy for EBMT mostly elucidated by Nagao's much quoted statement:

> *"Man does not translate a simple sentence by doing deep linguistic analysis, rather, Man does translation, first, by properly decomposing an input sentence into certain fragmental phrases ..., then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference." (Nagao, 1984).*

EBMT is also called "memory-based", "case-based", "experience-guided", "example-guided inference", or "analogy based" (Carl, 2002). EBMT is machine translation by example-guided inference. TM is an interactive tool for the human translator, while EBMT is an essentially automatic translation technique or methodology. He stressed the notion of detecting similarity. The basic processes of EBMT are analogy-based that is the search for phrases in the database which are similar to input source language strings (isolated by segmentation), their adaptation and recombination as target language phrases and sentences. Retrieving similar examples to the input is done by measuring the distance of the input to each of examples. The smaller a distance is, the more similar the example is to the input. EBMT uses real language data based on data-driven rather than theory-driven, overcoming constraints of structure preservation. The basic units for EBMT are thus sequences of words (phrases).

3.6.1 Stages of Example-Based Machine Translation

In general, there are four stages involved in EBMT, namely, example acquisition, example base management, example application and target sentence synthesis (Kit, 2001).

The first stage is Example Acquisition which is about how to acquire examples from existing parallel bilingual corpus. The examples can be collected from bilingual dictionaries at the word level, bilingual corpora at the multiple-word level and at sub-sentential levels including idioms and collocations, multi-word terminology, and phrases. Text alignment is a necessary step towards example acquisition at various levels. The approaches to text alignment can be again categorized into two types, namely, resource-poor approaches which relies mainly on sentence length statistics, co-occurrence statistics and some limited lexical information, and the resource-rich approaches which uses whatever available and useful bilingual lexicon and glossary.

The second stage is Example Base Management which is about how examples are stored and maintained. The Example Base Management is a crucial component in an EBMT system as it handles the storage, edition (including addition, deletion and medication) and retrieval of examples. Thus, an efficient EB must be capable of handling a massive volume of examples (both the source and the target language) at an adequately high speed searching strategy.

The third stage, Example Application, is about how to make use of existing examples to do translation which involves the decomposition of an input sentence into examples and the conversion of the decomposed source texts into target texts.

The fourth stage is known as the Sentence Synthesis and Smoothing which is to compose a target sentence by putting the converted examples into a smoothly readable order, aiming at enhancing the readability of the target sentence after conversion. Since different languages have different syntax to the sentential structures and word order, simple chain up the translated fragments may not work. The language modeling used

may include from the simple fixed-order n-gram models (e.g. bi-gram or tri-gram model) to more sophisticated probabilistic context free grammar models (Kit, 2001).

After sentence decomposition and example transfer, we have a sequence of translated fragments. The next task is to combine these translated chunks into a well-formed highly readable sentence in the target language. Since different languages have different syntax to govern the sentential structures and word order, it won't work in most cases if we simply chain up the translated fragments in the same order as in the source language.

Lexical EBMT systems use the surface form of texts directly. Because finding very similar sentences in the surface form is rare, lexical EBMT systems typically use partial matches (Brown et al, 2009) or phrase unit matches (Veale, 1997). To find hypothesis translations, they collect the translations of the matches for use in decoding. To increase coverage, lexical EBMT systems optionally perform generalization on the surface form to find translation templates.

Other EBMT systems use linguistic structures to calculate similarity. Some convert both source and target sentences in the example database into parse trees, and when they are given an input sentence, they parse it and calculate similarity to the stored example parse trees. They then select the most similar source parse trees with their corresponding target trees to generate target sentences after properly modifying them by the difference (Kurohashi, 2004). Or they find source sub tree matches with their aligned target sub trees and combine the target parts to generate target sentences (Menezes, 2006). In EBMT, the ‘classical’ similarity measure is the use of a thesaurus to compute word similarity on the basis of meaning or usage (Alexander et al, 2010).

3.7 Statistical Machine Translation

Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. Statistical Machine Translation (SMT) is a probabilistic framework for translating text from one natural language to another based on models induced automatically from analysis of parallel corpus (Axelrod, 2006). The general objective of SMT is to extract general translation rules from a given corpus consisting of sufficient number of sentence pairs which are aligned to each other (Mukesh et al, 2010).

Interest in SMT can be attributed to the convergence of several factors. The first factor is the growth of internet that is escalating the interest in the dissemination of information in multiple languages. The other factor is the availability of fast and cheap computing hardware has enabling applications that depend on large data and billions of statistics taken under translation. The development of automatic translation metrics and advance in freely available SMT toolkits are the other factors (Chang, 1992) and (Lopez, 2007).

The first statistical approach to MT was suggested by Warren Weaver in 1949 but pioneered by a group of researchers from IBM in the late 1980s (Brown et al, 1990). The idea behind SMT comes from information theory and is one of the applications of Noisy Channel Model which is proposed by Claude Shannon in 1948 in the field of Information Theory. It is based on statistical finding of the most probable translation from a large of pairs of equivalent source sentences and target sentences. For every pair of strings (a, g) a number Pr (g|a) is assigned which is the probability that a translator will produce g as his translation given a (Brown et al, 1993). By analogy with communication theory, Pr(a)

is a known “source” distribution, Pr(g|a) is a model of the process that encodes (or corrupts) it into the observed sentence g , and the argmax is a decoding operation (Goutte et al, 2009). Using Bayes' theorem, we can write:

$$\mathbf{Pr(a|g) = \frac{Pr(a)Pr(g|a)}{Pr(g)}}$$

Where Pr(a) is the language model probability, and where Pr (g|a) is the translation model probability. A document is translated according to the probability distribution p(a|g) that a string a in the target language (for example, Amharic) is the translation of a string g in the source language (for example, Geez). For each sentence in *A* is a translation of *g* with some probability, and the sentence that we choose as the translation (ê) is the one that has the highest probability. In mathematical terms [Brown et al., 1990], because Pr(g) is fixed, the maximization of ê is thus equivalent to maximization of $\mathbf{Pr(a)Pr(g|a)}$ and we get.

In mathematical terms (Brown et al., 1990), because $\mathbf{Pr(g)}$ is fixed, the maximization of $\boldsymbol{Pr(a|g)}$ denoted by $\hat{\mathbf{e}}$ is thus equivalent to maximization of $\mathbf{Pr(a)Pr(g|a)}$ and we get:

$$\hat{e} = argmax_a Pr(g|a)\Pr(a)$$

Decoding Algorithm — **Translation Model** — **Language Model**

Where:

- $\mathbf{Pr(a)}$ - The Language model - provides a probability to each unit of text.

- $\mathbf{Pr(a|g)}$ - The Translation model - that provides the probabilities of possible translation pairs of the source sentence g given the translated sentence a.
- argmax_a - The Search algorithm (Decoder) - searching for the best translation from the given all possible translations based on the probability estimates $\mathbf{Pr(a)}$ and $\mathbf{Pr(a|g)}$ and performs the actual translation.

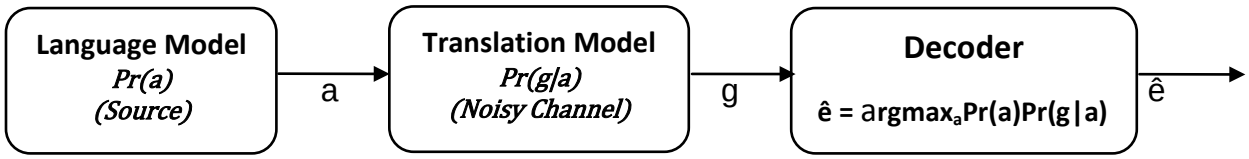

Figure 3.7-1 The Noisy Channel Model for Machine Translation

The Shannon's goal was to maximize the amount of information that could be transmitted over an imperfect (noisy) communication channel. It assumes that the original text has been accidentally scrambled or encrypted and the goal is to find out the original text by decoding the encrypted/scrambled version. In a probabilistic framework, finding the closest possible text can be stated as finding the argument that maximizes the probability of recovering the original input given the noisy text (Specia, Fundamental and New Approaches to Statistical Machine Translation, 2010).

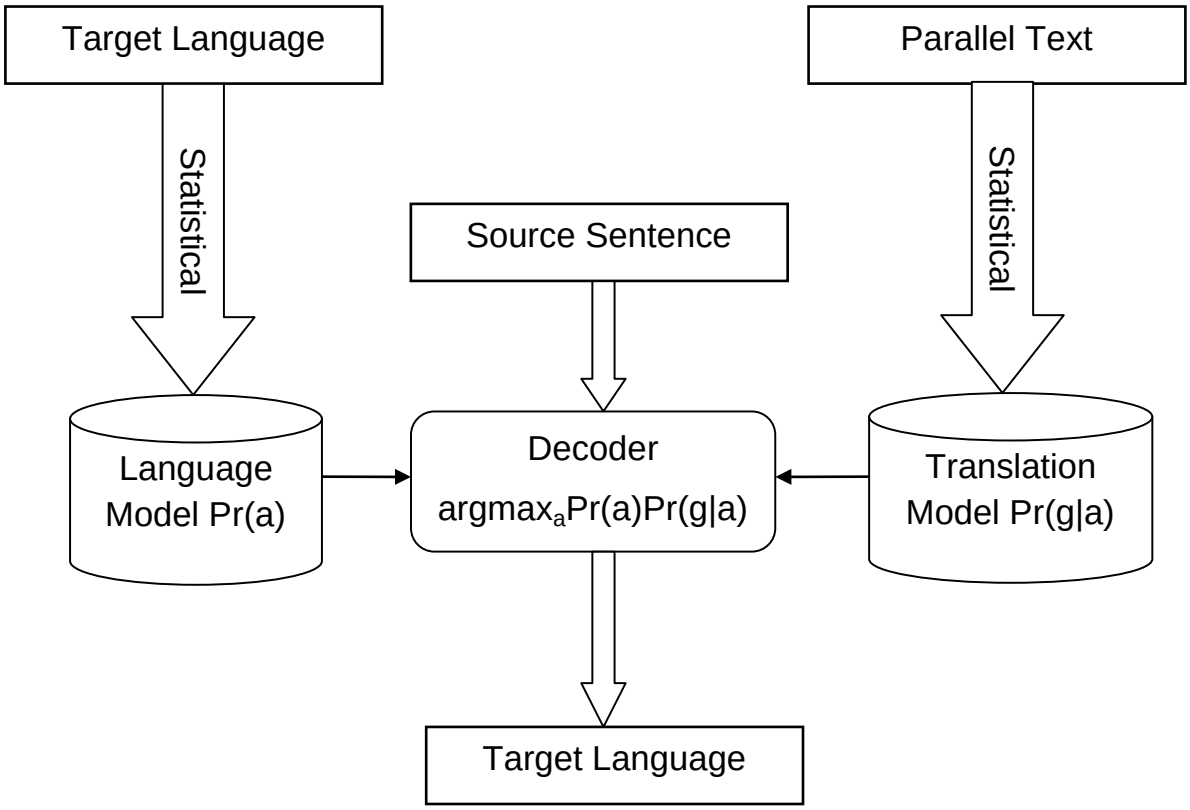

Figure 3.7-2 The SMT Process

A document is translated according to the probability distribution $\mathbf{Pr(a|g)}$ that a string $\mathbf{a}$ in the target language (for example, Amharic) is the translation of a string $\mathbf{g}$ in the source language (for example, Geez).

3.7.1 Components of Statistical Machine Translation

Statistical Machine Translation (SMT) usually consists of three components: a translation model Pr(g|a), a language model Pr(a) and a distortion model Pr(g,a) where g is an input sentence of a source language and a is an output sentence of a target language.

3.7.1.1 The Language Model

Language modeling is the process of determining the probability of a sequence of words. It has variety of applications in the area speech recognition, optical character recognition (OCR), handwriting recognition, Machine Translation, Information Retrieval, and Spelling Correction (Goodman, 2001). A statistical Language Model is a probabilistic way to capture regularities of a particular natural language in the form of word-order constraint (Rosenfeld, 2000). The Language Modeling component takes the monolingual corpus and produces the Language Model for the target language where plausible sequences of words are given high probabilities and nonsensical ones are given low probabilities. It generally reflects how frequently a string of words occur as a sentence.

Almost all language models decompose the probability of a sentence into conditional probability of the component words or phrases (Rosenfeld, 2000). Most language models are n-gram-based which is based on a sequence of n words. Given a word string $\mathbf{a}$ with n words $\mathbf{a} = w_i w_2 \ldots w_n$ the language model can be defined as the joint probability of a sequence of all words in the word string and can be written using chain rule as product of conditional probabilities as:

$$\Pr(\mathrm{a}) = \Pr(w_i, w_2, \ldots w_n) = \prod_{i=1}^{n} \Pr(w_i | w_i, w_2, \ldots w_{n-1})$$

Where w_i is the i^{th} word and n is the word length. The different language models used in SMT are discussed below with special focus on the n-gram model which is used in this research.

The N-gram Model

The n-gram model is the most dominant technique of Statistical Language Model which was proposed by Jelinek and Mercer (Bahl et al, 1983). The n-gram model assumes that the probability of the n^{th} word depends only on the ~~n~~1 preceding words based on the Markov assumptions. Markov assumes that only the prior local context consisting of last few words affects the next word. The N-gram thus has $(N-1)^{th}$ order of Markov Model (Jawaid, 2010). A high n provides a more information about the context of the specific sequence, but a low n provides more cases will have been seen in the training data and hence more reliable estimates. Most current open-domain systems consider n between 3 and 7 which actually varies according to the size of the corpus: the larger the corpus, the higher the n-grams that can be reliably counted.

Suppose we break an Amharic sentence ***a*** up into words $\boldsymbol{a=a_1a_2a_3 \ldots a_m}$ and assumed that the probability of seeing a word is independent of what came before it. Hence the probability to j^{th} word is given by:

$$\Pr(a_j) = \Pr(a_j|a_1, a_2, \ldots, a_{j-1}),$$

Then we can write the probability for ***a*** as a product of conditional probabilities:

$$\Pr(a) = \Pr(a_1)\Pr(a_2|a_1) \ldots \Pr(a_j|a_{j-1}) = \prod_{j=1}^{m} \Pr(a_i|a_1, a_2, \ldots, a_{j-1})$$

So, for example, in the sentence fragment “yihonal weym . . . ”, you would probably assign a very high conditional probability to the final word being “ayhonem”, certainly much higher than the probability of its occurrence at a random point in a piece of text.

We could estimate the probabilities $\Pr(a_j)$ by taking a very large corpus of Amharic text, and counting words. The bigram model, when n=2, assumes that the probability of a word occurring depends only on the word immediately before it:

$$\Pr(a_j|a_1, a_2, \ldots, a_{j-1}) = \Pr(a_j|a_{j-1}).$$

And the trigram model, when n=3, assumes that the probability of a word occurring depends only on the two words immediately before it:

$$\Pr(a_j|a_1, a_2, \ldots, a_{j-1}) = \Pr(a_j|a_{j-2}, a_{j-1}).$$

For instance, the trigram model considers two consecutive previous words as:

$$\Pr(\mathrm{a}) = \Pr(a_1)\Pr(a_2|a_1)\Pr(a_3|a_1a_2)\Pr(a_4|a_2\ a_3) \ldots \Pr(a_j|a_{j-2}, a_{j-1}).$$

The problem with this kind of training procedure is that it is likely to underestimate the probability of bigrams/trigrams which do not appear in the training set, and overestimate the probability of those which do. There are, for instance, n^2 possible bigrams and n^3 possible trigrams for a given n number of words in a training data. The next word following a given history can be reasonably predicted with the Maximum Likelihood Estimate (MLE) which predicts the next word based on the relative frequency of word sequences observed in the training corpus and the Count function used measures the number of times word was observed in the training corpus (Axelrod, 2006).

$$\mathbf{Pr_{MLE}}(a_j|a_{j-1}) = \frac{\mathrm{Count}(a_{j-1}, a_j)}{\mathrm{Count}(a_{j-1})}$$

Due to data sparseness and some uncommon words, the MLE is still unsuitable for statistical inference because of n-grams containing sequences of these words are unlikely

to be seen in any corpus. In addition, since the probability of a sentence is calculated as the product of the probabilities of component subsequences, these errors propagate and produce zero probability estimates for the sentence (Christopher et al, 1999). Hence, in order to address this problem the discounting or smoothing methods are devised which decrease the probability of previously seen events and assign the rest probability to the previously unseen events. The Smoothing methods allows for better estimators that allow for the possibility of sequences that did not appear in the training corpus. There are different smoothing methods including adding one, Good Turning Estimate, General Linear interpolation etc. The simplest smoothing algorithm is add-one to the count.

$$\Pr(a_j|a_{j-1}) = \frac{\text{count}(a_{j-1}a_j) + 1}{\text{count}(a_{j-1}) + V}$$

A common way to compare language model scores for different translation sentences is to compute the perplexity of such sentence as:

$$\frac{1}{N}\log_2 P(a)$$

Where N is the number of words in the translation sentence.

3.7.1.2 The Translation Model

Translation models define the bilingual relationship between the source and target strings of corresponding parallel corpora. Sometimes, a word in one language may have a cardinality of one-to-one, one-to-many, many-to-many or even one-to-zero. Despite these complications, the notion of a correspondence between words in the source language and in the target language is so useful. Most of state-of-the-art translation models used for regular text translation can be grouped into three categories: word-based models, phrase-

based models, and syntax-based models. The translation model probability cannot be reliably calculated based on the sentences as a unit, due to sparseness of the data. Instead, the sentences are decomposed into a sequence of words (Gao, 2011).

Determining the word alignment probabilities given sentence aligned training corpus is performed using the Expectation-Maximization (EM) algorithm. The key intuition behind Expectation Minimization is to determine the word translation probabilities from the number of times a word align with another in the corpus. The Expectation Minimization is formalized using techniques like IBM Models (Ramanathan et.al, 2002).

The statistical translation models were initially word based (Models 1-5 from IBM, Hidden Markov model from Stephan Vogel and Model 6 from Franz-Joseph Och, but significant advances were made with the introduction of phrase based models. Recent work has also incorporated syntax or quasi-syntactic structures.

I. Word - Based Translation

Word Based Translation Model is the original model for SMT where the fundamental unit of translation is a word in some natural language. The objective of word alignment is to discover the word to word translational correspondences in a bilingual corpus. It handles translation and alignment at word level with the assumption of all positions in the source sentence, including position zero for the null word, are equally likely to be chosen. The classical approaches to word alignment are based on series of IBM Models, Model 1 - 5, proposed from the IBM group a pioneer work at the very beginning of SMT in the early 1990s with increasing complexity, performance and assessment and the HMM based

alignment model and syntax based approaches for word alignment are also studied (Brown et al., 1993).

1. IBM Model 1

IBM Model 1, also called a lexical translation model, is the simplest and the most widely used word alignment model among the models that the IBM group has proposed. It uses an Expectation Maximization (EM) algorithm which works in an iterative fashion to estimate the optimal value for each alignment and translation probabilities in parallel texts. The IBM Model 1, given a Geez sentence G = *(g_1, . . . g_l)* of length *l* and Amharic sentence A = *(a_1, . . . , a_n)* of length n, ignores the order of the words in the source and target sentence and the probability of aligning word g_j and a_i is independent of their positions in string G and A, j and i respectively.

According to the noisy channel, IBM Model 1 try to identify a position j in the source sentence from which to generate the ith target word according to the distribution.

$$\boldsymbol{Pr}(g|a) = \frac{\varepsilon}{(\mathrm{j}+1)^{\mathrm{m}}} \prod_{\mathrm{j}=1}^{\mathrm{m}} \sum_{\mathrm{i}=0}^{\mathrm{l}} \mathrm{t}(g_j|a_i)$$

Where $\mathrm{t}(g_j|a_i)$ denotes the translation probability of g_j given a_i and ε *denotes* $P(m|a)$. We assume that all positions in the source sentence, including position zero for the null word, are equally likely to be chosen and there are $(l-1)^m$ acceptable alignments.

2. IBM Model 2

According to IBM Model 1, the word order was not cognized and the translation probabilities of the target words in any order are all the same. The first word in the source

language may appear in the last of the target language irrespective of their order. IBM Model 2 improves the reordering in Model 1 by adding an alignment model in addition to the lexical translation (IBM Model 1) such that words that follow each other in the source language have translations that follow each other in the target language. The alignment model S on a translation of a word in the i^{th} position of the source language to a word in the j^{th} position of the target language is given by:

$$s(i|j,l,m)$$

Which is the probability of connecting j^{th} word of Geez sentence of length m to i^{th} word of Amharic sentence of length l. Hence, the question become

$$\boldsymbol{Pr}(g|a) = \varepsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(g_j|a_i) S(i|j,l,m)$$

For example, the Amharic word “Neger gin/bengracene lay” cannot be translated word-by-word because the meaning of “Neger gin/bengracene lay” which cannot be constructed from the meanings of the component words of and course.

3. IBM Model 3

The IBM model 3 introduces the notion of fertility model to the IBM model 2. Typically, the number of words in translated sentences is different, because of compound words, morphology and idioms (Project, 2009). The ratio of the lengths of sequences of translated words is called fertility, which tells how many source words may be aligned to a specific number of target words.

Fertility is a mechanism to augment one word into several words or none and it is a conditional probability depending only on the lexicons. Often one word in the source is

aligned to one word in the target (fertility = 1) or to n multiple target words (fertility = n), or even zero target words (fertility = 0) (Gros, Survey of Machine Translation Evaluation, 2007). The IBM model 3 comprises of parameters of fertility probability $n(\varphi|a_i)$, translation probability $t(g|a_i)$ and distortion probability $d(j|i,m,l)$. Model 3 is deficient as it does not concentrate on all of its probability in for the sake of simplicity (Brownr et al, 1993).

$$\sum_{s}(m-\phi_s)p_1^{\phi_0}p_0^{m-2\phi_0}\prod_{i=1}^{l}n(\phi_i|e_i)\phi_i!\prod_{j=1}^{m}t\left(g_j\middle|e_{s_j}\right)d(j|s_j,m,l)$$

4. IBM model 4

IBM model 4 is one of the most successful alignment procedures so far with very complex distortion Model and many parameters that make it very complex to train (Cromières et al, 2009). IBM Model 4 further improves IBM Model 3 by providing a better formulation of the distortion probability distribution $s(i|j,l,m)$. The IBM Model 4 translation model is further decomposed into four sub models. Lexicon Model that represents probability of a word g in the Geez language being translated into a word a in the Amharic language, Fertility model which represent probability of a source word g generating n words, the Distortion Model which is concerned with the probability of distortion and the NULL Translation Model which is a fixed probability of inserting a NULL word after determining each target word (Watanabe et al, 2002). Model 4 replaces model 3's distortion parameters with the ones designed to model the way the set of source words generated by a single target word tends to behave as a unit for the purpose of assigning positions. In empirical

evaluations the IBM Model 4 has outperformed the other IBM Models and a Hidden Markov Model (HMM) (Lopez et al, 2005).

5. IBM Model 5

Model 5 is very much like Model 4, except that it is not deficient and the Models 1-4 are as stepping stones to the training of Model 5. IBM Models 3 and 4 are deficient (non-normalized) in that they can place multiple target words to the same position. But, IBM Model 5 eliminates this deficiency by keeping track of the number of vacant word positions and allowing for placement only into these positions (Specia, 2010). The IBM model 5 altered the distortion probability of model 4 to take into account all information about vacant positions which also brought a problem due to sparse data that have different permutations for vacant positions. The IBM models do not consider structural aspects of language, and it is suspected that these models are not good enough for structurally dissimilar languages.

In addition to the IBM models, there have been other models proposed including the more popular Hidden-Markov Models (HMM). The HMM is the other word-by-word alignment model where words of the source language are first clustered into a number of word classes, and then a set of transition parameters is estimated for each word class. The HMM models are similar to model 2 and use the first-order model. The characteristic feature of HMM is to reduce the number of parameters and make the alignment probabilities explicitly dependent on the alignment position of the previous word (Vogel, 1996).

All IBM Models are relevant for SMT since the final word alignment can be produced iteratively starting from Model 1 and finishing with Model 5. The limitations of word based models are their capability in managing word reordering, fertility, null words, contextual information, and non-compositional phrases (Axelrod, 2006). To make word-based translation systems manage, for instance, high fertility rates, and the system could be able to map a single word to multiple words, but not vice versa. For instance, if we are translating from Geez to Amharic, each word in Amharic could produce zero or more Geez words. But there's no way to group Amharic words to produce a single Geez word. Nowadays, the word-based translation is not widely used and has been improved upon by recent phrase based approaches to SMT, which use larger chunks of language as their basis (Project, 2009).

II. Phrase-based Translation

In real translation, it is common for adjacent sequences of words to translate as a unit. In phrase-based translation, the restrictions produced by word-based translation have been tried to reduce by translating any contiguous sequence of words. The multi-word segment of words are called blocks or phrases which are not linguistic phrases, such as a noun phrase but phrases found using statistical methods from the corpus. The segment phrases of the given source sentence are translated and then reordered to produce the target sentence (Zens et al, 2004). Restricting the phrases to linguistic phrases has been shown to decrease translation quality. Phrase-based translation models use the automatically generated word level alignments from the IBM models to extract phrase-pair alignments.

The phrase-based method has become the widely adopted one among all the proposed approaches in SMT due to its capability of capturing local context information from adjacent words to a one-to-many and many-to-many alignment which word-aligned translation models do not permit. It bases on phrase alignments instead of word alignments. One of the advantages of phrase based SMT systems is that the local reordering is possible and each source phrase is nonempty and translates to exactly one nonempty target phrase.

III. Syntax-based Translation

Syntax-based translation is based on the idea of translating syntactic units, rather than single words (as in word-based MT), or strings of words (as in phrase-based MT). The idea of syntax-based translation is quite old in MT, though its statistical counterpart did not take off until the advent of strong stochastic parsers in the 1990s. The syntax-based statistical translation model that includes in addition to word translation probabilities, the probabilities of nodes being reordered and words being added to particular nodes were proposed (Ramanathan, 2009). Examples of this approach include DOP-based MT and, more recently, synchronous context-free grammars. Syntax-based MT systems are slow to train and decode because the syntactic annotations further add a level of complexity.

Generally, Phrase-level alignments have been the state of the art and recent focus in SMT research outperforming the syntax-based translation model and word-based models.

3.7.1.3 Decoding

Decoding is the process of determining the most probable translation among all possible translations based on a searching algorithm. The search space is so huge because of

different possible translation for each word (phrase) with different ordering in sentence. Different decoding algorithms were proposed for SMT. Most of these decoding algorithms are based on partial sentence evaluation as it is not possible to find the best translation.

In order to solve decoding problem, most decoding algorithms are finding optimum solution instead of best solution. The Beam search algorithm, Greedy decoder and stack decoding algorithm are some examples. Most decoders in the SMT are based on the best-first search (Jurafsky et al, 2006). The A* was the first of the best-first search that was proposed by IBM group and implemented on word to word SMT where the search hypotheses are managed in a priority queue (stack) ordered by their scores (Casacuberta, 2004). The beam search is the other best-first search that is implemented on the phrase based decoding in the Moses[5] system.

3.7.2 Evaluation

Evaluating the quality of a translation is an extremely subjective task, and disagreements about evaluation methodology are widespread. Nevertheless, evaluating MT results is important to know how good an MT system is and identify new development area to improvement in translation quality. White et al (1994) suggested three aspects of evaluating a translation through how well the translation represents the source text (adequacy), the extent to which the translation is a well formed and correct sentence (fluency), and comprehensiveness of the information for readers (informativeness) (O'Connell et al, 1994). Generally, MT evaluation can be performed through human or

[5] Moses is a statistical machine translation system that allows us to automatically train translation models for any language pair.

automated system. The following is a brief description of the human and automatic MT evaluations.

3.7.2.1 Human Evaluation

In early days, MT evaluation is mainly subjective and scores are given by human judges assign to the MT output from various perspectives. The results of human evaluation are usually expensive, time consuming and not repeatable (Lopez, 2007). Manual evaluation generally scores output sentences based on their informativeness, fluency, fidelity, fluency and the accuracy of their content. Although human evaluation is accurate and reliable, they are too time-consuming and expensive to be used to compare many different versions of a system.

3.7.2.2 Automatic Evaluations

Automatic evaluations of Machine Translation are based on evaluation metrics like precision. It compares system translation output with reference translations from the parallel corpus. The automatic evaluations are important as they run frequently and cost efficient. There are different Machine Translation evaluation algorithms including BLUE, NIST and WER. The most widely used metric, namely the Bilingual Evaluation Understudy (BLEU), considers not only single word matches between the output and the reference sentence, but also n-gram matches, up to some maximum n (Lopez, 2007). Callison-Burch et al (2012) claim that automatic evaluations are an imperfect substitute for human assessment of translation quality.

3.8 Challenges in Machine Translation

Machine translation is hard for many reasons. The availability, collection and usage of huge amount of digital text and format types are some of the challenges. The language ambiguity that could arise from lexically differences where a word can have more than one meaning due to Semantic (out of context), Syntactic (in a sentence) and Pragmatic (situations and context) meanings, Technical Verbs, paragraphs with symbols and Equations, and Abbreviated Word are very difficult to translate.

In addition, different languages use different structures for the same purpose, and the same structure for different purposes. The challenges that could arise from Idiomatic and Collocation expressions where whose meaning cannot be completely understood from the meanings of the component parts. The different forms of a word, representation of a single word in one language with group of words in another, vocabulary difficulties in identifying direct equivalent word a particular word are also other challenges of machine translation (Ramanathan et al, 2002). The other big challenge of the MT is Vocabulary Differences which also arise from the Languages difference in the way they lexically divide the conceptual space, and sometimes no direct equivalent can be found for a particular word of one language in another.

The automatic translation from one language to another is an extremely challenging task, mainly due to the fact that natural languages are ambiguous, context-dependent and ever-evolving.

CHAPTER FOUR

CORPUS PREPARATION AND SYSTEM ARCHITECTURE

The chapter discusses the experimental setup, software tools used, the hardware environment, architecture of the system, the data used for the experimentation of the research. Moreover, the process of the experimentation, the result and the analysis of the results are discussed in detail.

4.1 Experimental Setup

4.1.1 Data Collection and Preparation

As discussed in the literature review part, SMT requires a huge amount of monolingual and bilingual data. The monolingual corpus is required to estimate the right word orders that target language should look like and the bilingual, which is sentence-aligned, is used to build the translation model training and decoding purpose that determine the word (phrase) alignment between the two aligned sentences. Finding parallel corpus with good quality and plausibly enough size were major challenges faced in this study. The corpus cleanup, correction of sentence level alignment and correction of errors is time consuming, expensive and language expert requesting task.

There are very few digital data available in the bilingual data as the Geez language is mainly used as spoken language in the digital era. It is understood that the size of the corpus is a major performance bottleneck for corpus-based machine translation systems. SMT systems would be able to attain better performance with more training sets (Kashioka, 2005). So, the researcher has tried his best to collect as many parallel

documents (written in Geez and Amharic) as possible to make the system perform well. The researcher found electronic version of some books of the Old Testements of Geez Bible including Genesis (1582 sentenses), Exodus (1102 sentenses), Leviticus (887 sentenses), Numbers (1322 sentenses), Deuteronomy (994 sentenses), Joshua (671 sentenses), Judgth (640 sentenses), Ruth (90 sentenses) and Psalms (5127 sentenses) and the all books of the Amharic Bible on the web[6] [7]. In addition, other religious resources like the Praises of St. Mary (Wedase Mariam), Arganon and some editions of Hamer Magazine comprising 425 sentenses are also used.

The materials were inherently verse level aligned, with some exceptions in the Geez versions, which has reduced the task of sentence level alignment. The part of the collected data which were not aligned at sentence (verse) level were aligned manually to sentence (verse) level, furthermore cross checking have been made between the corresponding sentences to confirm that the parallel sentences are same. The researcher found that most of the corresponding sentences were the same but some verses were misaligned due to a verse in one of the Geez (Amharic) document were broken into more than one verse in the corresponding document. Cross checking and correction of the verse level alignment were done manually. The language expert is used for the cross ckeeking the correct alignment of the corpuses.

The collected data were in different formats as they are collected from different sources. Some were in HTML, MS-word, MS-Publisher and MS-Excel format. Subsequently, all the collected data are merged to MS-word format and subsequently aligned to

[6] http://bible.org/foreign/amharic

[7] http://www.ranhacohen.info/Biblia.html

verse/sentence level, cleaned for noisy characters and converted to plain text in UTF-8 format to suit with the data type requirement of the training tools to be used.

The bilingual corpuses available for the training is comprised of a total of 12, 840 Geez sentences (146, 320 words) and 12,840 Amharic Sentences (144,815 words). The size of the number of sentences is not comparable to the resource available in European Parliament Proceedings Parallel Corpus[8] 1996-2011 which contains about 2 million parallel sentences (French-English). In additional to the target sentences in the parallel corpus supplementary monolingual corpus used for the language modeling is collected from the Amharic version of Bible, praise St. Marry (Wedase Mariamand Arganon), and website which contains 26, 818 sentence (contain 328,140 words).

4.1.2 Organization of Data

Once all the necessary preliminary preparation and formatting tasks are done on the corpus, 90% of the bilingual data is allocated for training and the remaining 10% is allocated for testing considering the training requires a larger amount of data to learn better. The training set consists of 11,560 Geez sentences (126,650 words) and 11,560 Amharic sentences (125,252 words) which are used for training the translation model. Whereas, the test set data consists of 1,280 sentence of Geez and 1,280 sentences of Amharic which are used to do tuning and evaluate the accuracy of the translation.

4.1.3 Software Tools Used

The experiment of this research was conducted on 32bit Linux machine (Ubuntu 14.04) as an operating system platform. The Moses SMT system, which is a full-fledged,

[8] http://www.statmt.org/europarl/

dominant and the state-of-the-art tool for SMT that automatically train translation models for any language pair, used for the translation and modeling purpose (Philipp, 2007). In addition Moses can also use the language modeling tool IRSTLM, the word-alignment tool GIZA++, and resulting translations evaluating model BLUE (Koehn, 2007).

4.1.4 Language Model Training

The IRSTLM[9] language modeling toolkit is used to train language model for this research. The IRSTLM is a free and open source Language Modeling Toolkit that has features of different algorithms and data structures suitable to estimate, store, and access very large language models (Federico et al, 2007). The IRSTLM is Lesser General Public License (LGPL) licensed (like Moses) and therefore available for commercial use. It is compatible with language models created with other tools, such as the [10]SRILM Toolkit.

4.1.5 Word Alignment

Many of the challenges encountered in parallel text processing are related to sentence length and complexity, the number of clauses in a sentence and their relative order. The task of word-based alignment was done by finding relationship between words based on the statistical value they have in a given Geez-Amharic parallel corpus. GIZA++[11] is a freely available, widely used SMT toolkit that is used to train IBM Models 1-5 and an HMM word alignment model. This package also contains the source for the mkcls[12] tool which

9 *http://sourceforge.net/projects/irstlm/*

10 *www.speech.sri.com/projects/srilm/download.html*

11 http://www.statmt.org/moses/giza/GIZA++.html

12 http://www.statmt.org/moses/giza/mkcls.html.

generates the word classes necessary for training some of the alignment models (Och, 2003). In this research, the GIZA++ toolkit is used for the word alignment.

4.1.6 Decoding

Decoding is done using Moses decoder. The job of the Moses decoder is to find the highest scoring sentence in the target language (according to the translation model) corresponding to a given source sentence. The decoder also output possible ranked list of the translation candidates, and also supply various types of information about how it came to its decision (for instance the phrase-phrase correspondences) (Dechelotte, 2007). An efficient search algorithm finds quickly the highest probability translation among the exponential number of choices. Moses phrase-based decoder is used for this experiment.

4.1.7 Tuning

In order to find the optimal weights from the given possible translation Moses tuning algorithm is used. The optimal weights are those which maximize translation performance on a small set of parallel sentences (the tuning set). About 1000 bilingual Geez - Amharic sentences from the corpus identified for the testing set. The bilingual corpora used for the tuning are preprocessed with tokenization and cleaning processes.

4.1.8 Evaluation

There are different Machine Translation evaluation algorithms including BLUE, NIST and WER. BLUE (Bilingual Evaluation Understudy) is one of the famous evaluation methods that can be used to have a comparison among different Machine Translation systems (Zhu, 2001). BLEU scoring tool is used for the evaluation of the quality of the translation

system based on the familiarity to the researcher and applicability with Moses. The BLUE evaluate the quality of text which has been machine-translated from one natural language to another based on the degree of correspondence between a machine's output and that of a human professional human translation.

4.2 Architecture of the System

The architecture of the system is illustrated in Figure 5.1. The system takes bilingual and monolingual corpuses as inputs which are represented by pile of sheets and the processes are represented by Rounded rectangle. The data are preprocessed with different preprocessing tools in order to fit the tools' requirement. The preprocessing is discussed in detail the experimentation part.

The models are represented with rectangular cube. The translation modeling takes the bilingual corpus (both the Geez and Amharic sentences) and then segmented into a number of sequences of consecutive words (so-called phrases). Each Geez phrase is translated into an Amharic phrase, based on the noisy channel model translation model. The language model takes the target language, Amharic corpus, to determine the word order in the sentence formation.

During decoding, the decoder searches for the best translation from the given all possible translations based on the probability. Tuning finds the optimal weights for the linear model, where optimal weights are those which maximize translation performance on a small set of parallel sentences (the tuning set).

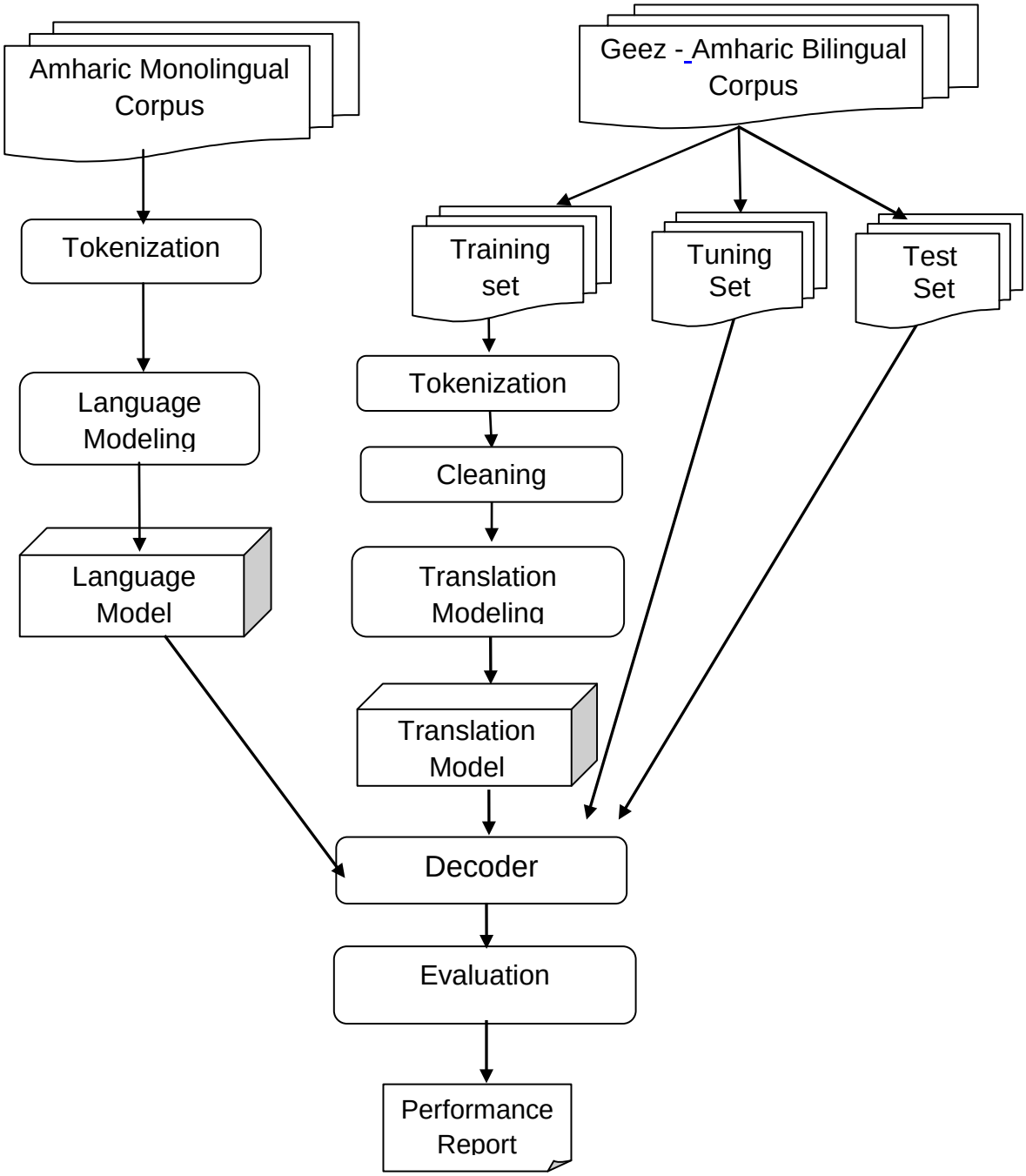

Figure 4.2-1 Architecture of Geez - Amharic SMT System

4.3 Preprocessing

Once data is converted into the right format (see section 5.2.1), it needs to be tokenized, and cleaned before it can be used to train a SMT system. Both the monolingual and parallel documents pass through a tokenization process to separate the words and make space between the words and punctuation marks which resolve the confusion of punctuation marks in a sentence. Cleaning is the other process of removing long sentences and empty sentences as they can cause problems with the training pipeline, misalignment of sentence. Cleaning is also important to reduce longer sentences which take extensive time in the process of translation training.

The Amharic monolingual corpus pass through tokenization only as the language modeling is not affected by long sentences as that of translation modeling. The preprocessing of both monolingual and bilingual corpus was done using the scripts written for this purpose[13] (http://www.statmt.org/moses).

[13] http://www.statmt.org/moses

CHAPTER FIVE

EXPERIMENT AND ANALYSIS

As discussed in section 5.2 phrase based SMT is used for this study. The portion of the Bible will be used to train and test the system performance. In this chapter the experimental procedures with the analysis of the experiment results will be presented.

5.1 Building and Testing the System

Once all the necessary preprocessing tasks are done, as discussed in section 5.4, on the corpus, out of the 12,840 parallel sentences 32 sentences, which are longer than 80 characters, are removed before the training as GIZA++ takes very long time and memory to train longer sentences.

5.2 Analysis of the Result

Part of the bilingual corpus (1280 bilingual sentences) left for the testing set was used for testing the system performance. Since the corpuses are already sentence level aligned and some cross checking has been made, the result was almost free of miss matched alignments. This has positively contributed to the performance of the result and BLUE score obtained were 8.14%. Further investigation has been done in order to crosscheck and improve the performance score as discussed below.

As there is a shortage of training data, a 10-fold cross validation (CV) method was used to determine the generalizability of the performance of the system. In 10-fold cross validation the data is iteratively divided in to 90% training and 10% testing set. The BLUE

score result obtained on the trails are 9.11%, 7.44%, 7.61%, 6.36%, 10.26%, 9.39%, 8.01%, 8.54% and 7.72%. The obtained result confirms the performance is relatively varied. Although the test data and the training set data are in similar domain, which is religious, the parts of the document varies in their content. The highest score obtained was 10.26% when the test data are taken from the part of the psalm whose part also available in the training set. The lowest score 6.36% was observed when the testing set from which contains the praise of Saint Mary and part of the Bible. The result verifies that the performance is highly dependent of the training and testing data domain. The discrepancy in the result could arise because of the part of document used for training may not contain the word in the test set and the system were not able to build the vocabulary of these sentences. The graphical representation of the 10-fold CV is shown below.

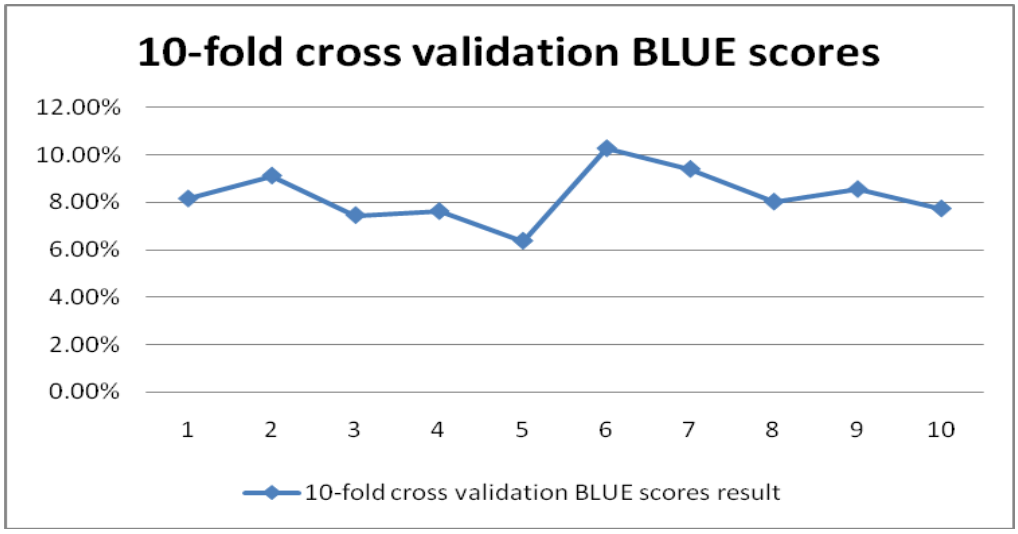

Figure 5.2-1 10-fold cross validation BLUE scores result

As discussed earlier on the literature review, performance of the system is dominant highly dependent of the training and testing data domain. A further experiment is conducted to test the assumption by splitting each version of the Bible in to 10% testing

set and the rest 90%for training set. The trials have been done three times in order to see the result and an average accuracy is calculated to compare with 10-fold CV test result.

Trail	**Performance of the system after splitting the each book of the Bible in to training and testing set.**
Trial 1	8.56%
Trial 2	8.23%
Trial 3	9.05%
Average performance	**8.61%**

Table 5.2-1 Performance of the system after splitting the each book of the Bible in to training and testing set.

As it is indicated in table 5.2-1 the average performance of the system shows, after splitting the each book of the Bible in to training and testing set, a better performance than the 10-fold CV result in which the each book of the Bible are not spirited in to training and testing. In addition the results after splitting each book of the bible in to training and testing sets has showen relative consistent. So the accuracy of this experiment confirms the performance is highly dependent on the training and testing set used.

5.2.1 Effect of the Language modeling corpus size

In an attempt to see the effect of the size of the language modeling data, the researcher has made the training by adding additional 13, 978 sentences (179,674 words) monolingual data to the original data used language modeling. The total monolingual corpus consisting of 26, 818 sentence (contain 328, 140 words) which is double of the

original monolingual data used for the language modeling. The monolingual data was obtained from Amharic Bible New Testament part, Praise of Saint Mary (Judases Miriam and Aragon) and Mahibere Kidusan websites[14]. The experiments are performed on four selected trial of the previous experiment after the addition of the monolingual corpus. The four trails are selected from the least, average and top Blue scores of the 10-CV. The resulting BLUE score positively fevered from 8.14% to 8.58%, 6.36% to 6.54%, 10.26% to 10.78% and 7.72% to 8.21%. A summary of the result obtained from the experiment and percentile difference is shown in the table Table 5.2.1-1.

Trail	**Before addition of Language modeling corpus size**	**After addition of Language modeling corpus size**	**Difference in percentile**
Trial 1	8.14%	8.58%	5.41%
Trial 5	6.36%	6.54%	2.83%
Trial 6	10.26%	10.78%	5.07%
Trial 10	7.72%	8.21%	6.35%
Average performance	**8.12%**	**8.53%**	**4.91%**

Table 5.2.1-1 Effect of language modeling corpus size

The result shows an average of an average 4.91% increment in the performance from the first training. From this, one can conclude that the performance of the system fevered by the increase the size of monolingual data used of the language modeling.

[14] www.eotcmk.org

Sample testing data set used	Obtained translation result before addition of Language modeling corpus size	Obtained translation result after addition of Language modeling corpus size	Reference sentence
እግዚአ ቦኡ አሕዛብ ውስተ ርስትከ	አሕዛብ ወይፍርሁ አቤቱ ስምህን	አቤቱ ስምህን አሕዛብ ወይፍርሁ	አቤቱ አሕዛብ ስምህን ይፍሩ
ወአውረድከኒ ውስተ መሬተ ሞት	ወደ ወአውረድከኒ በበሬዎችም ትቢያ	ወደ በበሬዎችም ትቢያ ወአውረድከኒ	ወደ ሞትም አፈር አወረድኸኝ
ወበመዝሙር ዘዐሠርቱ አውታሪሁ ዘምሩ ሎቱ	ወበመዝሙር ዘዐሠርቱ ዘምሩ አውታሪሁ ለእርሱ	ወበመዝሙር ዘዐሠርቱ አውታሪሁ ለእርሱ ዘምሩ	ዐሥር አውታርም ባለው በገና ዘምሩለት፡፡
ጎቤከ እግዚአ ጸራኅኩ አምላኪየ ወኢተጸመመኒ	አቤቱ ÷ ወደ አንተ ጮኽሁ ጮኽሁ ወኢተጸመመኒ አምላኬ ሆይ	አቤቱ ÷ አምላኬ ሆይ ÷ ወደ አንተ ጮኽሁ ወኢተጸመመኒ	አቤቱ አምላኬ ሆይ ወደ አንተ እጮኻለሁ ቸልም አትበለኝ፡፡
እስከ ማእዜኑ ትመይጥ ገጽከ እምኔየ	እስከ መቼ ትመይጥ ከእኔ ገጽከ ድረስ	እስከ መቼ ድረስ ትመይጥ ከእኔ ገጽከ	እስከ መቼ ፊትህን ከእኔ ትመልሳለህ
ወተንሥአ ንጉሥ ካልእ ዲበ ግብጽ ዘአያአምሮ ለዮሴፍ ።	ዮሴፍን ፥ ከዙፋኑም ተነሣ ፥ ሌላ ዘአያአምሮ በግብፅ ላይ ።	ከዙፋኑም ተነሣ ፥ ሌላ ዘአያአምሮ ዮሴፍን በግብፅ ላይ ።	በግብፅም ዮሴፍን ያላወቀ አዲስ ንጉሥ ተነሣ።
ወጎወጸሙ እግዚአብሔር ለደቂቀ እስራኤል ወተአምረ ሎሙ	እግዚአብሔር ወጎወጸሙ ልጆች ወተአምረ ።	እግዚአብሔር ወጎወጸሙ ለእስራኤልም ልጆች ወተአምረ ።	እግዚአብሔር የእስራኤልም ልጆች

Table 5.2.1-2 Comparison of the sample testing sentences translated before and after increase language modeling size

As it is shown in the table 5.2.1-2, the sample translations extracted from the testing set translation are in a better word order when the additional monolingual corpus is added to the original. The result agrees with the literature review that as a larger size language model corpus used result in better-performing language models and estimate good parameter values (Tucker, 1997). In addition, the data used for the language model are homogenous and it maintain better language model (Rose, 1997).

5.2.2 Effect of the Normalization of the Target Language

In the Amharic language writing system some words are written in different character combinations, as there are characters with the same sound having different symbol. For example, the characters 'ሀ', 'ሃ', 'ሐ', 'ሓ', 'ኀ and 'ኃ represent same sound. As a result the Amharic word "ስም" can also be written as "ሥም" where both refers to same word "Name" in English. Similarly one word 'እህት', can be written as 'እሕት', 'እኅት', 'ዕህት,' 'ዕሕት', ዕኅት' and 'ዕኽት' which all refer to one meaning in English "sister" While, in Ge'ez language writing system characters with same sound may produce different words. For example, the word "ሰዐሊ" and "ሰአሊ" have the same sound but different meaning - "draw a picture" and "beg for us" respectively. Similarly, "ሰረቀ" and "ሠረቀ" have same sound but different meaning - "He came" and "He stolen" respectively. Hence, normalization can be done for Amharic but not for Geez. The normalization of the Amharic words will reduce the data sparseness. The process of normalization was done by using a modified script written for this purpose by Solomon Mekonnen (Solomon, 2010). The normalization algorithm is depicted in the figure 5.3.2-1

```
Open corpus
While not end of corpus is reached do
        If a character is in normalization list
            Replace with its normalized char
        End if
 End while
close file
```

Figure 5.2.2-1 Normalization algorithm

The normalization has been performed on the previous original monolingual corpus and the training and testing also performed to see the effect. The results obtained are 8.28%, 6.44%, 10.46% and 7.84% respectively. As is in the table 5.2.2-1 the result show that here an average 1.62% increment on the performance. The finding supports the findings in other studies that the decrease in the data sparcity increases the performance of the translation (Sarkar, 2004).

Trial	**Before normalization**	**After normalization**	**Difference in percentile**
Trial 1	8.14%	8.28%	1.72%
Trial 5	6.36%	6.44%	1.26%
Trial 6	10.26%	10.46%	1.95%
Trial 10	7.72%	7.84%	1.55%
Average performance	**8.12%**	**8.26%**	**1.62%**

Table 5.2.2-1 Effect of Normalization

The number of unique words (vocabulary terms) before normalization of the target language was 27578 and it has decreased to 27376 after normalization. As presented in the Table 5.3.5-2, for example, the words “ሰው”, “ሠው” and “ሠዉ” which are same word, meaning “man” in English, can be represented by “ሰው”. It is conceivable that this causes the vocabulary table to have rather reduced sparse data.

Same words with different symbol before normalization		Words after normalization	
Word	**Number of occurrence**	**word**	**Number of occurrence**
ሰው	503	ሰው	510
ሠው	2		
ሠዉ	5		
ሄዱ	38	ሄዱ	40
ሔዱ	2		
አሥር	58	አስር	62
ዐሥር	4		
ዕጣን	16	እጣን	18
እጣን	2		
ዓይን	14	አይን	16
ዐይን	2		
ሺህ	131	ሺህ	135
ሺሕ	4		

Table 5.2.2-2 Sample same words with different symbol before and after normalization

The summary of the respective results before and after change in language model corpus size and normalization of the target language corpus are shown graphically in the figure 5.2.2-2

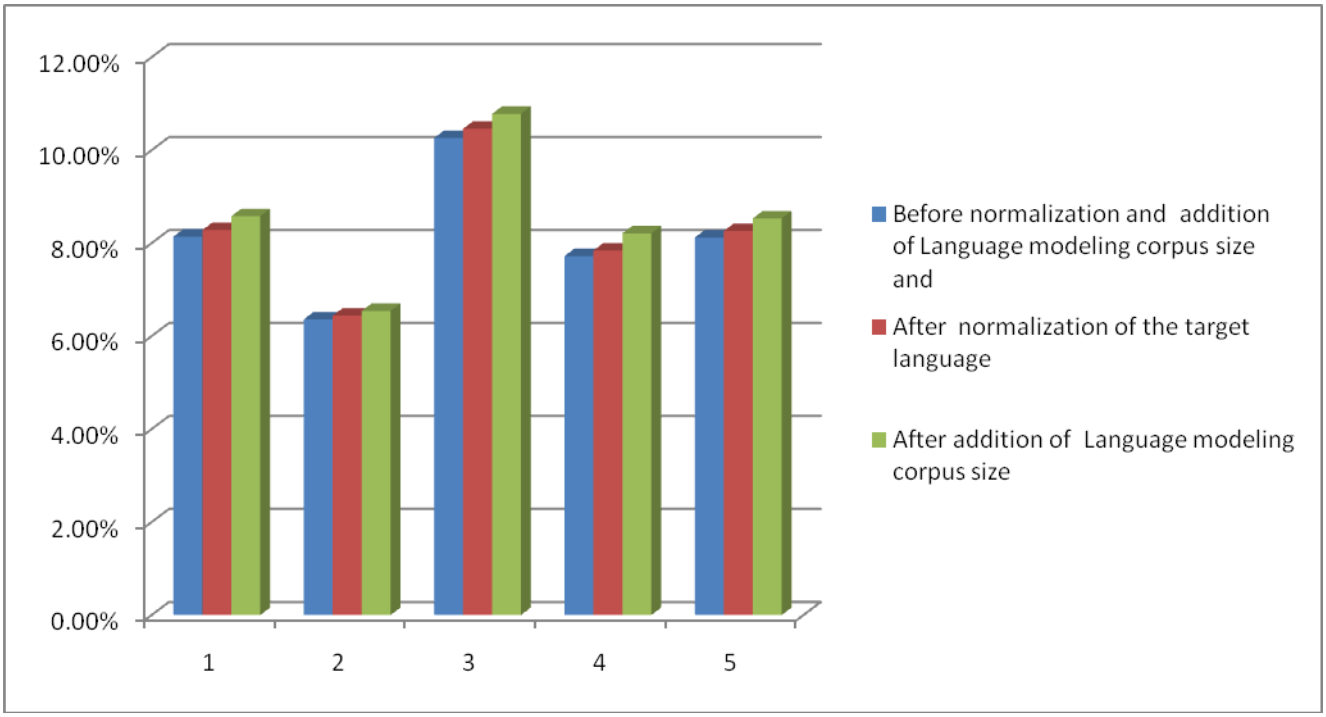

Figure 5.2.2-2 Performance of the system before and after addition of language model corpus size and normalization of target language

As is in the table 5.2.2-1 and table 5.2.2-2, the increase of additional monolingual corpus has showed a better performance as that of the normalization of the target language. Both results depicts the morphological synthesis can help for the better performance of the system.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The overall focus of this research is a Statistical Machine Translation (SMT) experiment from Geez to Amharic. SMT is the state-of-art in Machine Translation that is require huge amount of data and applicable approach. Although SMT requires a large amount of data in order to archive a good performance, the research was conducted with relatively small amount of data due to lack of fairly adequate amount of digital data in the languages. This has greatly affected the result. In this study, a phrase based SMT method was applied using the MOSES open source toolkit for SMT.

Accordingly, the average result that was achieved at the end of the experimentation was 8.26%. We have found that increasing the Amharic monolingual corpus can enhance the accuracy of the language modeling and the translation result. The accuracy is increased after normalization is applied to language model corpus. We also found that the normalization of the target language is a crucial factor in improving the accuracy of the translation by reducing the data scarcity. The performance of the system appears relatively low as compared to the performance of other experiments performed on huge amount of data. First reason for the low performance is the morphological complexity of the two languages. For instance, the BLUE score for Hebrew to Arabic translation (Shilon, 2012), which are both morphologically rich languages, is 14.3%. As well, the BLUE score for English to Affaan Oromo (Sisay, 2009) was 17.74%.

It is understood that corpus based translating between two morphologically rich languages poses challenges in morphological analysis, transfer and generation. The complex morphology induces inherent data scarcity problems, magnifying the limitation imposed by the dearth of available parallel corpora (Habash, 2006). Thus, as the two languages (both Amharic and Geez) are morphologically rich, less studied languages, and have little digital resources, the performance is relatively reasonable. The other reason for this is the size of data used for the training, as the larger the size of the corpus used the better the machine is able to do a high-quality translation (Kashioka, 2005).

6.2 Recommendations

Researches in statistical machine translation, generally in corpus based machine Translation, requires huge amount of bilingual and monolingual data in which the researcher faced a significant challenge to find digitally available data for the two languages especially Geez. Therefore; the researcher forwards the following recommendations as the extension of the current work and development of resources in both Amharic and Geez languages.

- As most of the available scripts in Geez are not converted into electronic format, the development of Optical Character Recognition (OCR) for languages will facilitate conversion of the scripts in both languages to digital format and hence easy access to this huge amount of manuscripts resources for smooth process of corpus based translation.
- This system does not perform well due to the limited size of the corpus. In addition, the training and testing data used are specific to religious contents.

Therefore, the researcher strongly recommends extending this research using a larger corpus size and various domains of contents other than the religious one.

- SMT from morphological complex language to morphological simple language yields better performance than between two morphological complex languages. The researcher believes that translation of Geez to other Ethiopian and International languages (like English) should be done to promote and extract the indigenous knowledge accumulated in the literatures of the Geez languages for the last thousands of years.
- Geez and Amharic are related but with scarce parallel corpora. Machine translation between the two languages is therefore challenging and requires exploring different approaches. Due to time constraints the researcher was not able to test the approach. The researcher recommends future research of Geez – Amharic translation should be undertaken using Example-based Machine Translation approach which is the other corpus based machine translation approach and requires relatively small amount of bilingual data for training (Dandapat, 2010).
- Geez and Amharic are related but morphology complex and limited researches have been done on the morphological segmentation and synthesizing of the two languages. The development of the languages' morphological synthesizers and segmenting tools can help for better performance. The researcher recommends extension of this research using the different morphological segmentation and synthesizing mechanisms.

Reference

Adam Lopez, Philip Resnik. (2005). Improved HMM Alignment Models for Languages with Scarce Resources. College Park: University of Maryland.

Adejumobi, S. A. (2007). The History of Ethiopia. Westport: Greenwood Press.

Alexander Clark, Chris Fox, Shalom Lappin. (2010). The Handbook of Computational Linguistics and Natural Language Processing. John Wiley & Sons.

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya and M. Sasikumar. (2002). Statistical Machine Translation. Mumbai.

Avik Sarkar, Anne De Roeck, Paul H Garthwaite. (1997). Technical Report on Easy measures for evaluating non-English corpora for language engineering. Some lessons from Arabic and Bengali. The Open University. United Kingdom.

Axelrod, A. E. (2006). Factored Language Models for Statistical Machine Translation. Edinburgh: University of Edinburgh.

Bahl, L. R., Jelinek, F., & Mercer, R. L. (1983). A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Transactions , PAMI-5 (2), 179 - 190.

Baye Yimam. (1992). Ethiopian Writing System. Addis Ababa University, Addis Ababa, Ethiopia. http://www.ethiopians.com/bayeyima.html on June 05, 2015

Bender, M. L., Sydeny W. Head, and Roger Cowley. (1976). The Ethiopian Writing System. In Bender et el (Eds.) Languages of Ethiopia. London: Oxford University Press.

Bj¨orn Gamb¨ack, F. O. (2009). Methods for Amharic Part-of-Speech Tagging. Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages (pp. 104-111). Athens, Greece: Association for Computational Linguistics.

Bjorn Gamback and Lars Asker. (2010). Experiences with developing language processing tools and corpora for Amharic. IST-Africa, 2010. Durban: Swedish Insitute of Computer Science AB.

Bonnie J. Dorr, Eduard H. Hovy and Lori S. Levin. (2004). Natural Language Processing and Machine Translation Machine Translation: Interlingual Methods.

Brown, P., S., Della Pietra, V., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics , 263-311.

Ceausu, A. (2011). Rich morpho-syntactic descriptors for factored machine translation with highly inflected languages as target. Dublin: unpublished .

Christopher D. Manning and Hinrich Schütze. (1999). Foundations of statistical natural language processing. Cambridge: MIT Press.

Chunyu Kit, H. P. (1991). Example-Based Machine Translation:A New Paradigm. Hong Kong: City University of Hong Kong.

Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster. (2009). Learning Machine Translation. London, England: Massachusetts Institute of Technology Press.

D.J. Arnold, L. B. (1994). Machine Translation: an Introductory Guide. London: Blackwells-NCC.

Daniel Jurafsky and James H. Martin. (2006). Speech and Language Processing: An introduction to natural language processing,. Chain: Prentice Hall.

Denkowski, M. C. Dyer, and A. Lavie. (2014). “Learning from post-editing: Online model adaptation for statistical machine translation,” in Proceedings of EACL, Gothenburg, Sweden.

Desie Keleb. (2003). Tinsae Geez - The Revival of Geez. Addis Ababa: EOTC Mahibere Kidusan.

Desta Berihu, Sebsibe Hailemariam, Zeradawit Adhana. (2011). Geez Verbs Morphology and Declaration Model. Addis Ababa: Department of computer science.

Dillmann, August. (2005). Ethiopic Grammar. Wipf & Stock Publishers. London: Williams and Norgate.

Fabien Cromières, Sadao Kurohashi. (2009). An alignment algorithm using belief propagation and a structure-based distortion model. EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 166-174). Stroudsburg: Association for Computational Linguistics .

Gambäck, S. E. (2005). Classifying Amharic News Text Using Self-Organizing Maps. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (pp. 71-78). Ann Arbor, Michiga: Workshop on Computational Approaches to Semitic Languages.

Gao, J. (2011). A Quirk Review of Translation Models.

Gerlach, J., V. Porro, P. Bouillon, S. Lehmann (2013). Combining pre-editing and post-editing to improve SMT of user-generated content. Preceding. of Workshop on Post-editing Technology and Practice, Nice, France.

Getahun Amare. (2001). Towards the Analysis of Ambiguity in Amharic. Journal of Ethiopian Studies, Vol. 34, No. 2. Institute of Ethiopian Studies. Ethiopia.

Goodman, J. T. (2001). A Bit of Progress in Language Modeling. Washington: Microsoft Research.

Gros, X. (2007). Survey of Machine Translation Evaluation. Saarbrucken, Germany: The EuroMatrix Project Co-ordinator.

Habash N, S. F. (2006). Arabic preprocessing schemes for statistical machine translation. The Association for Computational Linguistics. Moore RC, Bilmes JA, Chu-Carroll J, Sanderson M (eds) HLT-NAACL.

He, X. (2007). Using Word Dependent Transition Models in HMM based Word Alignment for Statistical Machine Translation. Association for Computational Linguistics (ACL), (pp. 80 - 87). Prague, Czech Republic.

Hetzron, Robert. (1997). The Semitic Languages. Taylor & Francis Group Publishing. New York. USA.

Hutchins, J. (2003). Example Based Machine Translation – a review and commentary.

Ilana Tahan.(2012), Ethiopic language collections, Retrieved on 10 May, 2015 from: http://www.bl.uk/reshelp/findhelplang/ethiopic/ethiopiancoll/

J. M. Harden. (1926). An Introduction to Ethiopic Christian Literature. The Macmilan company publishers.London. UK

Jawaid, B. (2010). Statistical Machine Translation between Languages with Significant Word Order Difference. Prague: Charles University in Prague.

Kim, J. D. (2010). Chunk alignment for Corpus-Based Machine Translation. Carnegie Mellon University.

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. Yorktown Heights, USA: IBM T. J. Watson Research Center.

Leslau, W. (1995). Reference Grammar of Amharic. Otto Harrassowit, Germany: Wiesbaden.

Lopez, A. (2007). A Survey of Statistical Machine Translation. Maryland: University of Maryland.

Marcello Federico, M. C. (2007). Efficient Handling of N-gram Language Models for Statistical Machine Translation. Proceedings of the Second Workshop on Statistical Machine Translation (pp. 88-95). Prague: Association for Computational Linguistics.

Michael Carl and Andy Way. (2002). Recent Advances in Example Based Machine Transaltion. Kluwer Academic Publishers.

Michel Galley, Daniel Cer, Daniel Jurafsky and Christopher D. Manning. (2009). Phrasal: A Toolkit for Statistical Machine Translation with Facilities for Extraction and Incorporation of Arbitrary Model Features, Stanford University, USA

Mukesh, G.S. Vatsa, Nikita Joshi, and Sumit Goswami. (2010). Statistical Machine Translation. DESIDOC Journal of Library & Information Technology,30(4),25-32.

Osborne, M. (2004). Statistical Machine Translation. United Kingdom : University of Edinburgh, .

Peter E Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert L. Mercer. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics Journal , 467-479.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar Alexandra Constantin and Evan Herbst (2007) Moses: Open Source Toolkit for Statistical Machine Translation, Proceedings of the ACL 2007

Demo and Poster Sessions, pages 177–180, Association for Computational Linguistics, Prague.

Popovic, Maja Hermann Ney. (2006). Statistical Machine Translation with a Small Amount of Bilingual Training Data. 5th LREC SALTMIL Workshop on Minority Languages, Genoa, Italy.

Project, P. L. (2009). Statistical Machine Translation for Bahasa Indonesia-English (BI-E) and English-Bahasa Indonesia (E-BI). Indonesia: Agency for the Assessment and Application of Technology.

Ramanathan, A. P. (2002). Statistical Machine Translation. Bombay: unpublished thesis paper Indian Institute of Technology.

Reshef Shilon, N. H. (2012). Machine Translation between Hebrew and Arabic. Machine Translation , 26 (1-2), 177-195.

Richard Zens, Franz Josef Och, and Hermann Ney. (2004). Phrase-Based Statistical Machine Translation. Germany: RWTH Aachen – University of Technology.

Robson, C. (1993) Real world research: A resource for social scientists and practitioner-researchers. Blackwell: Oxford; Cambridge.

Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? Proceedings of the IEEE , 88 (8), 1270–1278.

Rubin, A. D. (2010). A Brief Introduction to the Semitic Languages. USA: Gorgias Press.

Saba Amsalu and Dafydd Gibbon. (2004). Finite State Morphology of Amharic. Germany: Universit¨at Bielefeld.

Sameh Alansary, Magdy Nagi and Noha Adly. (2006). Towards a Language-Independent Universal Digital Library. Alexandria, Egypt..

Sandipan Dandapat, S. M. (2010). Statistically Motivated Example-based Machine Translation using Translation Memory. the 8th International Conference on Natural Language Processing. Kharagpur, India: Macmillan Publishers.

Schmidt, A. (2007). Statistical Machine Translation Between New Language Pairs Using Multiple Intermediaries. Heidelberg, Germany

Sisay Adugna. (2009). English-Afaan Oromo Machine Translation: An expererment using Statistical approach. Addis Ababa: Unpublished Thesis.

Sisay Fissaha Adafre. (2007). Part of Speech tagging for Amharic using Conditional Random Fields. Amsterdam: University of Amsterdam.

Sarkhel, Sneha Tripathi and JuranKrishna. (2010). Approaches to Machine Translation. Annals of Library and Information Studies .

Seretan V., Pierrette Bouillon, Johanna Gerlach. (2014). Large-Scale Evaluation of Pre-editing Strategies for Improving User-Generated Content Translation.

Solomon Mekonnen. (2010). Word Sense Disambiguation For Amharic Text: A Machine Learning Approach. Masters Thesis, Addis Ababa University.

Specia, L. (2010). Fundamental and New Approaches to Statistical Machine Translation. Wolverhampton: Unviersity of Wolverhampton.

Stephan Vogel, H. N. (1996). HMM-Based Word Alignment in Statistical Translation.

Taddesse Tamrat. (1972). Church and State in Ethiopia 1270 - 1527. Oxford University Press. London. UK

Taro Watanabe, Eiichiro Sumita. (2002). Statistical Machine Translation Decoder Based on Phrase. 7th International Conference on Spoken Language Processing , (pp. 1889-1892). Colorado, USA.

Thomas O. Lambdin. (1978). Introduction to Classical Ethiopic (Ge'ez). Harvard University. USA

Thurmair, G. (1991). Recent Developments in Machine Translation. Computers and the Humanities , 115-128.

Tony Rose, Tucker Roger and Nicholas Haddock. (1997). The Effects of Corpus Size and Homogeneity on Language Model Quality, Proceedings ACL-SIGDAT workshop on very large corpora, pp178-191, Hong Kong.

Ullendorff Edward. (1955). Semetic Lnguages of Ethiopia: A comparative Phonology. Taylor's Forign Press, London, England, UK.

Vogel, H. Ney, and C. Tillmann. (1996). HMM-based Word Alignment In Statistical Translation. In Proceedings of COLING.

W.J.Hutchins. (1994). Machine Translation: History and General Principles. The Encyclopedia of Languages and Linguistics , 5, 2322-2332.

White, J. S., O'Connell, T. and O'Mara, F. (1994). The ARPA MT Evaluation Methodologies:Evolution, Lessons, and Future Approaches. Proceedings of the First, (pp. 193–205). Columbia.

Wilker Ferreira Aziz, Thiago Alexandre Salgueiro Pardo and Ivandre Paraboni. (2007). an Experiment in Spanish-Portuguese Statistical Machine Translation. University of Sao Paulo, Brazil.

Zhu, K. P.-J. (2001). Bleu: a Method for Automatic Evaluation of Machine Translation. Yorktown Heights, USA: IBM T. J. Watson Research Center.

2 Appendix I

List of Amharic Normalization list

ዐ አ ዓ ኣ

ዑ ኡ

ዒ ኢ

ዔ ኤ

ዕ እ

ዖ ኦ

ሠ ሰ

ሡ ሱ

ሢ ሲ

ሣ ሳ

ሤ ሴ

ሥ ስ

ሦ ሶ

ፀ ጸ

ፁ ጹ

ፂ ጺ

ፃ ጻ

ፄ ጼ

ፅ ጽ

ፆ ጾ

ው ዉ

ሀ ሐ ሃ ሓ ኀ ኃ ኻ

ሁ ሑ ኁ ኹ

ሂ ሒ ኂ ኺ

ሄ ሔ ኄ ኼ

ህ ሕ ኅ ኽ

ሆ ሖ ኆ ኾ

3 Appendix II

Sample list of Geez Sentences used for testing with their Amharic equivalent translation

1. Translating: **እመንም ወሰብሑ ለስሙ**

   BEST TRANSLATION: **የተቀደሰውን ወሰብሑ**|UNK|UNK|UNK **እመንም**|UNK|UNK|UNK

2. Translating: **እስመ ኄር እግዚአብሔር እስመ ለዓለም ምሕረቱ**

   BEST TRANSLATION: **እግዚአብሔር ቸር ነውና ምሕረቱ ለዘላለም ነውና** [111111]

3. Translating: **መዝሙር ዘዳዊት**

   BEST TRANSLATION: **የዳዊት መዝሙር** [11]

4. Translating: **እዜምር ወእሌቡ ፍኖተ ንጹሕ**

   BEST TRANSLATION: **እዜምር**|UNK|UNK|UNK **መንገድ ወእሌቡ**|UNK|UNK| **ንጹሕ ነው**

5. Translating: **ወአሐውር በየዋሃተ ልብየ በማእከለ ቤትየ**

   BEST TRANSLATION: **ልቤ በየዋሃተ**|UNK|UNK|UNK **ወአሐውር**|UNK|UNK**በቤቴ መካከል**

6. Translating: **ወኢረሰይኩ ቅድመ አዕይንትየ ግብረ እኩየ**

   BEST TRANSLATION: **ወኢረሰይኩ**|UNK|UNK|UNK **ጌታዬ ዐይኖቼ ፊት ክፉ ሥራ**

7. Translating: **ጸላእኩ ገበርተ ዐመፃ**

   BEST TRANSLATION: **ዐመፅ ጠላሁ** [111]

8. Translating: **ሶበ ተግሕሠ እኩይ እምኔየ ኢያእመርኩ**

   BEST TRANSLATION: **ክፉ ጊዜ ተግሕሠ**|UNK|UNK|UNK **ከእኔ ኢያእመርኩ**|UNK|UNK|UNK [11111]

9. Translating: **ከመ ኣንብሮሙ ምስሌየ**

   BEST TRANSLATION: **ከእኔ ዘንድ ኣንብሮሙ**|UNK|UNK|UNK [111]

10. Translating: **ዘየሐውር በፍኖት ንጹሕ ውእቱ ይትለአከኒ**

    BEST TRANSLATION: **በመንገድ ዘየሐውር**|UNK|UNK|UNK **ይትለአከኒ**|UNK|UNK|UNK **ንጹሕ ነው** [11111]

11. Translating: **ወኢይነብር ማእከለ ቤትየ ዘይገብር ትዕቢተ**

    BEST TRANSLATION: **ይሁንልህ አንዳች በቤቴ መካከል ጋር አልቀመጥም** [11111]

12. Translating: **ወኢያረትዕ ቅድሜየ ዘይነብብ ዐመፃ**

    BEST TRANSLATION: **ዐመፃን ዘይነብብ**|UNK|UNK **ወኢያረትዕ**|UNK|UNK|UNK **በፊቴ**

13. Translating: **ስምዐኒ እግዚኦ ጸሎትየ**

    BEST TRANSLATION: **አቤቱ ጸሎቴን ስማኝ** [111]

14. Translating: **ወይብጻሕ ቅድሜከ ገዐርየ**

    BEST TRANSLATION: **በፊትህ ወይብጻሕ**|UNK|UNK|UNK **ገዐርየ**|UNK|UNK|UNK [111]

15. Translating: **ወኢትሚጥ ገጸከ እምኔየ በዕለተ መንዳቤየ**

    BEST TRANSLATION: **በመከራዬ መንዳቤየ**|UNK|UNK| **ፊትህን ከእኔ አትመልስ** [11111]

16. Translating: **አፅምእ እዝነከ ኀቤየ**

    BEST TRANSLATION: **ጆሮህን ወደ እኔ አዘንብል** [111]

17. Translating: **ወነገሥትኒ ከመ ይትቀነዩ ለእግዚአብሔር**

    BEST TRANSLATION: **እግዚአብሔርን እንደ ወነገሥትኒ**|UNK|UNK|UNK **ይትቀነዩ**|UNK|UNK|UNK [1111]

18. Translating: **እማንቱሰ ይትሐጐላ ወአንተሰ ትሄሉ**

    BEST TRANSLATION: **አንተ ይትሐጐላ**|UNK|UNK|UNK **እማንቱሰ**|UNK|UNK|UNK **ትሄሉ**|UNK|UNK|UNK [1111]

19. Translating: **ወኵሉ ከመ ልብስ ይበሊ**

    BEST TRANSLATION: **እንደ ልብስ ይበሊ**|UNK|UNK|UNK **ሁሉ** [1111]

20. Translating: **ወደቂቀ አግብርቲከ ይነብርዋ**

    BEST TRANSLATION: **እኛ ባሪያዎችህ ልጆች ይነብርዋ**|UNK|UNK|UNK [111]

21. Translating: **መዝሙር ዘዳዊት**

    BEST TRANSLATION: **የዳዊት መዝሙር** [11]

22. Translating: **ተባርኮ ነፍስየ ለእግዚአብሔር**

    BEST TRANSLATION: **ነፍሴ እግዚአብሔርን ተባርኮ**|UNK|UNK|UNK [111]

23. Translating: **ወኵሉ አዕጽምትየ ለስሙ ቅዱስ**

    BEST TRANSLATION: **አጥንቶቼ ሁሉ የተቀደሰውን** [1111]

24. Translating: ተባርኮ ነፍስየ ለእግዚአብሔር

BEST TRANSLATION: ነፍሴ እግዚአብሔርን ተባርኮ|UNK|UNK|UNK [111]

25. Translating: ዘይኬልለከ በሣህሉ ወበምሕረቱ

BEST TRANSLATION: ዘይኬልለከ|UNK|UNK|UNK በሣህሉ|UNK|UNK|UNK ወበምሕረቱ|UNK|UNK|UNK [111]

26. Translating: ዘይሔድሳ ከመ ንስር ለውርዙትከ

BEST TRANSLATION: እንደ ንስር የንጋትን ክንፍ ብወስድ ዘይሔድሳ|UNK|UNK|UNK ለውርዙትከ|UNK|UNK|UNK [1111]

27. Translating: ገባሬ ሣህል እግዚአብሔር

BEST TRANSLATION: እግዚአብሔር ሣህል|UNK|UNK|UNK ገባሬ|UNK|UNK|UNK [111]

28. Translating: ወለደቂቀ እስራኤል ሥምረቶ

BEST TRANSLATION: የእስራኤልም ልጆች ሥምረቶ|UNK|UNK|UNK [111]

29. Translating: መሓሪ ወመስተሣህል እግዚአብሔር

BEST TRANSLATION: እግዚአብሔር መሓሪና ይቅር ባይጻድቅ ነው [111]

30. Translating: ወታነብር መንጸረ ይወፅእ ፍትሕ አዝፋሪሁ ለስንሳሌሃ እምኵልሄ እምስማካቲሃ ወእምኵናኔ ታነብር

BEST TRANSLATION: ወታነብር|UNK|UNK|UNK መንጸረ|UNK|UNK|UNK ዘይትም እስኪወጣ አዝፋሪሁ|UNK|UNK|UNK ለስንሳሌሃ|UNK|UNK|UNK እምኵልሄ|UNK|UNK|UNK እምስማካቲሃ|UNK|UNK|UNK ወእምኵናኔ|UNK|UNK|UNK ታነብር|UNK|UNK|UNK

31. Translating: ወተሐርድ ላህሞ በቅድመ እግዚአብሔር በጎበ ኖኅተ ደብተራ ዘመርጡር

BEST TRANSLATION: ላህሞ|UNK|UNK|UNK ወተሐርድ|UNK|UNK|UNK በእግዚአብሔር ፊት በምስክሩ ድንኳን ደጃፍ አጠገብ [11111111]

32. Translating: ወታነብር መንገለ ምኵናን ዘፍትሕ ትእዛዘ ወርትዐ ወይትገበር ውስተ እንግድዓሁ ለአሮን ወሶበ ይበውእ ቤተ መቅደስ ቅድመ እግዚአብሔር ያብእ አሮን ፍትሐሙ ለውሉደ እስራኤል በውስተ እንግድዓሁ ቅድመ እግዚአብሔር ለዘልፍ

BEST TRANSLATION: ወታነብር|UNK|UNK|UNK ምኵናን|UNK|UNK|UNK በኩል ዘፍትሕ|UNK|UNK|UNK ወይትገበር|UNK|UNK|UNK ወርትዐ|UNK|UNK|UNK ትእዛዝ ወደ ጊዜ አሮንን በደረቱ ኪስ ወገኖች ወዳለት ያብእ|UNK|UNK|UNK ፍትሐሙ|UNK|UNK|UNK በእግዚአብሔር ፊት አሮንም ለእስራኤል ልጆች እንግድዓሁ|UNK|UNK|UNK ውስጥ በእግዚአብሔር ፊት ሁልጊዜ [11111111111111111111111111]

33. Translating: ወንሣእ በግዐ ካልአ ወያነብሩ አሮን ወደቂቁ እደዊሆሙ ላዕለ ርእሱ

    BEST TRANSLATION: ጠቦት ውሰድ ወያነብሩ|UNK|UNK|UNK ሊላ አሮንና ልጆቹ ራስ ላይ እጃቸውን [111111111]

34. Translating: ወይኩኖሙ ለአሮን ወለደቂቁ ሕገ ለዓለም በጎበ ውሉደ እስራኤል እስመ ፍልጣን ውእቱዝ ወፍልጣን ለይኩን በጎበ ውሉደ እስራኤል እምዝብሐተ ይዘብሑ ለፍርቃኖሙ ፍልጣን ለእግዚአብሔር

    BEST TRANSLATION: ለአሮንና ለልጆቹ ፊንታቸው ፡ ሕግ ዘንድ ከእስራኤል ልጆች ፍልጣን|UNK|UNK|UNK ወፍልጣን|UNK|UNK|UNK ይህ ዘንድ ይሁን የእስራኤልም ልጆች ይዘብሑ|UNK|UNK|UNK እምዝብሐተ|UNK|UNK|UNK ለፍርቃኖሙ|UNK|UNK|UNK ለእግዚአብሔር ፍልጣን|UNK|UNK|UNK [11111111111111111111]

35. Translating: ወትገብር ለአሮን ወለደቂቁ ከመዝ ኵሎ በከመ አዘዝኩከ ሰቡዐ ዕለተ ከመ ትፈጽም እደዊሆሙ

    BEST TRANSLATION: ለአሮንና ለልጆቹ ዘርፍ ላይ እንዲህ እንደ ሁሉ አዘዝኩከ|UNK|UNK|UNK ሰባት ቀን ትፈጽም|UNK|UNK ዘንድ በሚቃጠለው መሥዋዕት

36. Translating: ወአሐደ በግዐ ትገብር በጽባሕ ወአሐደ ፍና ሰርክ

    BEST TRANSLATION: ጠቦት ይጫኑ ታደርጋለህ ይጫኑ ማልዶ ተነሣ በመሸ ጊዜ

37. Translating: ወነበቦ እግዚአብሔር ለሙሴ ወይቤሎ

    BEST TRANSLATION: እግዚአብሔርም ሙሴን እንዲህ ብሎ ተናገረው [1111]

38. Translating: ወመሥዋዕተ ወማእደ ወኵሎ ንዋያ ወተቅዋመ ማኅቶት ንጽሕተ ወኵሎ ንዋያ

    BEST TRANSLATION: ወማእደ|UNK|UNK|UNK መሥዋዕት ሁሉ ፊት ለፊት ሁሉ ንዋያ|UNK|UNK|UNK ንጽሕተ|UNK|UNK|UNK ንዋያ|UNK|UNK|UNK [111111111]

39. Translating: ወአንተኒ አዝዞሙ ለደቂቀ እስራኤል ወበሎሙ ውቁ ከመ ትዕቀቡ ሰንበትየ እስመ ትእምርት ውእቱ በጎቤየ ወበጎቤክሙኒ በትውልድክሙ ከመ ታእምሩ ከመ አነ ውእቱ እግዚአብሔር ዘእቄድሰክሙ

    BEST TRANSLATION: የተናገረውን የእስራኤልን ልጆች እንዲህ ብለህ እዘዛቸው ያዘዝሁህን ነውና የምትቀመጡ እንደ ሆነ ሰንበትየ|UNK|UNK|UNK እርሱ ወበጎቤክሙኒ|UNK|UNK|UNK ታላቅ ነው በትውልድክሙ|UNK|UNK|UNK ታእምሩ|UNK|UNK|UNK ዘንድ ዘንድ የእርሱን ዘእቄድሰክሙ|UNK|UNK|UNK

40. Translating: ቡርክት አንቱ እምአንስት ወቡሩክ ፍሬ ከርሥኪ

    BEST TRANSLATION: አንቱ|UNK|UNK|UNK መልካም እምአንስት|UNK|UNK|UNK ፍሬ በነሣሽ በእግዚአብሔር የተባረከ ነው [111111]

41. Translating: ዐቢይ ውእቱ ስብሐተ ድንግል ናኪ ኦ ማርያም ድንግል

    BEST TRANSLATION: ክብር ታላቅ ድንግል|UNK|UNK|UNK ናኪ|UNK|UNK|UNK ኦ|UNK|UNK|UNK ድንግል|UNK|UNK|UNK ማርያምም [11111111]

42. Translating: እስመ ወለድኪ ለነ ንጉሥ መንክር ምሥጢር ኃደረ ላዕሌኪ

    BEST TRANSLATION: ወለድኪ|UNK|UNK|UNK ነውና ንጉሥ|UNK|UNK|UNK ማን ነው ? መንክር|UNK|UNK|UNK ምሥጢር|UNK|UNK|UNK ኃደረ|UNK|UNK|UNK ላዕሌኪ|UNK|UNK|UNK [11111111]

43. Translating: አንቲ ውእቱ ዕፅ ዘርእየ ሙሴ በነደ እሳት ወኢትውዒ

    BEST TRANSLATION: ነው ዘርእየ|UNK|UNK|UNK ከአንቺም ዛፍ ሙሴም እሳት ነበልባል ውስጥ ዐረገ ወኢትውዒ|UNK|UNK|UNK [11111111]

44. Translating: ኦ አዳም መሬት አንተ ወትገብእ ውስተ መሬት

    BEST TRANSLATION: ኦ|UNK|UNK|UNK መሬት|UNK|UNK|UNK አዳምም አንተ ወደ ትሄዳለህ መሬት|UNK|UNK|UNK [1111111]

45. Translating: ፈቀደ እግዚእ ያግዕዞ ለአዳም ጎዙነ ወትኩዘ ልብ ወያግብኦ ጎበዘትካት መንበሩ

    BEST TRANSLATION: እግዚአብሔርም አዳምም ሊገድለን ያግዕዞ|UNK|UNK|UNK ጎዙነ|UNK|UNK|UNK ወትኩዘ|UNK|UNK|UNK ልብ|UNK|UNK|UNK ወያግብኦ|UNK|UNK|UNK ጎበዘትካት|UNK|UNK|UNK መንበሩ|UNK|UNK|UNK

46. Translating: ነአምን በአሐዱ አምላክ እግዚአብሔር አብ አጋዜ ኩሉ

    BEST TRANSLATION: ነአምን|UNK|UNK|UNK እግዚአብሔርም በአንድ አብ|UNK|UNK|UNK አጋዜ|UNK|UNK|UNK ኩሉ [1111111]

47. Translating: ወትከብሪ እምድር

    BEST TRANSLATION: ወትከብሪ|UNK|UNK|UNK ከምድር ከፍ እንደሚል ÷ [11]

48. Translating: ላዕለ ጻድቃኑ ወላዕለ እለ ይመይጡ ልቦሙ ጎቤሁ

    BEST TRANSLATION: ጻድቃን ሆይ ÷ ላይ ላይ ይመይጡ|UNK|UNK| ወደ ልብ ሰዎች

4 Appendix III

Sample Sentenses used for training and testing

Geez sentences

እግዚኦ ቦኡ አሕዛብ ውስተ ርስትከ

ወአርኰሱ ጽርሐ መቅደስከ

ወረሰይዋ ለኢየሩሳሌም ከመ ልገተ ዐቃቤ ቀምሐ

ወረሰዩ አብድንቲሆሙ ለአግብርቲከ መብልያሙ ለአዕዋፈ ሰማይ

ወሥጋሆሙኒ ለጻድቃኒከ ለአረዊተ ገዳም

ከዐዉ ደሞሙ ከመ ማይ ዐውዳ ለኢየሩሳሌም

ወኀጥኡ ዘይቀብሮሙ

ወኮነ ጽእለተ ለጎርነ

ሣሕቀ ወስላቀ ለአድያሚነ

እስከ ማእዜኑ እግዚኦ ትትመዓዕ ለዝሉፉ

ወይነድድ ከመ እሳት ቅንአትከ

ከዐው መዐተከ ላዕለ አሕዛብ እለ ኢያአምሩከ

ወላዕለ መንግሥት እንተ ኢጸውዐት ስመከ

እስመ በልዕዎ ለያዕቆብ ወአማሰኑ ብሔሮ

ኢትዝክር ለነ አበሳነ ዘትካት

ፍጡነ ይርከበነ ሣህልከ እግዚኦ

እስመ ተመንደብነ ፈድፋደ

ርድአነ አምላክነ ወመድኀኒነ በእንተ ስብሐተ ስምከ

እግዚኦ ባልሐነ ወስረይ ኀጢአተነ በእንተ ስምከ

ከመ ኢይበሉነ አሕዛብ

አይቴ ውእቱ አምላኮሙ

ወይርአዩ አሕዛብ በቅድመ አዕይንቲነ

በቀለ ደሞሙ ለአግብርቲከ ዘተክዕወ

ይባእ ቅድሜከ ገዐሮሙ ለሙቁሓን

ወበከመ ዕበየ መዝራዕትከ

ተሣህሎሙ ለደቂቀ ቅቱላን

ፍድዮሙ ለጎርነ ምስብዒተ ውስተ ሕዕኖሙ

ትዕይርቶሙ ዘተዐየሩከ እግዚኦ

ወንሕነሰ ሕዝብከ ወአባግዐ

መርዔትከ ንገኒ ለከ ለዓለም

ወንነግር ስብሐቲከ ለትውልደ ትውልድ

ፍጻሜ ዘበእንተ እለ ተበዐዱ ስምዕ ዘአሳፍ

ንፍኑ ቀርነ በዕለተ ሠርቅ

በእምርት ዕለት በዓልነ

እስመ ሥርዐቱ ለእስራኤል ውእቱ

ወፍትሐኩ ለአምላከ ያዕቆብ

ወአቀመ ስምዐ ለዮሴፍ አመ የሐውር ብሔረ ግብጽ

ወሰምዐ ልሳነ ዘኢ.ያአምር

ወሜጠ ዘባኖ እምሕራማቲሆሙ

ወተቀንያ እደዊሁ ውስተ አክፋር

ወምንዳቤክ ጸዋዕከኒ ወአድኀንኩከ

ወተሰጠውኩከ በዐውሎ ኀቡእ

ወአመከርኩከ በኀበ ማየ ቅሥት

ስምዐኒ ሕዝብየ ወእንግርከ

እስራኤል ወኣስምዕ ለከ

እመሰ ሰማዕከኒ ኢይከውነከ አምላከ ግብት

ወኢትስግድ ለአምላክ ነኪር

ዘአውፃእኩከ እምድረ ግብጽ

እስመ አነ ውእቱ እግዚአብሔር አምላክከ

አርሕብ አፉከ ወእነልኦ ለከ

ወኢሰምዑኒ ሕዝብየ ቃልየ

ወእስራኤልኒ ኢያፅምኡኒ

ወፈነውኩ ሎሙ በከመ ምግባሮሙ

ወሖሩ በሕሊና ልቦመ

ሶበሰ ሰምዑኒ ሕዝብየ ቃልየ

ወእስራኤልኒ ሶበ ሖሩ በፍኖትየ

እምአኀሰርክዎሙ በኵሉ ለጸላእቶሙ

ወእምወደይኩ እዴየ ዲበ እለ ይሣቅይዎሙ

ጸላእቱሰ ለእግዚአብሔር ሐሰውዎ

ወይከውን ጊዜሆሙ እስከ ለዓለም

አቡነ ዘበሰማያት ይትቀደስ ስምከ ትምጻእ መንግሥትከ ወይኩን ፈቃድከ በከመ በሰማይ ከማሁ በምድር

ሲሳየነ ዘለለ ዕለትነ ሀበነ ዮም ኀድግ ለነ አበሳነ ወጌጋየነ ከመ ንህነኒ ንኀድግ ለዘአበሰ ለነ

ኢታብአነ እግዚኦ ውስተ መንሱት አላ አድኅንነ ወባልሓነ እምኩሉ እኩይ እስመ ዘአከ ይእቲ መንግሥት ኃይል ወ ስ ብ ሐ ት ለዓለመ ዓለም

በሰላመ ቅዱስ ገብርኤል መልአክ ኦ እግዝእትየ ማርያም ሰላም ለኪ

ድንግል በኀሊናኪ ወድንግል በሥጋኪ

እመ እግዚአብሔር ጸባዖት ሰላም ለኪ

ቡርክት አንቲ እምአኀስት ወቡሩክ ፍሬ ከርሥኪ

ጸጋ እግዚአብሔር ምስሌኪ ሰአሊ

ወጸልዩ ኀበ ፍቁር ወልድኪ ኢየሱስ ክርስቶስ ከመ ይሥረይ ለነ ኃጣውኢነ

ፈቀደ እግዚእ ያግዕዞ ለአዳም ኀዙነ ወትኩዘ ልብ ወያግብኦ ኀበዘትካት መንበሩ

ሰአሊ ለነ ቅድስት

ሠረቀ በሥጋ እምድንግል ዘእንበለ ዘርዐ ብእሲ ወአድኃነነ

ለሔዋን እንተ አስሓታ ከይሲ ፈትሐ ላዕሌሃ እግዚአብሔር እንዘ ይብል ብዙኀን አበዝኖ ለሕማምኪ ወለጸዕርኪ ሥምረ ልቡ ኀበ ፍቅረ ሰብእ ወአግዓዛ

ሰአሊ ለነ ቅድስት

ኢየሱስ ክርስቶስ ቃል ዘተሰብአ ወኃደረ ላዕሌነ

ወርኢነ ስብሓቲሁ ከመ ስብሐተ አሐዱ ዋሕድ ለአቡሁ ሥምረ ይሣሃለነ

ወይቤሎ እግዚአብሔር ለአብራም ፃእ እምነ ምድርከ ወእምነ ዘመድከ ወእምቤተ አቡከ ውስተ ምድር እንተ አነ ኣርእየከ

ወእፈስየከ ሕዝበ ዐቢየ ወእባርከከ ወአዐቢ ስመከ ወትከውን ቡሩከ

ወእባርኮሙ ለእለ ይባርኩከ ወእረግሞሙ ለእለ ይረግሙከ ወይትባረክ ኵሉ አሕዛበ ምድር በእንቲአከ

ወሖረ አብራም በከመ ይቤሎ እግዚአብሔር ወሖረ ሎጥኒ ምስሌሁ ወአመ ወፅአ አብራም እምነ ካራን ፸ወ፭ክረምቱ

ወነሥኣ አብራም ለሶራ ብእሲቱ ወሎጥሃ ወልደ እኁሁ ወኵሎ ንዋዮሙ ዘአጥረዩ በካራን ወወፅኡ ወሖሩ ምድረ ከናአን

ወያዳ አብራም ለይእቲ ምድር እስከ ሲኬም ኀበ ዕፅ ነዋኅ ወሰብአ ከናአንሰ ሀለው ይእቲ አሚረ ውስተ ይእቲ ምድር

ወአስተርአዮ እግዚአብሔር ለአብራም ወይቤሎ ለዘርእከ እሁባ ለዛቲ ምድር ወነደቀ አብራም በህየ መሥዋዕት ለእግዚአብሔር ዘአስተርአዮ

ወግዕዘ እምህየ ውስተ ምድረ ቤቴል ዘመንገለ ሠረቅ ወተከለ ህየ ዐጸደ ውስተ ቤቴል አንጻረ ባሕር ዘመንገለ ሠረቅ ወኀደረ ህየ ወነደቀ በህየ ምሥዋዐ ለእግዚአብሔር ወጸውዐ ስሞ

ወተንሥአ ወሖረ ወግዕዘ አብራም ውስተ ገዳም ከመ ይኅድር ህየ

እስመ ጸንዐ ረኀብ ውስተ ብሔር ወወረደ አብራም ውስተ ግብጽ ከመ ይኅድር ህየ እስመ ጸንዐ ረኃብ ውስተ ብሔር

ወኮነ ሶበ ቀርበ አብራም ከመ ይባእ ውስተ ግብጽ ይቤላ አብራም ለሶራ ብእሲቱ ኣአምር ከመ ብእሲት ለሓየ ገጽ አንቲ

ወእምከመ ርእዩኪ ሰብአ ግብጽ ይብሉ ብእሲቱ ይእቲ ወይቀትሉኒ ወኪያኪስ ያሐይውኪ

ወበሊ እንክ እኅቱ አነ ከመ ያሠንዩ ሊተ በእንቲአኪ ወትሕዮ ነፍስየ በዕብሬትኪ

ወኮነ ሶበ በጽሐ አብራም ውስተ ግብጽ ወርእይዋ ለብእሲቱ ሰብአ ግብጽ ከመ ሠናይት ጥቀ

ወርእይዋ መላእክተ ፈርዖን ወወሰድዋ ኀበ ፈርዖን ወአብጽሕዋ ቤቶ

ወአሠነዩ ለአብራም በእንቲአሃ ወአጥረየ አባግዐ ወአልህምተ ወአእዱገ ወአግማለ ወአግብርተ ወአእማተ ወአብቅለ

ወሣቀዮ እግዚአብሔር ለፈርዖን ዐቢየ ሥቃየ ወእኩየ ወለቤቱሂ በእንተ ሶራ ብእሲቱ ለአብራም

ወጸውዖ ፈርዖን ለአብራም ወይቤሎ ምንትኑዝ ዘገበርከኒ ከመ ብእሲትከ ይእቲ

ለምንት ትቤለኒ እኅትየ ይእቲ ወነሣእክዋ ትኩነኒ ብእሲተ ወይእዜኒ ነያ ቅድሜከ ንሥኣ ወሑር

ወአዘዘ ፈርዖን ይፈንውዎ ዕደው ለአብራም ወለብእሲቱ ወለኵሉ ንዋዮሙ ወለሎጥ ምስሌሁ ውስተ አሕቀል

ምዕራፍ

ወዐርገ አብራም እምግብጽ ውእቱ ወብእሲቱ ወኵሉ ንዋዩ ወሎጥሂ ምስሌሁ ውስተ አዜብ

ወአብራምሰ ብዙዕ ጥቀ ወባዕል ፈድፋደ እምእንስሳ ወእምወርቅ ወእምብሩር

ወገብአ እምኀበ ወዕአ ውስተ ሐቅል ውስተ ቤቴል ውስተ መካን ኀበ ሀሎ ቀዲሙ ዐጸዱ ማእከለ ቤቴል ወማእከለ ሐጌ

ውስተ መካን ኀበ ገብረ ምሥዋዐ ህየ ቀዲሙ ወጸውዐ አብራም ስመ እግዚአብሔር በህየ

ወሎጥኒ ዘሖረ ምስሌሁ ለአብራም አጥረየ አባግዐ ወአልህምተ ወእንስሳ

ወኢአከለሙ ምድር ከመ ይኅድሩ ኀቡረ

ወኮነ ጋእዝ ማእከለ ኖልት ዘሎጥ ወዘአብራም ወህለው ይእተ አሚረ ሰብአ ከናአን ወፈርዜዎን ኀዱራን ውስተ ይእቲ ምድር

ወይቤሎ አብራም ለሎጥ ኢይኩን ጋእዝ ማእከሌከ ወማእከሌየ ወማእከለ ኖሎትከ ወማእከለ ኖሎትየ እስመ አኀው ንሕነ

ወናሁ ኵላ ምድር ቅድሜከ ይእቲ ተሌለይ እምኔየ እማእኮ የማነ አንተ ወአነ ፀጋመ ወእማእከ አንተ ፀጋመ ወአነ የማነ

ወአልዐለ ሎጥ አዕይንቲሁ ወርእየ ኵሎ አሕቃላቲሁ ለዮርዳንስ ርውይ ውእቱ ኵሉ ምድር ዘእንበለ ይገፍትዖን እግዚአብሔር ለሶዶም ወለኀሞራ ከመ ገነተ እግዚአብሔር ውእቱ ወከመ ምድረ ግብጽ

ወኀርየ ሎቱ ሎጥ ኵሎ አሕቃላተ ዮርዳንስ ወግዕዘ ሎጥ እመንገለ ሠርቅ ወተሌለዩ አሐዱ ምስለ ካልኡ

አብራም ኀደረ ምድረ ከናአን ወሎጥ ኀደረ ውስተ አድያም ወኀደረ ውስተ ሶዶም

ወሰብአ ሶዶምሰ እኩያን ወኃጥኣን ጥቀ በቅድመ እግዚአብሔር

ወይቤሎ እግዚአብሔር ለአብራም እምድኅረ ተሌለየ እምኔሁ ሎጥ ነጽር በአዕይንቲከ ወርኢ እምነ ዝንቱ መካን ኀበ ሀለውከ ለመንገለ መስዕ ወአዜብ ወሠርቅ ወባሕር

እስመ ኵለንታሃ ለዛቲ ምድር እንተ ትሬኢ ለከ እሁባ ወለዘርእከ እስከ ለዓለም

ወእሬስዮ ለዘርእከ ከመ ኖጻ ባሕር እመቦ ዘይክል ኆልቆ ለኖጻ ባሕር ይኄልቆ ለዘርእከሂ

ዕርግ ወዑዳ ለይእቲ ምድር ውስተ ኑኀ ወርሕባ እስመ ለከ እሁባ

ወግዕዘ አብራም ኀበ ዕፅ እንተ ውስተ ኬብሮን ወነደቀ በህየ መሥዋዕተ ለእግዚአብሔር

Amharic sentences

አቤቱ አሕዛብ ወደ ርስትህ ገቡ

ቤተ መቅደስህንም አረከሱ÷

ኢየሩሳሌምንም እንደ ተክል ጠባቂ ጎጆ አደረጉአት፡፡

የባሪያዎችህንም በድኖች ለሰማይ ወፎች መብል አደረጉ÷

የጻድቃንህንም ሥጋ ለምድር አራዊት

ደማቸውንም በኢየሩሳሌም ዙሪያ እንደ ውኃ አፈሰሱ÷

የሚቀብራቸውም አጡ፡፡

ለጎረቤቶቻችንም ስድብ ሆንን÷

በዙሪያችንም ሳለ ሳቅና መዘበቻ፡፡

አቤቱ እስከ መቼ ለዘለዓለም ትቈጣለህ?

ቅንዐትህም እንደ እሳት ይነድዳል?

በማያውቁህም አሕዛብ ላይ

ስምህንም በማትጠራ መንግሥት ላይ

መዓትህን አፍስስ ያዕቆብን በልተውታልና÷

የቀደመውን በደላችንን አታስብብን÷

አቤቱ ምሕረትህ በቶሎ ያግኘን÷

እጅግ ተቸግረናልና፡፡

አምላካችንና መድኀኒታችን ሆይ÷ ርዳን÷

ስለ ስምህ ክብር አቤቱ÷ ታደገን÷

ስለ ስምህም ኀጢአታችንን አስተሥርይልን፡፡

አሕዛብ÷ “አምላካቸው ወዴት ነው?” እንዳይሉን÷

የፈሰሰውን የባሪያዎችህን ደም በቀል

በዐይኖቻችን ፊት አሕዛብ ይዩ፡፡

የእስረኞች ጩኸት ወደ ፊትህ ይግባ

እንደ ክንድህም ታላቅነት የተገደሉትን

ሰዎች ልጆች ጠብቃቸው።።

አቤቱ÷ የተገዳደሩህን መገዳደራቸውን÷

ለጎረቤቶቻችን ሰባት እጥፍ በብብታቸው ክፈላቸው።።

እኛ ሕዝብህ ግን÷ የማሠማሪያህም በጎች÷

ለዘለዓለም እናመሰግንሃለን

ለልጅ ልጅም ምስጋናህን እንናገራለን።።

በረዳታችን በእግዚአብሔር ደስ ይበላችሁ÷

ለያዕቆብም አምላክ እልል በሉ።።

ዝማሬውን አንሡ ከበሮውንም ስጡ÷

ደስ የሚያሰኘውን በገና ከመሰንቆ ጋር

በዓላችን ቀን መለከትን ንፉ

በመባቻ ቀን በታወቀችው

ለእስራኤል ሥርዐቱ ነውና÷

የያዕቆብም አምላክ ፍርድ ነውና።።

ወደ ግብፅ ሀገር በሔደ ጊዜ ለዮሴፍ ምስክርን አቆመ።።

የማያውቀውን ቋንቋ ሰማ።።

ጀርባውን ከመስገጃው መለሰ÷

እጆቹም በቅርጫት ተገዙ።።

በመከራህ ጊዜ ጠራኸኝ÷ አዳንህህም÷

በተሰወረ ዐውሎም መለስሁልህ÷

በክርክር ውኃ ዘንድም ፈተንሁህ።።

ሕዝቤ ሆይ÷ ስማኝ እነግርሃለሁም

እስራኤል ሆይ÷ እመሰክርልሃለሁ፡፡

አንተስ ብትሰማኝ የድንገት አምላክ አልሆንህም÷

ለሌላ አምላክም አትስገድ፡፡

ከግብፅ ምድር ያወጣሁህ

እኔ እግዚአብሔር አምላክህ ነኝና

አፍህን አስፋ÷ እሞላዋለሁም፡፡

ሕዝቤ ግን ቃሌን አልሰሙኝም÷

እስራኤልም አላደመጡኝም፡፡

እንደ ሥራቸው ላክሁባቸው÷

በልባቸው አሳብ ሔዱ፡፡

ሕዝቤስ ቃሌን ሰምተውኝ ቢሆን÷

እስራኤልም በመንገዴ ሔደው ቢሆን

ጠላቶቻቸውን ፈጥኜ ባዋረድኋቸው ነበር÷

በሚያስጨንቋቸውም ላይ እጄን በጣልሁ ነበር÷

የእግዚአብሔር ጠላቶችስ ዋሽተውት ነበር

ዘመናቸውም ለዘለዓለም ይሆናል

አባታችን ሆይ በሰማያት የምትኖር ስምህ ይቀደስ መንግሥትህ ትምጣ ፈቃድህ በሰማይ እንደሆነች እንዲሁም በምድር ትሁን

የዕለት እንጀራችንን ስጠን ዛሬ በደላችንንም ይቅር በለን እኛ የበደሉንን ይቅር እንደምንል

ወደፈተናም አታግባን ከክፋ ሁሉ አድነን እንጂ መንግሥት ያንተ ናትና ኃይል ምስጋና ለዘላዓለሙ አሜን

እመቤቴ ማርያም ሆይ በመልአኩ በቅዱስ ገብርኤል ሰላምታ ሰላም እልሻለሁ

በሃሳብሽ ድንግል ነሽ በሥጋሽም ድንግል ነሽ

የአሸናፊ የእግዚአብሔር እናት ሆይ ላንቺ ሰላምታ ይገባል

ከሴቶች ሁሉ ተለይተሽ አንቺ የተባረክሽ ነሽና የማኅፀንሽም ፍሬ የተባረከ ነው

ጸጋን የተመላሽ ሆይ ደስ ይበልሽ እግዚአብሔር ካንች ጋር ነውና

ከተወደደው ልጅሽ ከጌታችን ከኢየሱስ ክርስቶስ ዘንድ ይቅርታን ለምኝልን ኃጢአታችንን ያስተሠርይልን ዘንድ አሜን

ጌታ ልቡ ያዘነና የተከዘ አዳምን ነፃ ያወጣውና ወደቀድሞ ቦታው ይመልሰው ዘንድ ወደደ

ቅድስት ሆይ ለምኝልን

ከድንግል ያለወንድ ዘር በሥጋ ተወለደና አዳነን

ከይሲ ያሳታት ሔዋንን እግዚአብሔር ምጥሽንና ጻርሽን አበዛዋለሁ ብሎ ፈረደባት ሰውን ወደደና ነፃ አደረጋት

ቅድስት ሆይ ለምኝልን

ሰው የሆነና በኛ ያደረ ቃል ኢየሱስ ክርስቶስ ነው

ክብሩንም ለአባቱ አንድ እንደመሆኑ ክብር አየን ይቅር ይለን ዘንድ ወደደ

እግዚአብሔርም አብራምን አለው ከአገርህ ከዘመዶችህም ከአባትህም ቤት ተለይተህ እኔ ወደማሳይህ ምድር ውጣ

ታላቅ ሕዝብም አደርግሃለሁ እባርክሃለሁ ስምህንም አከብረዋለሁ ለበረከትም ሁን

የሚባርኩህንም እባርካለሁ የሚረግሙህንም እረግማለሁ የምድር ነገዶችም ሁሉ በአንተ ይባረካሉ

አብራምም እግዚአብሔር እንደ ነገረው ሄደ ሎጥም ከእርሱ ጋር ሄደ አብራምም ከካራን በወጣ ጊዜ የሰባ አምስት ዓመት ሰው ነበረ

አብራምም ሚስቱን ሦራንና የወንድሙን ልጅ ሎጥን ያገኙትን ከብት ሁሉና በካራን ያገኙአቸውን ሰዎች ይዞ ወደ ከነዓን ምድር ለመሄድ ወጣ ወደ ከነዓንም ምድር ገቡ

አብራምም እስከ ሴኬም ስፍራ እስከ ሞሬ የአድባር ዛፍ ድረስ በምድር አለፈ የከነዓን ሰዎችም በዚያን ጊዜ በምድሩ ነበሩ

እግዚአብሔርም ለአብራም ተገለጠለትና ይህችን ምድር ለዘርህ እሰጣለሁ አለው እርሱም ለተገለጠለት ለእግዚአብሔር በዚያ ስፍራ መሠውያን ሠራ

ከዚያም በቤቴል ምሥራቅ ወዳለው ተራራ ወጣ በዚያም ቤቴልን ወደ ምዕራብ ጋይን ወደ ምሥራቅ አድርጎ ድንኳኑን ተከለ በዚያም ለእግዚአብሔር መሠውያን ሠራ የእግዚአብሔርንም ስም ጠራ

አብራምም ከዚያ ተነሣ እየተጓዘም ወደ አዜብ ሄደ

በምድርም ራብ ሆነ አብራምም በዚያ በእንግድነት ይቀመጥ ዘንድ ወደ ግብፅ ወረደ በምድር ራብ ጸንቶ ነበርና

ወደ ግብፅም ለመግባት በቀረበ ጊዜ ሚስቱን ሦራን እንዲህ አላት አንቺ መልከ መልካም ሴት እንደ ሆንሽ እነሆ እኔ አውቃለሁ

የግብፅ ሰዎች ያዩሽ እንደ ሆነ ሚስቱ ናት ይላሉ እኔንም ይገድሉኛል አንቺንም በሕይወት ይተዉሻል

እንግዲህ በአንቺ ምክንያት መልካም ይሆንልኝ ዘንድ ስለ አንቺም ነፍሴ ትድን ዘንድ እኅቱ ነኝ በዪ

አብራምም ወደ ግብፅ በገባ ጊዜ የግብፅ ሰዎች ሴቲቱን እጅግ ውብ እንደ ሆነች አዩ

የፈርዖንም አለቆች አዩአት በፈርዖንም ፊት አመሰገኑአት ሴቲቱንም ወደ ፈርዖን ቤት ወሰዱአት

ለአብራምም ስለ እርስዋ መልካም አደረገለት ለእርሱ በጎችም በፍየችም አህዮችም ወንዶችና ሴቶች ባሪያዎችም ግመሎችም ነበሩት

እግዚአብሔርም በአብራም ሚስት በሦራ ምክንያት ፈርዖንንና የቤቱን ሰዎች በታላቅ መቅሠፍት መታ

ፈርዖንም አብራምን ጠርቶ አለው ይህ ያደረግህብኝ ምንድር ነው? እርስዋ ሚስትህ እንደ ሆነች ለምን አልገለጥህልኝም?

ለምንስ እኅቴ ናት አልህ? እኔ ሚስት ላደርጋት ወስጄአት ነበር አሁንም ሚስትህ እነሆት ይዘሃት ሂድ

ፈርዖንም ሰዎቹን ስለ እርሱ አዘዘ እርሱንም ሚስቱንም ከብቱንም ሁሉ ሸኙአቸው

ምዕራፍ

አብራምም ከግብፅ ወጣ እርሱና ሚስቱ ለእርሱ የነበረውም ሁሉ ሎጥም ከእርሱ ጋር ወደ አዜብ ወጡ

አብራምም በከብት በብርና በወርቅ እጅግ በለጠገ

ከአዜብ ባደረገው በጒዞውም ወደ ቤቴል በኩል ሄደ ያም ስፍራ አስቀድሞ በቤቴልና በጋይ መካከል ድንኳን ተክሎበት የነበረው ነው

ያም ስፍራ አስቀድሞ መሠውያ የሠራበት ነው በዚያም አብራም የእግዚአብሔርን ስም ጠራ

ከአብራም ጋር የሄደው ሎጥ ደግሞ የላምና የበግ መንጋ ድንኳንም ነበረው

በአንድነትም ይቀመጡ ዘንድ ምድር አልበቃቸውም የነበራቸው እጅግ ነበረና በአንድነት ሊቀመጡ አልቻሉም

የአብራምንና የሎጥን መንጎች በሚጠብቁት መካከልም ጠብ ሆነ በዚያም ዘመን ከነዓናውያንና ፌርዛውያን በዚያች ምድር ተቀምጠው ነበር

አብራምም ሎጥን አለው እኛ ወንድማማች ነንና በእኔና በአንተ በእረኞቼና በእረኞችህ መካከል ጠብ እንዳይሆን እለምንሃለሁ

ምድር ሁሉ በፊትህ አይደለችምን? ከእኔ ትለይ ዘንድ እለምንሃለሁ አንተ ግራውን ብትወስድ እኔ ወደ ቀኝ እሄዳለሁ አንተም ቀኙን ብትወስድ እኔ ወደ ግራ እሄዳለሁ

ሎጥም ዓይኑን አነሣ በዮርዳኖስ ዙሪያ ያለውንም አገር ሁሉ ውኃ የሞላበት መሆኑን አየ እግዚአብሔር ሰዶምንና ገሞራን ከማጥፋቱ አስቀድሞ እስከ ዞዓር ድረስ እንደ እግዚአብሔር ገነት በግብፅ ምድር አምሳል ነበረ

ሎጥም በዮርዳኖስ ዙሪያ ያለውን አገር ሁሉ መረጠ ሎጥም ወደ ምሥራቅ ተጓዘ አንዱም ከሌላው እርስ በርሳቸው ተለያዩ

አብራም በከነዓን ምድር ተቀመጠ ሎጥም በአገሩ ሜዳ ባሉት ከተሞች ተቀመጠ እስከ ሰዶምም ድረስ ድንኳኑን አዘዋወረ

የሰዶም ሰዎች ግን ክፉዎችና በእግዚአብሔር ፊት እጅግ ኃጢአተኞች ነበሩ

ሎጥ ከተለየው በኋላም እግዚአብሔር አብራምን አለው ዓይንህን አንሣና አንተ ካለህበት ስፍራ ወደ ሰሜንና ወደ ደቡብ ወደ ምሥራቅና ወደ ምዕራብ እይ

የምታያትን ምድር ሁሉ ለአንተና ለዘርህ ለዘላለም እሰጣለሁና

ዘርህንም እንደ ምድር አሸዋ አደርጋለሁ የምድርን አሸዋን ይቈጥር ዘንድ የሚችል ሰው ቢኖር ዘርህ ደግሞ ይቈጠራል

ተነሣ በምድር በርዝመትዋም በስፋትዋም ሂድ እርስዋን ለአንተ እሰጣለሁና

አብራምም ድንኳኑን ነቀለ መጥቶም በኬብሮን ባለው በመምሬ የአድባር ዛፍ ተቀመጠ በዚያም ለእግዚአብሔር መሠውያን ሠራ