Addis Ababa University

College of Natural Sciences

Geez to Amharic Machine Translation

Biruk Abel

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

Addis Ababa, Ethiopia

May, 2018

Addis Ababa University

College of Natural Sciences

Biruk Abel

Advisor: Mulugeta Libise (PhD)

This is to certifiy that the thesis prepared by Biruk Abel, titled: *Geez to Amharic Machine Transaltion* and submitted in partial fulfillment of the requirments for the Degree of Master of Science in Computer Science complies with the regualtions of the Univesity and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signiture	Date
Advisor:			
Examiner:			
Examiner:			

Abstract

Natural Language Processing (NLP) is defined as ways for computers to analyze, understand, and derive meaning from human language in a smart and useful way. Machine Translation (MT) is one of the applications of NLP. It is the use of computers to translate from one natural language like Geez to another say Amharic. Natural languages may follow different word ordering during sentence formation for example Geez follows Subject + verb + object (SVO) and Verb + subject + object (VSO) while Amharic only follows SOV so alignment of the right Geez word with the Amharic word is of paramount importance to improve the translation quality. The purpose of this study to develop a Hybrid Geez to Amharic Machine Translation system using serial coupling of rule based Geez language word reordering followed by a standard Statistical Machine Translation (SMT) system.

The proposed system is composed of two main components a Rule Based Geez Corpus Preprocessor and a Baseline SMT. The Rule Based Preprocessor takes the manually Part of Speech (POS) tagged Geez corpus and produces another corpus that contains reordered Geez sentences having similar structure with that of Amharic sentences. This component contains set of activities that process each Geez sentence in the input corpus one by one to determine POS pattern and subsequently apply the corresponding reordering rule. It first reads all sentences from the input file and iterates through all sentences and it first determines POS pattern and applies the corresponding reordering rule. After each sentence is processed the output corpus along with the Amharic corpus will be supplied as an input to the Baseline SMT. Then using the input corpora the actual translation of Geez sentence to Amharic sentences will be performed by the Decoder of the Baseline SMT by using the Language model of Amharic and Translation model.

The translation quality of the proposed system is evaluated using BLEU evaluation metrics and compared with that of the Baseline SMT. Two experiments were conducted one to test the Baseline SMT and the other to test the proposed system. To test the Baseline SMT both Geez and Amharic corpus without POS were used while to test the proposed system Geez corpus with POS and Amharic corpus with no POS were used. Based on the test results the Baseline SMT scored a BLEU of 72% and the proposed system outscores it by 4% and scored 76% owing to the reordering rules applied on Geez corpus.

Keywords: Geez to Amharic Machine Translation, Hybrid Machine Translation, Rule Based Word Reordering, Statistical machine Translation, Part of Speech Tagging.

Acknowledgement

First and foremost I would like to thank God and St. Mary for giving me the endurance and wisdom to accomplish this study.

Second my deep gratitude goes to Dr. Mulugeta Libsie, for his constructive support and guidance during the course of the thesis starting from proposal till completion. Meticulousness, dedication and punctuality are the most important life time lessons that I learnt from you.

Thirdly, I would like to thank Metasebia Demessie (MTH) for reviewing the Geez and Amharic corpora, part of speech tagging, reordering rules, and editing translation input/output/reference for correctness, which helped me in evaluating the translation quality of the proposed system.

Table of Contents

List of Figures

List of Tables

List of Acronyms

API	Application Programming Interface
BLEU	Bilingual Evaluation Understudy
CAT	Computer Assisted Tools
EBMT	Example Based Machine Translation
HMT	Hybrid Machine Translation
MT	Machine Translation
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NMG	Normalized Mean Grams
POS	Part of Speech
RBMT	Rule Based Machine Translation
SMT	Statistical Machine Translation
SPE	Statistical Post Editing
SVO	Subject Verb Object
VSO	Verb Subject Object
WER	Word Error Rate
XDG	Extensible Dependency Grammar

Chapter One: Introduction

Machine Translation is the automatic translation of one natural language to the other using computers. Interest in Machine Translation is nearly as old as the electronic computer. It is a key application in the field on Natural Language Processing (NLP) [1]. In general, paradigms to Machine Translation (MT) could be either rule based or corpus based [2]. Statistical Machine Translation (SMT) is a corpus based approach to MT using Machine Learning methods. According to [1], SMT treats translation as a Machine Learning problem. This means that we apply a learning algorithm to a large body of previously translated text, known variously as parallel corpus, parallel text, bi-text, or multi-text. The learner is then able to translate previously unseen sentences.

Ethiopia is the only countriy in Africa having its own script (*Fidel*) and numeral system (*Ahaze*). Different manuscripts justify that Ethiopia is a nation having a lot of wisdom regarding literatures of varying kind [3]. Geez is a family of Semitic languages [4]. The ancient philosophy, tradition, history and knowledge of Ethiopia were being written in Geez and also there are different books which are written in this language. However, Geez is not known by the current generation. It has almost ceased to be a spoken language of Ethiopia, i.e., it is only used and known by scholars/*liqawent* of Ethiopian Orthodox Church. Moreover, it is being used as a language of literature and of liturgy in the religion. The literature includes religious texts (such as the Bible, Apocrypha, Pseudepigrapha, liturgical literature, homiletic, theological, and magical texts, stories of martyrs and saints, religious poetry, hymns in honor of Christ, the Virgin, the martyrs, the saints, and angels), as well as secular writings (histories and romances, legal, mathematical, and medical texts) [4].

Amharic is a family of Semitic languages. It uses a unique script called ‘fidel’ which is conveniently written in a tabular format of seven columns. The first column represents the basic form and the other orders are derived from it by more or less regular modifications indicating the different vowels. Amharic has 34 base characters and this leads to have a total of 238 (=34*7) Amharic characters. In addition, there are about two scores of characters representing labialized sounds [5].

Concerned bodies that are working on languages of Ethiopia and society of the nation as a whole have the responsibility of passing on the treasure found in the aforementioned manuscripts to the forthcoming generation. This proposal is aimed at proposing an automatic machine translation system that translates Geez text to Amharic text.

1.1 Motivation

The majority of manuscripts in Ethiopia that are from ancient times are written in this language [6]. Great effort has been made by our ancestors in order to pass Ethiopian culture, knowledge and wisdom to the generation of our time. As to our fathers and mothers this generation must contribute something in order to pass the legacy to the coming generation plus, stop the language becoming extinct. One solution would be developing an automatic system that translates Geez text to Amharic. This is the main idea behind this proposal.

1.2 Statement of the Problem

Ethiopia is a rich country in literature of various kinds. For the past thousands of years our ancestors have contributed a lot in maintaining and building the country's tradition and passing it onto the next generation. This is one of the reasons that Ethiopia is a pride not only to its citizens but also for our continent Africa [3]. In the Ethiopian Orthodox church all liturgical services are predominantly conducted by geez language. The liturgy of the 14 Anaphoras is celebrated in this language. Moreover, there are lots of scripts written in the language and are being used in Ethiopian Orthodox church.

These religious texts include but are not limited to, 'Gedle' (different books on the history of different saints); 'Dirsan' (different books on the history and works of different Holy Angeles of God); 'Sinsikar' (the lives of different saints including their birth days, afflictions, miracles, death days, etc. classified in the 365 days of the year. It is common to read the story of three or five or more saints life story in any day of the year.); 'Teamire Mariam' (the book of the miracles of the Blessed St. Virgin Mary); 'Teamire Eyesus' (the book of the miracles of our Lord Jesus Christ); 'Haymanot Abew' (Book of the Dogmatic teachings of the fathers); ' Metsehafe kdasse ' (books of the 14 Anaphoras); 'Fitiha Negest' (A book which states the former constitution of the state and the early and current constitution of the church); 'Zena Hawaryat' (the story and the works of

the Apostles); 'Negere Mariyam' (the story of the Blessed Virgin St. Mary); 'Kibre Negest' (Glory of the Kings); and numerous books of prayers [6].

The distinctive attainment of Ethiopian history lies in the vast collection of manuscripts, compiled and preserved in the monasteries and churches. Almost all these scriptures and other religious works of the religion are done by the Geez language [6]. Therefore, the Geez language means more to the Ethiopian Orthodox Tewahido Church and Ethiopia. Overall, there are numerous church scriptures in the church which were written in Geez and not yet translated into the Amharic Language. Hence, an automatic translation system that translates Geez to a language being spoken at national level like Amharic is of paramount importance.

1.3 Objectives

General Objective

The general objective of this thesis is to design and implement machine translation system that automatically translates Geez sentences to corresponding Amharic sentences using hybrid machine translation techniques.

Specific Objectives

- To review related works in machine translation for different languages,
- To prepare parallel corpus for the translation,
- To prepare Part of Speech (POS) tag set for Geez language
- To prepare reordering rule for Geez sentences
- To formulate algorithm for performing Geez sentences reordering from POS
- To develop a Rule Base Geez Corpus Preprocessor component
- To develop a Baseline Statistical Machine Translation component
- Evaluate the translation quality of the hybrid machine translation system against the baseline SMT.

1.4 Methods

Literature Review

For the purpose of finding up-to-date methodologies in machine translation domain, thorough literature review will be conducted. Any peer reviewed publications including books, articles, journals and other scholarly publications will be reviewed.

Data Collection

Two sets of corpora will be prepared for this study First, corpora with no POS information and it will contain two text file one containing Geez sentences the other one is an Amharic corpus that contains Amharic sentences/phrases which are the translation of the corresponding Geez sentences in the Geez corpus. The second corpora contains Geez corpus with POS information and corresponding Amharic corpus with no POS. Regardless of the POS tags both corpora will contain identical sentences and will entirely be collected from [7].

Tools

In order to achieve the objectives of the study a number of tools are needed. Tools that are needed to develop the Rule based Geez Corpus Preprocessor and those for the Baseline SMT development. The Rule Based Geez Corpus Preprocessor component will be a Windows based application that will be developed using Microsoft Visual Studio 2013, Dot Net Frame Work Version 4.0, and C# Programming language. Tools that will be used while developing the Baseline SMT will be VMware 10.0, Ubuntu Linux 14.04, Moses, IRSTLM, and MGIZA.

Testing

To evaluate the translation quality of the proposed hybrid Geez to Amharic translation system, two experiments will be conducted. The first to test the translation quality of the hybrid system and the second to test the Baseline SMT system and BLEU (Bilingual Evaluation Understudy) evaluation metrics will be used.

1.5 Scope and Limitations

The machine translation system that is going to be developed will be used to translate Geez phrase to corresponding Amharic phrase not the other way round.

1.6 Application of Results

The research work will have two application areas. First it will bring new knowledge into the domain of machine translation. Second, it will be used as a teaching platform for learning both Geez and Amharic languages.

1.7 Thesis Organization

The rest of the thesis is organized as follows. Literatures on Machine Translation and Geez language constructs like part of speech, word ordering rules are presented in Chapter 2. Chapter 3 reviews related Machine Translation works on various languages specially those that use Hybrid architecture. A detailed description of the proposed Rule Based Geez Corpus Preprocessor and Baseline SMT component will be presentenced in Chapter 4. Chapter 5 presents the experimentation result along with the environment used to conduct the experiment. Moreover, BLEU results are also discussed. Finally, Chapter 6 concludes the thesis with the research findings, conclusions, and future works.

Chapter Two: Literature Review

This Chapter discusses about Natural Language Processing (NLP), approaches to NLP and its applications like question answering, machine translation, text summarization and speech recognition. It further explains machine translation, history of machine translation and available approaches to machine translation. Finally, features of Geez and Amharic languages are discussed.

2.1 Natural Language Processing

According to [8] language is one of the key parts of human conduct and is a significant segment of our lives. In composed frame it serves as a long haul record of learning starting with one era then onto the next. In talked frame it serves as our essential method for organizing our everyday conduct with others. The term natural language refers to human languages like English, Geez, Amharic, etc. Language is studied in several different academic disciplines. Each discipline defines its own set of problems and has its own methods for addressing them. The linguist, for instance, studies the structure of the language itself, considering questions such as why certain combinations of words form sentences but others do not, and why a sentence can have a few implications yet not others.

The psycholinguist, then again, concentrates on the procedures of human language creation and understanding, considering inquiries, for example, how individuals recognize the suitable structure of a sentence and when they settle on the fitting significance for words. The philosopher considers how words can mean anything at all and how they identify objects in the world. Philosophers also consider what it means to have beliefs, goals, and intentions, and how these cognitive capabilities relate to language. The goal of the computational linguist is to develop a computational theory of language, using the notions of algorithms and data structures from computer science. In order to develop a computational model, it is a must to take the advantages of what is known from all the other disciplines [8].

NLP is defined in [9] as ways for computers to analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic

segmentation. Apart from common word processor operations that treat text like a mere sequence of symbols, NLP considers the hierarchical structure of a language: several words makeup a phrase, several phrases makeup a sentence and, ultimately, sentences convey ideas. By analyzing language for its meaning, NLP systems have long filled useful roles, such as grammar correction, converting speech to text and the automatic translation between languages.

NLP is the field of study that focuses on the interactions between human language and computers. It sits at the intersection of computer science, artificial intelligence, and computational linguistics. NLP is used to analyze text, allowing machines to understand how humans speak. This human-computer interaction enables real-world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more. NLP is commonly used for text mining, machine translation, and automated question answering [8].

Natural language is the preferred medium of communication between people. Scientific articles, magazines and billions of web pages are also written in natural languages. On the other hand, computers can do a lot of useful things for us like storing data in structured form, for example, databases and knowledge bases [9]. Plus, they are used to specify tasks in a formal way using programming languages. NLP bridges the gap between people and computers and this leads to a better and a more natural communication with computers and process an ever increasing amount of natural language data generated by people like extracting information from the web.

2.1.1 Approaches to Natural Language Processing

There are four approaches to NLP: symbolic, statistical, connectionist and hybrid. Symbolic and statistical approaches coexisted starting from the earliest times and later connectionist approaches comes to existence [9].

a. Symbolic

This approach performs deep analysis of the language based on formulated rules by linguists and lexicon. An example of symbolic approach is seen in rule based systems. Rule-based systems usually consist of a set of rules, an inference engine, and a workspace or working memory. Knowledge is represented as facts or rules in the rule-base. The inference engine repeatedly selects a rule whose condition is satisfied and executes the rule.

Semantic networks are another example of symbolic approach. Semantic networks represent knowledge through a set of nodes that represent objects or concepts and the labeled links that represent relations between nodes. The pattern of connectivity reflects semantic organization, that is, highly associated concepts are directly linked whereas moderately or weakly related concepts are linked through intervening concepts. Semantic networks are widely used to represent structured knowledge and have the most connectionist flavor of the symbolic models. Symbolic approaches have been used for a few decades in a variety of research areas and applications such as information extraction, text categorization, ambiguity resolution, and lexical acquisition [10].

b. Statistical Approach

Statistical approaches employ various mathematical techniques and often use large text corpora to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant linguistic or world knowledge. In contrast to symbolic approaches, statistical approaches use observable data as the primary source of evidence. Statistical approaches have typically been used in tasks such as speech recognition, lexical acquisition, parsing, part-of-speech tagging, collocations, statistical machine translation, and statistical grammar learning [10].

c. Connectionist Approach

Similar to the statistical approaches, connectionist approaches also develop generalized models from examples of linguistic phenomena. What separates connectionism from other statistical methods is that connectionist models combine statistical learning with various theories of representation - thus the connectionist representations allow transformation, inference, and manipulation of logic formulae. In addition, in connectionist systems, linguistic models are harder to observe due to the fact that connectionist architectures are less constrained than statistical ones [9].

d. Hybrid Approach

The above three approaches may not be adequate for complex NLP tasks. In some cases it may be difficult to categorize an approach as pure symbolic, connectionist or statistical rather current trends show that combination of these approaches to come up with a more robust hybrid approach to NLP that combines the best of all existing approaches [11].

2.1.2 Applications of Natural Language Processing

The following are applications of NLP.

1. **Question Answering**: a system capable of understanding questions formulated in one natural language such as Geez and responding with exactly the requested information with the same language [10].
2. **Text Summarization**: This area includes applications that can take a collection of documents or emails, and produce a coherent summary of their content. Such tasks also aim to provide brief summaries of longer documents [11].
3. **Machine Translation (MT)**: is the use of computers to translate from one natural language say Geez to another like Amharic. Translations of high quality require a deep and reach understanding of the source language and a sophisticated, poetic, and creative command of the target language. The problem of automatically producing high-quality translation of arbitrary text from one language to another is thus far too hard to automate completely [12]. Since our work is on MT we will discuss MT in more detail in the next section.
4. **Speech Recognition**: it is the automatic recognition of spoken language via computers. This is one of the most difficult problems in NLP. There has been great progress in building models that can be to recognize spoken language utterances that are questions and commands [11].
5. **Document Classification**: This is one of the most successful areas of NLP, wherein the task is to identify which category (or *bin*) a document should be put in. This has proved enormously useful for applications such as spam filtering, news article classification, and movie reviews, among others. One reason this has had such a big impact is the relative simplicity of the learning models needed for training the algorithms that do the classification [1].
6. **Text to Speech**: the application of NLP that automatically converts text given in one natural language and converts the input text to an acoustic signal. Four sub processes are performed in this application. Text analysis, word pronunciation, phonetic interpretation and finally signal generation [13].

2.1.3 Machine Translation

Machine translation is an automatic translation of one language into another by means of a computer or another machine that contains a dictionary, along with the programs needed to make logical choices from synonyms, supply missing words and rearrange word order as required for the new language [14].

a. History of Machine Translation

Interest in automatic machine translation started in the late forties after World War II. MT was constrained by several factors: limitation of hardware particularly, inadequacy of memories and slow access and unavailability of high level programming language. The linguistic study was not correlated with machine translation research, so researchers relied on the dictionary- based approach and the application of statistical methods [1].

Researchers of that time were faced with a lot of technical constraints and realized that there could be no perfect high quality translation, and suggested the involvement of humans in the process. They also proposed the development of controlled languages and restriction of systems to specific domains.

Criteria concerning the success and failure of machine translation were set in its first 50 years of research and development. These criteria are the conceptual, engineering, operational, commercial and communicative criteria [14].

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

An American mathematician and scientist named Warren Weaver in 1947 had a belief that a computer is capable of translating one natural language to another by using logic, cryptography, frequencies of letter combinations, and linguistic patterns. He published a memorandum that

outlines his belief. In the 1950s a research program at Georgetown University teamed up with IBM to perform a research on MT. Later in 1954, they demonstrated a system that translates few phrases from Russian to English. The research resulted in a wide acceptance and interest to the field [15].

In 1970's and 80's researches shifted their focus to assisting machine translation rather than replacing human translators. That resulted in the development of translation memory and many computer assisted tools (CAT) for MT. with the emergence of internet and cheap and powerful computers plus advances in speech recognition software were a few factors that accelerated the progress of MT. Nowadays researches are focused on improving quality of a MT Translation system [15].

b. Approaches to Machine Translation

Paradigms to MT could be rationalism or empiricist [16]. However, both methods have both advantages and limitations. Hybrid techniques exist that combine the benefits of both approaches. The following sub sections will give explanation on paradigms to MT namely Rule Based machine translation, Statistical Machine translation, example based machine translation and Hybrid machine translation.

i. Rule Based Machine Translation

Rule based machine translation (RBMT), which is a rationalism approach, requires analysis and representation of the meaning of source language texts and the generation of equivalent target language texts. Representation should be unambiguous lexically and structurally [14]. There are two major approaches:

1. The transfer approach in which translation process operates in three stage-analyses into abstract source language representations, transfer into abstract target language representations and generation or synthesis into target language text.
2. The two stage Interlingua model where analysis into some language-neutral representation starts from this Interlingua representation.

These two models incorporate batch processing with post-editing and non-interactive components, essentially syntax oriented, with analyses and generation passing through a series of levels (morphological, syntactic, deep syntactic, semantic) and making little use of pragmatic or discourse information. Knowledge-based approach is founded on the assumption that translation must go beyond linguistic knowledge and must involve understanding.

ii. Statistical Machine Translation

Statistical machine translation (SMT), which is an empiricist approach, is a paradigm to MT that is characterized by Machine Learning methods [1]. SMT treats translation as a Machine Learning problem. This means that we apply a learning algorithm to a large body of previously translated text, known variously as a parallel corpus. It allows faster prototyping of MT systems. The general architecture of SMT includes three components: language model, translation model, and decoder. Language model ensures that words come in the right order, i.e., Subject (S)-Object (O) -(Verb) V.

SMT faces the following problems specifically with the translation model [18].

1. **Fertility is asymmetric**: a single word from the target language often matches with two or more words of the source language.
2. **Sensitivity to training data:** minor changes to training data and probability model results in huge changes in the estimates of the parameters.
3. **Efficiency:** sentences containing more than 30 words will took much time during decoding.

Since SMT lacks any linguistic knowledge, the following problems are faced with this paradigm.

1. **No notion of phrases:** it only deals with words, no notion of phrases hence, absence of context information for translation.
2. **Morphology:** morphologically related words are treated separately.
3. **Data sparseness:** since only training corpus is used and no other linguistic knowledge is used, estimation of rare words becomes unreliable.

iii. Example Based Machine Translation

An empiricist approach of machine translation, Example Based Machine Translation (EBMT) retrieves similar examples (pair of source and target sentences) from a database in order to translate new sentences [19]. System configuration of EBMT includes example database, thesaurus and three transfer modules namely analysis, example based transfer and generation.

There are four stages in EBMT, namely, example acquisition, example base management, example application and target sentence synthesis [20]. Example acquisition is about how to acquire examples from parallel bilingual corpus (, i.e., existing translation), and example base management is about how examples are stored and maintained. The example application concerns itself with how examples are used to facilitate translation, which involves the

decomposition of an input sentence into examples and the conversion of source texts into target texts in terms of existing translation. The sentence synthesis 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.

iv. Hybrid Machine Translation

Theoretically, Hybrid Machine Translation (HMT) is an MT paradigm that combines the best of rule based machine translation, statistical machine translation and example based machine translation [20]. The most famous hybrid models for MT are the following.

1. RBMT and EBMT Hybrid Approach

This approach combines the best of RBMT and EBMT paradigms. An algorithm was designed in [21] that combines the RBMT and EBMT for translation of English to Japanese. The algorithm has three major steps first, selecting a set of candidate sentences which are similar to the input sentence; second, selection of the most typical translation out of those corresponding to the candidates, and the third step is using this translation and its source as templates to translate the input sentence. By discarding candidates with a typical translation, the algorithm filters out free, incorrect or context dependent translations.

2. RBMT and SMT Hybrid Approach

This is a hybrid model that combines the pros of RBMT and SMT machine translation paradigms. The output of a RBMT is fed to a statistical post editing component to make domain specific corrections [22]. This work shows a great BLEU score improvement owing to the statistical post editing component. Even when the training data is scarce, the proposed MT outperforms the direct based MT.

3. EBMT and SMT Hybrid Approach

The third hybrid model uses EBMT and SMT. The work in [17] proposed architecture in which multiple EBMT engines work in parallel and their outputs are passed to a post-process statistical selector that selects the best candidate according to SMT models. Then the output of the selection is fed to a new Statistical Machine Translation using Hidden Markov Model (HMM) based statistical machine translation, and finally applying a paraphrasing technique.

2.2 The Geez Language

Ethiopia is a country in Africa having its own characters and numerals. A lot of scriptures are evidence that Ethiopia is a country that is enriched with literatures of various kinds [4]. For the

past thousands of years the effort exerted by our ancestors on building the culture of our country is enormous. They have contributed a lot in passing on their wisdom to the generation of our time. This wisdom is a pride not only to Ethiopia but also to Africa as well [3].

Geez is a family of Afro-Asiatic language having its own characters called hohiyat (ሆህያት). These characters have their own meaning and pictorial representation that have been in use since ancient times. Ethiopia is very rich in literature of various kinds which are written in Geez language. These scripts have contributed a lot to the growth of literature not only to Ethiopia but to Africa as well [3]. One of the instruments that was used to pass the wisdom is the inborn and ancient language of Ethiopia, Geez. This language has embraced a lot like identity, history, religion, etc. of citizens of Ethiopia.

Geez has ceased to be a spoken language of Ethiopia except scholars of the Orthodox Church. It is used as the main language of liturgy and literature of the Church [23]. The literature includes religious texts (such as the Bible, Apocrypha, Pseudepigrapha, liturgical literature, homiletic, theological, and magical texts, stories of martyrs and saints, religious poetry, hymns in honor of Christ, the Virgin, the martyrs, the saints, and angels), as well as secular writings (histories and romances, legal, mathematical, and medical texts).

In addition to its own scripts Geez has its own numerals (Ahaz). The base of these numerals is the literature characters /hohiyat/. For example, in the Geez numerals six ፮ and seven ፯, if we omit the dash sign placed on top and bottom of the characters we will find characters ፯ and ፱. In general the language has its own punctuation marks and style of writing [3]. Since Geez is one of the natural languages, characters of other languages are also applied to it. It has passed various stages, for example, the shape and structure of Geez letters have changed through time. ሀ (used to have shorter leg on the right), ለ (the right leg was shorter and its vowel extension was on the top), ሐ (right leg was very shorter), ተ (the leg was towards the left) [3].

2.2.1 Word Class in Geez (ክፍላተ ቃል)

There are seven word classes in Geez [24]: noun (ስም), verb (ግስ), adjective (ቅጽል),preposition (መስተዋድድ), article (መስተፃምር), Adverb (ተውሳከ ግስ), and pronoun (ተውላጠ ስም).

i. Noun (ስም)

Noun is a word in Geez that is used to identify or address an object. All objects having a definite and indefinite volume are called by a name. Examples include እግዚአብሔር፣ ሰብእ፣ እንስሳ፣ መሬት፣ ነፋስ፣ እሳት፣ ማይ፣ አዳም፣ ኢየሩሳሌም፣ and ፍቅር [25].

ii. Verb (ግስ)

Verb is a word Geez that indicates an action has been done. Examples are ሐየለ/አየለ፣ በረታ፣ ጸና/, መልሐ/መዘዘ ፣ አወጣ/, ከፈለ/ከፈለ ፣ሰጠ/ [3].

Root (Modal) Verbs (ግስ አርእስት)

Root verbs are Geez verbs that are used as a base for other verb forms and those verbs that follow them use the same derivation rules as their modal verb. There are eight root verbs in Geez ቀተለ (ገደለ)፣ ቀደሰ (አመሰገነ)፣ ገብረ (ሠራ)፣ ፈጠረ፣ አእመረ (ዐወቀ)፣ ባረከ (ባረከ)፣ ሤመ (ሾመ)፣ ብህለ (አለ)፣ ቆመ (ቆመ). All verbs in geez start their derivation in past tense form (ቀዳማይ አንቀጽ) [3]. As an example Derivation of root verb ቀደሰ (አመሰገነ) is shown below.

ቀደሰ -- አመሰገነ

ይቄድስ -- ያመሰግናል

ይቀድስ -- ያመሰግን ዘንድ

ይቀድስ -- ያመስግን

ቀድሶ (ቀድሶት) -- ማመስገን

ቀዳሲ -- ያመሰገነ

ቀዳስያን -- ያመሰገኑ

ቀዳሲት -- ያመሰገነች

ቀዳስያት -- ያመሰገኑ /ሴቶች/

ቅዱስ -- ምስጉን

ቀዳሲ -- አመስጋኝ

ቅዳሴ -- ምስጋና

መቅደስ -- ማመስገኛ

ቅድስት -- የተመሰገነች

Table 2.1 shows root verbs and followers of root verbs which follow the same kind of derivation rules.

Table 2.1: *Root and Follower Verbs*

Root verb	**Follower verb having same derivational rule**
ቀተለ (ገደለ)	ነበረ፣ ወረደ
ቀደሰ (አመሰገነ)	ሰብሐ፣ ጸለየ
ገብረ (ሠራ ፣ ፈጠረ)	ሠምረ፣ ጎብረ፣ ነጽሐ
አእመረ (ዐወቀ)	አጥረየ፣ አመንተወ፣ አመኸንየ፣ አመድበለ
ባረከ (ባረከ)	ማህረከ፣ ፃመወ፣ ሳቀየ
ሤመ (ሾመ)	ጌሰ፣ ዜነወ
ብህለ (አለ)	ውጉዘ፣ ሥህወ፣ ጥዕየ
ቆመ (ቆመ)	ሖረ፣ ምርቅሐ (ላጠ)

iii. Adjective/ቅጽል

Adjective is a word in geez that is used to convey additional information about a noun. It gives detail like state of being, physical appearance, distance (near or far), structure and color of the noun under consideration [3]. The underlined words below are examples of adjectives.

<u>ሠናይ</u> ወልድ - A handsome boy

<u>ሠናይት</u> ወለት - A beautiful girl

<u>ንዑስ</u> ወልድ - A small boy

<u>ዐቢይ</u> ወልድ - A great boy

<u>አብድ</u> ወልድ - A lazy boy

<u>ሥሉጥ</u> ወልድ - Active boy

iv. Preposition (መስተዋድድ)

According to [24] prepositions are divided in to two.

a. Prepositions that fall onto nouns to indicate start and end, direction, comparison, time, place etc. examples are እም (ከ)፣ ጎበ (መንገለ፣ ወደ)፣ ከመ (እንደ)፣ በ (በቁሙ)፣ ለ (በቁሙ)፣ አሜ (ጊዜ) ፣ ውስተ (ውስጥ)፣ አፍኣ (ውጪ).

   ጎበ ቤተክርስቲያን -- ወደ ቤተክርስቲያን

   ከመ ቡሩክ -- እንደ ቡሩክ

   እም ገሊላ -- ከገሊላ

b. Prepositions having (of, ‘s) meaning for example ዘ ፣ እንተ ፣ እለ -- የ /of, ‘s/, those that bear the meaning ‘only after/immediately after’ (ከመ/መጠነ -- እንደ) and those used as a conjunction (እስመ ፣ አምጣነ ፣ አኮኑ ፣ ወ -- እና/ና)

ዮም ዘመጸአ ብእሲ --- ዛሬ የመጣው ሰው

ትማልም እንተ መጽአት ብእሲት -- ትናንት የመጣችው ሴት

ጌሠመ እለ ይመጽኡ ሰብእ -- ነገ የሚመጡት ሰዎች

ከመ ሰማዕከ ንግረኒ -- እንደ ሰማህ ንገረኝ /tell me just after you hear/

ተደልው እስመ ይመጽእ መርዓዊ -- ሙሽራው ይመጣልና ተዘጋጁ

አምጣነ ይመጽእ መርዓዊ

አኮኑ ይመጽእ መርዓዊ

v. Article (መስተፃምር)

Article is a word that joins sentences together such as ወ፣ አው፣ ሚመ፣ ዓዲ፣ ሂ፣ ሰ፣ ባህቱ፣ አላ፣ እንበይነዝ፣ በእንተዝ፣ እምዝ፣ እምድኅረዝ etc. Are examples of articles in Geez [24].

vi. Adverb (ተውሳከ ግስ)

As adjectives (ቅጽል) give additional meaning to nouns, adverbs (ተውሳከ ግስ) are words that give additional meaning to verbs. They give information like how, why, where, etc. about verbs [3]. The types of adverbs in Geez are described below.

- Adverb of manner (ኩነታዊ) for example ፍጡን -- በፍጥነት (quickly)
- Adverb of place (መካናዊ) for example ዝየ -- እዚህ (here)
- Adverb of time (ጊዜያዊ) for example ዮም -- ዛሬ (today)
- Adverb of frequency (የደጊመ ጊዜ) for example ኩለሄ -- ኹሌ (usually)
- Adverb of certainty (ርግጠኝነትን የሚገልጽ) for example እሙነ -- በርግጠኝነት (surely)
- Adverb of degree (የከፍታን ገላጭ) for example ጽድቀ -- በአግባቡ (fairly)
- Interrogative (መጠይቃዊ) for example ማዕዜ -- መቼ (when)
- Relative (ተዛማጅ) for example በጊዜ

vii. Pronoun (ተውላጠ ስም)

Pronouns/ተውላጠ ስም are substitute words for a noun. Table 2.2 shows list of pronouns in Geez.

Table 2.2*: List of Geez Pronouns*

		Pronouns (ተውላጠ ስም)	
		Singular	**Plural**
First person	Male	አነ (እኔ)	ንሕነ (እኛ)
	Female	አነ (እኔ)	ንሕነ (እኛ)
Second person	Male	አንተ (አንተ)	አንትሙ (እናንተ)
	Female	አንቲ (አንቺ)	አንትን (እናንተ)
Third person	Male	ውእቱ (እሱ)	ውእቶሙ (እነሱ)
	Female	ይእቲ (እሷ)	ውእቶን (እነሱ)

Below are list of sentences taken from [3] to demonstrate the use of the above pronouns with the verb ሖረ in past tense form (ቀዳማይ አንቀጽ).

አነ ሖርኩ ኀበ ቤተ እግዚአብሔር/እኔ ወደ እግዚአብሔር ቤት ሄድኩ

አንተ ሖርከ ኀበ ቤተ እግዚአብሔር

አንቲ ሖርኪ ኀበ ቤተ እግዚአብሔር

ውእቱ ሖረ ኀበ ቤተ እግዚአብሔር

ይእቲ ሖረት ኀበ ቤተ እግዚአብሔር

ንሕነ ሖርነ ኀበ ቤተ እግዚአብሔር

አንትሙ ሖርክሙ ኀበ ቤተ እግዚአብሔር

አንትን ሖርክን ኀበ ቤተ እግዚአብሔር

ውእቶሙ ሖሩ ኀበ ቤተ እግዚአብሔር

ውእቶን ሖራ ኀበ ቤተ እግዚአብሔር

From the above sentences it is observed that except for pronouns ውእቱ (ሖረ) and ይእቲ (ሖረት) in the remaining the verb's endining letter ረ is changed to (sixth order) ሳድስ (ር), (forth) ራብዕ (ራ) and (second order) ካብእ (ሩ).

2.2.2 Sentences in Geez

A sentence is a sequence of words of Geez language that is used by its speakers to express their thought/idea in spoken or written form. There are two ways to form a sentence in Geez as opposed to Amharic language which always follow one pattern, i.e., Subject + object + verb pattern [26].

a. Subject + verb + object (SVO)
   አልማዝ ገብረት ጽብአ -- አልማዝ ወጥ ሰራች
   ኤልያስ መሀረ ትምህርተ -- ኤልያስ ትምህርትን አስተማረ
   ከልብ በዘበልዓ ይግዕር -- ውሻ በበላበት ይፈጫል
b. Verb + subject + object (VSO)
   ርዕየ መላእክት በሰማይ -- መላእክትን በሰማይ አየ
   ቀተለ ጳውሎስ አርዌ -- ጳውሎስ አውሬ ገደለ
   ትከሉ ሐዋርያት ወይነ -- ሐዋርያት ወይን ተከሉ
   አብዝሃ ወልድየ ብካየ -- ወንድ ልጇ ለቅሶ አበዛ

2.3 The Amharic Language

Amharic is a family of Semitic languages. Amharic is the second most widely spoken Semitic language, next to Arabic. It uses a unique script called hohiyat (ሆህያት) which is conveniently written in a tabular format of seven columns. Both Geez and Amharic languages share the same scripts. The first column represents the basic form and the other orders are derived from it by more or less regular modifications indicating the different vowels. Amharic has 34 base characters and this leads to have a total of 238 (=34*7) Amharic characters. In addition, there are about two scores of characters representing labialized sounds [3].

Word Classes in Amharic

It is stated that words in Amharic belong to five classes [27]. These are noun (ስም), verb (ግስ), adjective (ቅጽል), preposition (መስተዋድድ) and adverb (ተውሳከ ግስ). However, there were additional 3 word classes they are pronoun (ተውላጠ ስም), article (መስተጸምር) and exclamation (ቃለ አጋኖ).

a. Noun (ስም)

Noun is class of words that is used to address thing(s). Nouns are used to indicate gender (two gender types exist in Amharic feminine and masculine) and numbers [5]. The following are examples.

ልጅ ናት -- She is a child/to indicate the gender is feminine.

ልጅ ነው -- He is a child/to indicate the gender is masculine.

ድመት ነው -- it is a cat/to indicate singularity

ድመቶች ናቸው -- they are cats/to indicate plurality

b. Verb (ግስ)

A verb possesses two behaviors that make it different from other word classes. The first one is, it is placed at the end of an Amharic sentence, and the second one is it has suffix attached to it indicating subject of the sentence [27]. The following are examples.

እሱ አንበሳ ገደለ - ኧ -- /he killed a lion/ the underlined word is the verb of the sentence and the suffix 'ኧ' indicates the subject of the sentence is he (እሱ) masculine.

እሷ ምሳዋን በላ - ች -- /she ate her lunch/ the underline word is a verb and the suffix 'ች' indicates the subject is of the sentence is she (እሷ) which is a feminine gender

c. Adjective /ቅጽል/

According to [5] an adjective modifies the noun that it precedes. The following is an example.

አሮጌ ልብስ ነው -- /the cloth is old/ the word 'አሮጌ' modifies the noun 'ልብስ'

d. Preposition /መስተዋድድ/

Author of [27] categorizes both preposition (መስተዋድድ) and article (መስተፃምር) under preposition (መስተዋድድ) since the words in both categories have common characteristics. Preposition possesses two unique characters that make them different from other word classes. Firstly, no word is formed from prepositions either via derivation or other means. Secondly, they don't add any suffix. The following underlined words are examples of prepositions.

አስቴር እንደ አልማዝ ቆንጆ ነች -- Aster is as beautiful as Almaz.

ካሳ ከጎጃም መጣ -- Kassa came from Gojam.

ስለ ትምህርት ጥራት በሰፊው መወያየት አለብን -- We need to have lots of discussions regarding the quality of education.

ለሃገር እድገት እንጨነቅ -- We must care about the development of our country.

e. Adverb /ተውሳከ ግስ/

Words of this class are modifiers of a verb [5]. The most common examples are quickly (ቶሎ), yet (ገና), today (ዛሬ), tomorrow (ነገ), yesterday (ትናንትና), now (አሁን), only (ብቻ), when (መቼ).

2.3.2 Sentences in Amharic

An Amharic sentence is formed from noun phrase (NP) and verb phrase (VP). Regarding their order in a given sentence, NP comes first then VP follows. A phrase is a collection of words arranged in a formal manner. A noun phrase in Amharic is a phrase with a noun as its starting word. አንበሳ ሁለት ላሞችን ገደለ is an example of a noun phrase in Amharic because the starting word አንበሳ/Lion is a noun. A verb in a sentence indicates action or the state of being of something. For example in ካሳ ብርጭቆ ሰበረ, the phrase 'ብርጭቆ ሰበረ' indicates the glass was broken by ካሳ. In ካሳ ነጋዴ ሆነ, the phrase 'ነጋዴ ሆነ' indicates the state of being, i.e., the person named ካሳ has become a merchant. In Amharic verb phrases are formed from Amharic verbs. These verbs indicate either action or the state of being of something. An example of a sentence in Amharic is ሁለት ትልልቅ ልጆች ትናንት በመኪና ወደ ጎጃም ሄዱ. In this sentence the NP is ሁለት ትልልቅ ልጆች and the VP is ትናንት በመኪና ወደ ጎጃም ሄዱ [27].

2.4 Summary

In this Chapter details about NLP like history, approaches to NLP (symbolic, statistical, connectionist and hybrid) and applications (text to speech, speech recognition, machine translation, question answering, and text summarization and document classification) were discussed. Since our research is on machine translation we have tried to review literatures on that area. Moreover, language specific constructs of Geez and Amharic have been presented.

Chapter Three: Related Work

This Chapter presents the review of published articles on Machine Translation that use hybrid architecture. Variants of hybrid Machine Translation systems developed so far will be presented first. Then works that have used word reordering to improve translation quality are discussed. The review of MT systems that are purely rule based and finally a summary that wraps up the Chapter is presented.

3.1 Variants of Hybrid Machine Translation

Hybrid approaches are aimed at combining the best of existing rule based or corpus based MT paradigms. Various researches that exist till now that use hybrid methodologies have simple or complex architectures to combine pure rule based or corpus based techniques [28]. Since MT uses techniques from multiple disciplines various approaches, both from linguistics and statistics perspectives, have been devised to tackle the challenges in the field.

According to [29] hybrid architecture falls into three categories. First those formed via coupling of serial and parallel coupling of exsiting MT approaches, i.e., RBMT, SMT, and EBMT. A good and the most researched example of serial coupling is statistical post editing (SPE) of a rule based machine translation. In parallel coupling the single best translation will be selected by some mechanism from different outputs of several systems. The work in [30] aimed at improving translation of a commercial rule based machine translation by using serial coupling technique. Statistical models were used for automatic post editing. It was tested on parliament and protocol corpus and a promising result was achieved.

By combining outputs of MT a best translation quality could be achieved by means of parallel coupling as implemented in [31]. A technique of searching the best n-grams of all output hypotheses was used and an improvement in BLEU was observed. The second architecture is formed by extension of existing architectures which means that the basic architecture of paradigms will be changed. The modification is done either during pre-processing or core modification, for example, extension of phrase tables. On the contrary during coupling the basic architecture of individual systems is not changed [29].

The third one is genuine hybrid architecture that takes the whole components of respective approaches and combines them to form a new hybrid system. This system uses three main

components. The first component is identification of source language constituents like words and phrases. And then the second component transforms these source language constituents to target language by using bilingual resources. Finally, the last component will generate target language sentences [29]. Works that applied the three mentioned architecture of hybrid machine translation are discussed hereunder.

3.1.1 Serial Coupling

The statistical post editing of a rule based machine translation's output is a serial coupling of hybrid machine translation which is applied in the development of Indonesian English hybrid machine translation. The editing by SMT is done on the output of the rule based machine translation, i.e., English language and produces another English sentence which is the real translation this kind of process can be seen as a target to target translation. The rule based machine translation is developed using the free and open source tool Apertium which follows transfer based approach for translation.

The transfer based approach follows three phases to translate the source sentence into the target sentence. Firstly, through analysis process source language test is transformed into intermediate representation of the source language. Then lexical and structural transfer process follows to transfer the intermediate representation of source language to intermediate representation of the target language. Finally, through generation process the intermediate representation of the target language is transformed into the target language. To test the translation quality of Indonesian English hybrid machine translation developed in [32] two experiments were conducted, one on SMT system the other on hybrid system and the result indicated that on average SMT still outperforms the proposed hybrid system by 8.01%.

A hybrid machine translation system using serial coupling of rule based and a phrase based statistical machine translation for typological divergent languages was developed to translate English string to Hindi string in [33]. The rule based approach follows three steps in the first stage is the source analysis of the English language is performed by running Brill's POS tagger and Stanford's parser on the input and changing it into a chunk based unordered parse tree. The transfer grammar stage then performs local and long distance reordering this reduces the number for rules that must be written in the grammar.

The third stage is the generation step that generates the output which will be used by the phrase based SMT system to produce the final translation. The proposed system was finally tested with

a BLEU metrics and it was reported that the hybrid system showed a significant improvement over baseline rule based machine transition and statistical machine translation with 0.84 and 7.14 respectively.

In [34] Japanese to English machine translation system that translates patent texts by using the coupling of rule based machine translation with statistical post editing was developed. Since paten sentences are long and translating them by only using statistical methods without the use of syntactic analysis often produces strange output, so the authors propose a rule based machine translation with statistical post editing. For the purpose of fluency evaluation two corpuses were used one with only reference translations and the other a large sized US patent corpus of the target language. And a new evaluation measure based on n-gram model called Normalized Mean Grams (NMG) was proposed.

Based on the evaluation results conducted by NMG on the two sets of corpus, both the rule based machine translation and the statistical post editing parts have significant impact on the fluency of the translation. The rule based translation has the advantage on the statistical transfer of the long and complex Japanese patent strings. The statistical post editing part has an advantage for the lexical transfer of technical terms that frequently occur in patent sentences.

A hybrid data driven approach composed of rule based machine translation, example based machine translation, and statistical machine translation that out performs the base line systems, i.e., the example based, rule based and statistical machine translation systems was developed in [35]. In the proposed system the rule based machine translation accepts partial input from both parts- the statistical machine translation part and the example based machine translation part.

To find the accuracy of the proposed approach word error rate (WER) a metric which is based on Levenshtein distance also known as the edit distance. It is computed as the summation of the minimum number for Insertions (I), deletions (D), and substitutions (S) applied to make a sequence similar to the other. The experiment conducted on four sentence type complex sentences, simple sentences, Idioms, and sentences with ambiguity. And the result was compared against Google translate, Babylon, and Microsoft Bing. Except for the result on complex sentence in which Google translate outperforms all, the proposed system has better accuracy compared to the other systems.

3.1.2 Parallel Coupling

A parallel hybrid syntax-based multi-system translation that translates English to Latvian was developed in [36]. It employs a parser to acquire syntactic chunks of English sentences and translates these chunks by using multiple online translation system Application Programming Interfaces (API) and produces a Latvian output string by combining translated chunks to obtain the best possible translation. The proposed system has three main components pre-processing of English sentences, the acquisition of translations by the online APIs, and Post-Processing the selection of the best translation of chunks and generation of the output.

The using the Berkley Parser the source sentence is divided into linguistically motivated chunks to be used by the online APIs. Three APIs were used for the translation of chunks generated Google translate, Bing Translate, and LetsMT. The reason for selecting these APIs was there availability to the public, descriptive documentation and support for the translation of source and target language. Finally the selection of the best translation was performed by calculating the perplexity of each translation hypothesis. The proposed system demonstrates an improvement in terms of BLEU and National Institute of Standards and Technology (NIST) evaluation scores as compared to the baseline hybrid MT without the syntax-based processing.

3.1.3 Pure Hybrid

Pure Hybrid architectures don't just incorporate an additional add-on rather they combine components of already existing architectures into a novel system [31]. The basic components of a pure hybrid system are identification of source language chunks-words or phrases-, transformation of the identified chunks into target language by using bilingual resources, and target language generation. Recently, there are a lot of works done using this architecture like [37] which circumvents the need of parallel corpora by simply using a full-form bilingual dictionary. An n-gram window is moved over the sentence, and all words in the window are translated using the full-form bilingual dictionary; based on these translations, the target language corpus is searched for the closest n-gram. The result is a lattice of n-gram translations. The translation with the strongest left and right overlaps, and the highest density of terms, are selected by the decoder.

3.2 Word Reordering

Nowadays researches on MT using syntactical and morphological information have started exploring tree based machine translation. The MT system developed in [38] focused on translation of English to Malayalam using syntactic based machine translation technique. Based on part of speech tagging of English source texts various rules for these texts were identified. For translation a bilingual English-Malayalam dictionary and a morphology generator was used. To change English pattern, i.e., Subject-Verb-Object (SVO) to Malayalam pattern, i.e., Subject-Object-Verb (SOV) word reordering was performed on the generated syntax tree of source English text in order to improve translation quality of the system.

The English to Malayalam MT system has four modules, the module responsible for preprocessing of source English text. This sub system, called syntax tree generator using Stanford parser, generates syntax tree, parts of speech tagging and dependency information. Syntax based reordering is used in the second module, i.e., word reordering module. In pattern recognition module 10 phrase patterns based on the dependency information generated from the syntax tree generator were identified, for example (subject + verb, subject + verb + indirectobject + directobject, etc.). Finally the translation module translates the reordered English sentence to Malayalam using a bilingual English Malayalam dictionary. Plus using the rules for translating nouns, pronouns and verbs morphological generator is used by the translation module.

Since source and target languages use different word orders, reordering is one of the fundamental issues that a statistical machine translation suffers from. If not handled in MT system that uses statistical approach the final translation quality will definitely deteriorate. In [39] a Chinese to Mongolian translation system with pre-ordering method by applying manually written reordering rules on phrase structure sub trees was proposed and a 1.7 increase in BLEU over the base line phrase based statistical machine translation was achived.

Phrase structures that correspond to source sentences were firstly constructed. Then for the implementation of reordering of source sentences manually prepared reordering rules that are prepared by linguistic experts are applied on the matched sub tress. These rules integrate phrase structure labels with POS tags. Using the manually prepared rules not only reordering among words but also the reordering between words and phrases were easily implemented. Three steps were followed in order to implement sub tree reordering. First, phrase structure tree was constructed for Chinese source text. Second, it finds subtrees that match with the original

linguistic rules of the constructed phrase structure tree. And finally, reordering rules are applied on the matched subtrees to swap left branches with right branches.

In order to get grammatically correct translation from a general purpose phrase based translation system, two core steps were followed that are aimed at matching the word ordering between source and target language. The first step is learning of reordering rules. In this step there are three core sub steps involved starting from parsing the source language in bilingual corpus and then discarding those with low word-aligning ratio to be excluded from training candidate. Secondly, collecting training instances and finally applying conditional random fields (CRFs) to estimate lexical and grammatical items in rules to determine target language word order. Before the source text is fed to the decoder a final preprocessing step that reorders source sentence during runtime is performed [40].

English to Malayalam phrase based machine translation using phrase based MT system was developed in [41]. Malayalam is an Indian language hence follows subject + object + verb pattern, as opposed to English which is subject + verb + object sentence pattern. So preprocessing of training data was proposed to improve the translation quality of phrase based systems. The preprocessing involves two core steps First using Stanford dependency parser reordering source English text was performed so that word order would be same as Malayalam. Second using a morph analyzer suffix was removed from root words for both source and target languages. Then the preprocessed training was used for phrase based translation and better result was achieved plus the training data was suitable for statistical machine translation. Owing to the heterogeneity in syntactic structure of natural languages, word ordering plays a huge role in improving translation quality of MT systems.

3.3 Rule Based Translation

A purely rule based English to Tamil translation was developed in [42]. English and Tamil follow different word ordering subject-object-verb pattern is followed by Tamil and subject-verb-object by English. This is the main reason for the work to follow syntax transfer based approach. A parser is used as the translation engine of the proposed system that analyzes English text and then by using transfer lexicon target Tamil text is generated.

The proposed rule based translation system is composed of five core modules. Root words and feature equations of source text are obtained by morph analyzer which is the first module of the

system. Parts of speech tagger and word sense disambiguation is the second module. The tagger assigns word class to each word of the source sentence and the disambiguation process identifies in which sense a word was meant in the given text. The parser is the third module for analyzing source text. From a single source structure one or more target sentences are generated from the forth module, i.e., target generator. The fifth and last module is the morph generator that handles target text morphology.

English to Amharic rule based machine translation using L^3 framework - a framework that is based on Extensible Dependency Grammar (XDG) is developed in [43]. The focus was to examine the advantages of the framework in handling the structural divergence between the two languages. XDG, which is the basic building block of the aforementioned framework, is based on a lexical, i.e., means the basic units are words. Analyzing a sentence using XDG results in a directed graph having nodes, that represent words of the source sentence plus a root node for a punctuation mark that represents the end of a sentence. The framework has also allowed the integration of shallow and deep analysis into a unified system.

3.4 Summary

In this Chapter we have reviewed research works that we think are useful in designing our proposed work on Geez to Amharic machine translation. To the best of our knowledge there doesn't exist any literature on MT of Geez language. Thus, we have tried to review works especially on languages which are morphologically complex/rich in nature like Geez and Amharic. Hence, the aforementioned researches are valuable input to the proposed work.

Chapter Four: The Proposed Solution

This Chapter details the architectural design of Geez to Amharic machine translation system. The proposed system is a hybrid of rule based technique and statistical machine translation system using serial coupling technique. Since Geez and Amharic languages use different sentence/phrase construction rules a rule based preprocessing technique is applied to preprocess a manually part of speech (POS) tagged Geez corpus which is composed of a set of activities. To perform this preprocessing a Rule Based Geez Corpus Preprocessor component is developed. After the POS tagged Geez corpus is supplied to this component it reads all sentences from the text file and stores each in an array. Then one by one each sentence is processed to determine its POS pattern after determining the POS a reordering rule that is suitable to the pattern determined is applied on the sentence being processed so that sentences in Geez and Amharic corpus will have similar structure which increases translation quality of the Baseline SMT. After the reordering rule is applied the reordered sentence is written to another text file without POS information to be later used in training as well as translation by the Baseline SMT. This process continues for each sentence stored in the array. Finally, after the preprocessing is completed the file containing reordered Geez sentences together with the Amharic corpus will be supplied to the Baseline SMT.

The Baseline SMT takes the parallel corpus (composed of preprocessed Geez corpus and Amharic corpus) and performs training to train the Translation Model. To translate a given sentence from Geez to Amharic the Translation Model calculates the probability of a Geez sentence (g) given an Amharic sentence (a) denoted as $p(g|a)$. The Language Model fills the deficiencies of the Training Model by ensuring fluency of target texts, i.e., Amharic texts. It only need one Amharic corpus in this case and calculates $p(a)$. The final translation of Geez to Amharic is done by the Decoder.

4.1 System Architecture

The proposed system is composed of two main components rule based Geez Corpus Preprocessor and Baseline statistical machine translation (SMT). The contribution of this research work is the rule based corpus preprocessor which is surrounded with dashed line and its components are indicated by shaded background in Figure 4.1. The Baseline SMT is the standard statistical machine translation system and is taken from [44].

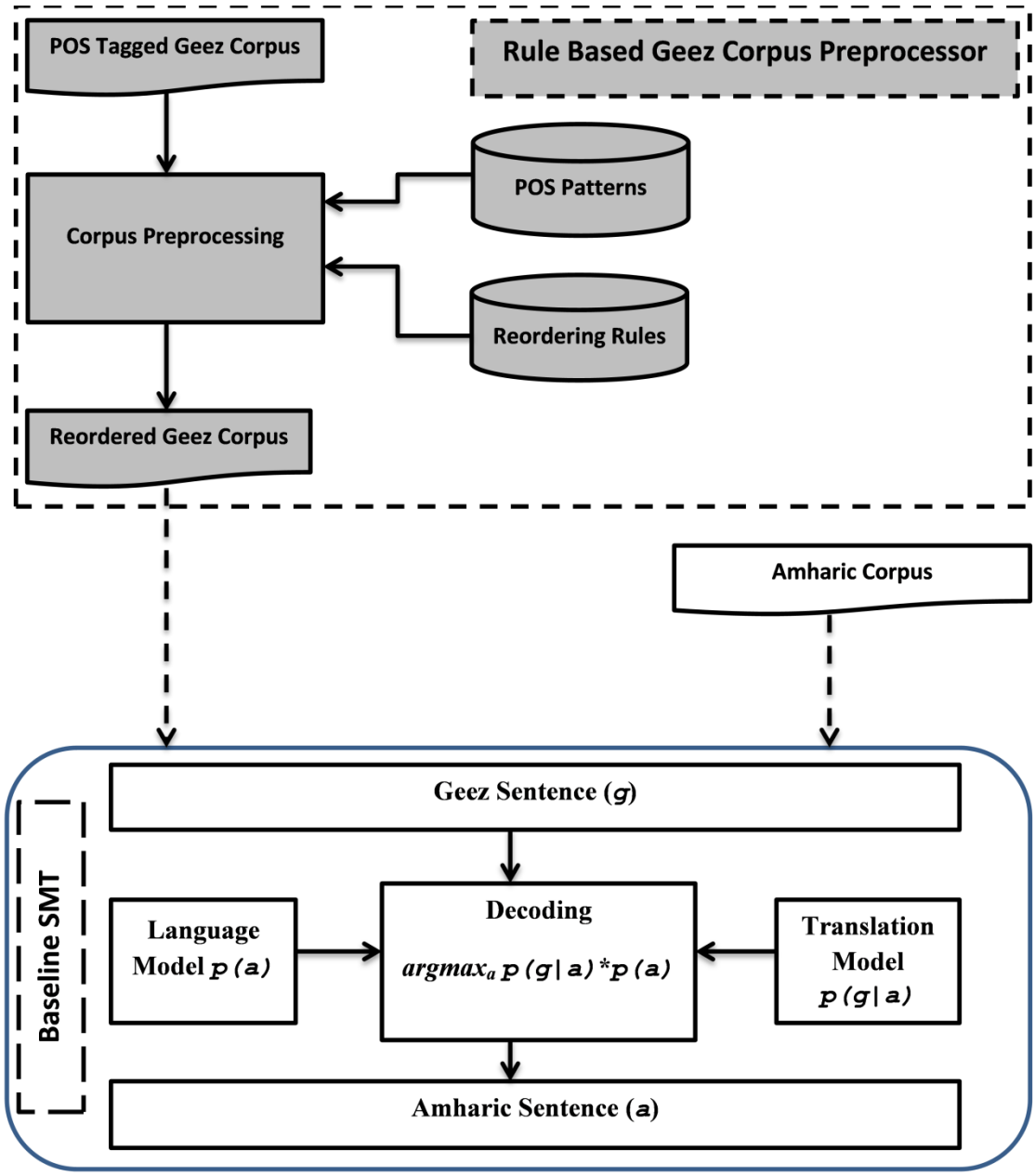

Figure 4.1: *Architectural Design of Geez to Amharic Machine Translation System*

4.2 Rule Based Geez Corpus Preprocessor

The Rule Based Geez Corpus Preprocessor is first component of the proposed Hybrid Machine Translation System that translates Geez text to Amharic text. The objective is to make the input Geez sentence similar in structure to that of an Amharic sentence in terms of word ordering. It

has five subcomponents POS Tagged Geez Corpus, POS Patterns, Reordering Rules, Corpus Preprocessing, and Reordered Geez Corpus. Each is described in detail in Sections 4.2.1, 4.2.2, 4.2.3, 4.2.4, and 4.2.5 respectively.

4.2.1 POS Tagged Geez Corpus

The first subcomponent of the Rule Based Geez Corpus Preprocessor which contains 976 manually POS tagged Geez sentences that are taken from [25]. The tagging is done manually as there is no automated POS tagger for the language. As the translation is done between local languages each tag is enclosed within < > to demark it from the word that it tags and written using Geez letters, so no need of translating tags to English. In the Geez sentence አብርሃም<ባለቤት> ተወክፎ<ግስ> ለነግድ<ተሳቢ>/Abraham welcomed his gust/አብርሃም እንግዳውን ተቀበለው that is taken from the corpus, contains three tags <ባለቤት><ግስ><ተሳቢ> each representing three different words <ባለቤት> is tag for አብርሃም, <ግስ> tags ተወክፎ, and <ተሳቢ> is tag of ለነግድ. There are 40 tags used in this work and each is explained hereunder.

i. Subject/አድራጊ.ስም/ባለቤት

In Geez language Subject/አድራጊ.ስም/ባለቤት is the person or thing about whom the statement is made. For example in the phrase ሐጸነት እምየ ኪያየ ወለታ/"I am raised by my mother" which means እናቴ እኔን ልጇን አሳድጋለች in Amharic 'እምየ/My mother' is the subject. It was the second word in the Geez phrase while it appears first in the Amharic phrase.

ii. Object/ተሳቢ

An object in Geez sentence shows who or what the action of the verb affects. For example ወነጸርኩ አነ ዳንኤል/"I have seen Daniel" which translates to the Amharic sentence አኔ ዳንኤልን አይቻለሁ. In the sentence ዳንኤል is the object that shows አነ/I the subject has seen/ወነጸርኩ Daniel/ዳንኤል.

iii. Objective pronoun/ተሳቢ.መራህያን

When pronouns in a sentence came after subject of a sentence they are referred to as objective pronouns. For example the phrase ፈጠረ እግዚአብሔር ኪያነ ሕዝቦ/"God created our society"/እግዚያብሔር እኛን ህዝቦቹን ፈጥሯል the word ኪያነ/our serves as an objective pronoun since it follows the subject እግዚአብሔር.

iv. Noun/ስም

A noun is the word that refers to a person, thing or abstract idea. A noun can tell who or what. Examples include, name of a person (ቡሩክ/Biruk), places (ኢየሩሳሌም/Jerusalem), or things ፍስሐ/Happiness).

v. Noun describing behavior/ስመ.ባሕርይ

This is a sub category of noun used for naming that also expresses the nature of the entity being named, for example እግዚአብሔር/God.

vi. Noun to address person or thing/ስመ.ተጸውዖ

A noun used to call/address something/someone without indicating the behavior of the entity, for example አዳም/Adam.

vii. Objective noun/ተሳቢ.ስም/ተደራጊ.ስም

It is the noun that receives the action of the verb. for example in the sentence ንጉሠ ዳዊትሀ ቀብአ ሳሙኤል/"Samuel Anointed David as king" the word ዳዊትሀ/David is the objective noun that indicates the action of anointment performed on him by Samuel.

viii. Noun to address person or thing used as an object/ተሳቢ.ስመ.ተጸውዖ

Is used to indicate the noun (the noun doesn't indicate the behavior of the thing/person/place) on which an action is performed by the subject. For example, the sentence እግዚአብሔር ፈጠረ አዳምሀ which means እግዚአብሔር አዳምን ፈጠረ in Amharic. The word አዳምሀ indicates that አዳም/Adam was created by God.

ix. State of being/ማሰ.ዐን/መዝጊያ

A word that describes a condition or situation exist for example አኮ ጻድቅ ይሁዳ/"Judah is not righteous" which means ይሁዳ ጻድቅ አይደለም in Amharic. The word 'አኮ/አይደለም/is not' is confirming that ይሁዳ/ Judah is not a righteous person.

x. Collective Nouns/መድበል.ስመ.ተጸውዖ

This is a sub class of a noun that refers to a group of people or things for naming that can be visible, invisible, and having no gender category. ሕዝብ/people ፣ ፍጥረት/creatures are some of the examples belonging to this category.

xi. Descriptive Adjective/ግልጽ.ዘ.ቅጽል

This is a sub class of an adjective used to name an attribute of a noun and having the Geez letter ዘ explicitly as the first letter of the adjective**.** Example ዘከብረ ብእሲ/"Respected Person" when translated to Amharic becomes የከበረ ሰው. In the Geez phrase ዘከብረ is playing the role of the descriptive adjective.

xii. Adverb/ዐንቀጽ

This is a word class that describes an action, doer of an action, and situation of the action in terms of gender (male/female), closeness to the writer/speaker (this/that/those), time (past/present/future), and number (singular/plural). For example, ኮነ ወልድየ ወሬዛ/"My Child has become a grown man" when translated to Amharic ልጄ ጎልማሳ ሆነ the word ኮነ indicates that the child has grown.

xiii. Verb with letter ዘ/ግልጽ.ዘ.ዐንቀጽ

This is a sub class of verb having the Geez letter ዘ as the first letter consider the sentence ዘኤይነውሙ ሚካኤልወገብርኤል ይሰእሉ ምሕረተ/"Both Michel and Gabriel beg for mercy continuously". The word ዘኤይነውሙ starts with the letter ዘ.

xiv. Verb to have/ነባር.ዐንቀጽ

This is a sub class of verb that is used to express positive (ዐ in Amharic አለ፣ኖረ፣ነበረ፣ ኖሯል)/negative (አልዐ in Amharic የለም፣አልኖረም) affirmation past, present, and future situation ውእቱ in Amharic ነው፣ሆኗል፣ይሁን).

xv. Adjective used in place of noun/በቂ.ውስጠ.ዘ

This is an adjective used in place of a noun that can be used as subject of the sentence. For example the word ዐብድ in the sentence ዐብድ ይብል አልዐ እግዚአብሔር/" a lazy person says God doesn't exist" plays that role. The Amharic translation is ሰነፍ እግዚአብሔር የለም ይላል.

xvi. Passive voice/ተገብሮ.ግስ

Its role is to indicate that a subject has been put in action and it doesn't need an objective noun. Example ወንጌል ተጽሕፈ the word ተጽሕፈ is the passive voice it indicates the ወንጌል/Gospel has been written; the translation in Amharic is ወንጌል ተጻፈ.

xvii. Pronoun/መራሕያን

A pronoun is defined as a word or phrase that may be substituted for a noun or noun phrase, for example I/አነ/እኔ.

xviii. Noun to address person or thing used as an object/ገቢር.ስመ.ተጸውዖ

A noun (noun to address person or thing describing the behavior) used as an object of the sentence. For example, in እግዚአብሔር ፈጠረ አዳምሀ ወሔዋንህ the compound word አዳምሀ ወሔዋንህ plays this role.

xix. Adjective/ቅጽል

Adjective in Geez is a word class that is used to indicate behavior, volume, and nature of a noun. For example ዘ/shows belongingness and means የ in Amharic. ክቡር/respected means የከበረ in Amharic.

xx. Adjective ending with sixth order letter/ሳድስ.ውስ.ቅጽል

This is a sub class of an adjective having its last letter on the sixth order (ሳድስ) of Geez alphabet "ሰ፣ሱ፣ሲ፣ሳ፣ሴ፣ስ" so ስ is the ሳድስ/"sixth order" for example ልዑል. The name "ውስጠዘ" is given to this category of adjective because it bears the meaning of የ in Amharic which has ዘ as its equivalent meaning in Geez, i.e., the Geez letter is not explicitly written rather understood implicitly.

xxi. Adjective ending with third order letter/ሣልስ.ው.ቅጽል

This is a sub class of an adjective having its last letter on the third order (ሣልስ) of Geez alphabet "ሰ፣ሱ፣ሲ"so ሲ is the ሣልስ/"third order" to mean on the third order for example ለባዊ. The name "ውስጠዘ" is given to this category of adjective because it bears the meaning የ in Amharic which

has ዘ as its equivalent meaning in Geez, i.e., the Geez letter is not explicitly written rather understood implicitly.

xxii. Collective adjective/መድበል.አሐዝ.ቅጽል

This is a sub class of an adjective that shows the number of entities mentioned/contained in a noun collectively. It is used when that number is more than two for example ብዙሕ/many/ which means ብዙ in Amharic.

xxiii. Pronoun as an adjective/ደቂቅ.ቅጽል

This is a sub class of a pronoun that appears before a noun in a sentence and is used to indicate a person, thing, place, and time for example ዝንቱ/ይህ/This and ዝኩ/ያ/That.

xxiv. Active voice/ገቢር.ግስ

An active voice describes a sentence where the subject performs the action stated by the verb. In active voice it needs both subject and object of the sentence. The word ጸሐፈ serves as an active verb in the sentence ማቴዎስ ጸሐፈ ወንጌለ/"Matthew has written the Gospel" when translated to Amharic ማቴዎስ ወንጌልን ጻፈ.

xxv. Root verbs/አርእስተ.ግስ

In Geez language there are eight root verbs. They are ቀተለ፣ቀደሰ፣ገብረ፣ተንበለ፣ባረከ፣ኤለ፣ከህለ፣ያደ. They are called roots because other regular verbs formation from their infinitive form follows the same procedure that they follow.

xxvi. Verbs starting with (አ፣ነ፣ተ፣የ)/አስራውና.አናቅጽ

This is a sub class of verbs that start with letters (አ፣ነ፣ተ፣የ) and thier forms. The sentence አነ አሐነግር ወልድየ/"I carry my child on my shoulder" the word አሐነግር is the verb starting with the letter አ. When translated in Amharic እኔ ልጄን እሸኮኮ እላለሁ.

xxvii. Possessives/ዝርዝር

Are used to expresses belonging to or ownership and are added at the end of Geez word. Examples include ከ፣ክሙ፣ዎ፣ዎሙ፣ኪ፣ክን etc.

xxviii. Possessive nouns/ስምና.ዝርዝር

Noun suffixed with ዝርዝር/Possessives. In the sentence ሰሎሞን ሐነጸከ ቤትየ the word ቤትየ plays this role ቤት being the noun and የ being the possessive when translated to Amharic will be ሰሎሞን ቤቴን ሠራህ.

xxix. **Preposition/አገባብ**

This is placed at the beginning, middle, and at the end and used to link words or phrases to other words or phrases within a sentence. Example is ኀበ/To/ወደ, እስመ/like/እንደ.

xxx. Prepositions/ዐብይ.አገባብ

This is a sub class of preposition that is placed before noun or verb and link words or phrases using past tense. Example is እስመ/like/እንደ.

xxxi. Prepositions/ደቂቅ.አገባብ

This is a sub class of preposition that is placed before noun and link words or phrases. Example is ማእከለ/in the middle/በመካከል.

xxxii. Preposition for noun and verbs/ንኡስ.አገባብ

This is a sub class of preposition that is placed before noun or verb and links words or phrases. Example is እፎ ይከውን ዝንቱ/"how could this be?"/ይህ እንዴት ይሁን?

xxxiii. Prepositions + to + objective pronouns/ተጨቃሽ

This is added at the end of a verb to link the verb to a pronoun (አነ፣አንተ፣አንቲ፣ ውእቱ፣ ይእቲ፣ንሕነ፣ አንትሙ፣አንትን፣ ውእቶሙ፣ ውእቶን). Examples are （ከ፣ክሙ፣ኪ፣ክን፣ዎ፣ዎሙ፣ዋ፣ዎን） to indicate to which pronoun that the subject performs the action on.

xxxiv. States of being for past and present/ነባር.ማሰ.አንቀጽ

This is a word that describes a condition or situation that exists in the past and present like ውእቱ/is/ነው. Example is ምንትየ ውእቱ አብርሃም/what is Abraham to me/አብርሃም ምኔ ነው. The word ውእቱ/is/ነው is the state of being for the present time.

xxxv. A noun following an adjective/ዘርፍ

When an adjective comes at the beginning of a sentence and two nouns follow consecutively the noun that the adjective that is describing the noun-the first noun- is called a noun just after the adjective. Example is ቅቱለ ቃኤል አቤል ኮነ ሰማዕተ/"Abel who was killed by Cain become a martyr" when translated to Amharic ቃኤል የገደለው አቤል ሰማዕተ ሆነ. The noun Abel is the noun after the adjective.

xxxvi. Adjective that appears as a noun/መስም.ው.ቅጽል

This is a sub class of an adjective that starts with the Geez letter መ. The name መስም is given because it looks like a noun. For example, the word መፈክር in the sentence መፈክር ዳንኤል ፈከረ ሕልመ/"The dream interpreter Daniel had a dream" plays this role and when translated into Amharic it would be ተርጓሚ ዳንኤል ሕልምን ተረጎመ.

xxxvii. Adjectival quantifier/አሐዝ.ቅጽል

This is a sub class of an adjective used to indicate number and physical size of a noun. Example is አሐዱ ብእሲ ይወርድ ጎበኢየሩሳሌም/"A man went to Jerusalem" translated into Amharic አንድ ሰው ወደ ኢየሩሳሌም ይወርድ ነበር. In the Geez sentence the word አሐዱ is the adjectival quantifier.

xxxviii. Descriptive adjective for a noun following and adjective/ዘርእ.ው.ቅጽል

This is a sub class of an adjective that is used to describe the so called "noun after an adjective/ዘርፍ". The word ክብረ in ክብረ ቅዱሳን ክርስቶስ ገብረ ሰላመ is the adjective describing the noun ቅዱሳን/saints. When the sentence is translated to Amharic it would be የቅዱሳን ክብር ክርስቶስ ሰላምን አደረገ.

xxxix. Adjective that gives meaning to a noun/መተርጉም.ው.ቅጽል

This is a sub class of an adjective that is used to give meaning to a noun. Example is in the sentence ያዕቆብ እስራኤል ወለደ ደቂቀ the word እስራኤል gives meaning to the subject ያዕቆብ/Jacob. The translation in Amharic is እስራኤል የተባለ ያዕቆብ ልጆቹን ወለደ which indicates that another name of ያዕቆብ is እስራኤል.

xl. Interrogative adjective/ንኡስቅጽል

This is a sub class of an adjective used to ask a question regarding persons, thing, time, and place. It includes words like መኑ፣ ምንት፣ አይ፣ ማእዜ. The sentence መኑ መልአክ ከሀለ ነጽሮቶ the word መኑ/ማን/who is the interrogative adjective. When translated to Amharic it would be ማን መልአክ ማየቱን ቻለ.

4.2.2 POS Patterns

Both Geez and Amharic sentences in the corpus are taken from [25], Geez sentences in the corpus are constructed from number of words varying between one and five. As explained in section 2.2.2, Geez sentence follows both type of patterns (SVO=subject + verb+ object and VSO=verb + subject+ object) while Amharic follows only one (SOV= Subject + object + verb). Sentences in the Geez corpus are annotated with POS information and 40 tags are used for this work as explained in section 4.2.1. The sequence of these tags in each sentence being preprocessed is used to indicate the POS pattern in which the sentence is written in. For ease of implementing this subcomponent which is not due to any linguistic nature of Geez, POS Patterns are grouped into four categories. The grouping only indicates the number of words that a sentence from a certain group is made up of. Sentences that are made up of only one word doesn't require any POS Pattern determination nor application of reordering rules hence sentences containing two or more words needs preprocessing. Each group is described hereunder. For this study 126 POS Patterns that are taken from [25] that are also checked by a person with Geez knowledge have been implemented by this subcomponent.

1. Group Two POS Patterns

Sentences in this Group are made up of two words and two tags. There are 22 POS Patterns in this group. For example, the sentence ንግበር<ማሰ.ዐን> ሰብአ<ስመ.ባሕርይ> is written with POS pattern <ማሰ.ዐን><ስመ.ባሕርይ>. Table 4.1 lists all patterns and reordering rules belonging to group two.

Table 4.1: *Group Two Sentence POS Patterns*

No.	POS Patterns
1	<ባለቤት> <ተሳቢ.ስም>
2	<ማሰ.ዐን> <ስመ.ባሕርይ>
3	<ስመ.ባሕርይ><ማሰ.ዐን>

No.	POS Patterns
4	<ስም><ማሰ.ዐን>
5	<ማሰ.ዐን><ስም>
6	<ስመ.ተጸውዖ><ማሰ.ዐን>
7	<መድበል.ስመ.ተጸውዖ><ማሰ.ዐን>
8	<ባለቤት><ማሰ.ዐን>
9	<ግልጽ.ዘ.ቅጽል><ስም>
10	<ንኡስ.አንቀጽ><ስም>
11	<በቂ.ውስጠ.ዘ><ማሰ.ዐን>
12	<አድራጊ.ስም><ግስ>
13	<ግስ> <አድራጊ.ስም>
14	<ተደራጊ.ስም><ተገብሮ.ግስ>
15	<ተደራጊ.ስም><ተገብሮ.ስም>
16	<ተገብሮ.ግስ><ተደራጊ.ስም>
17	<ባለቤት><ተገብሮ.ግስ>
18	<ተገብሮ.ግስ><ባለቤት>
19	<ተገብሮ.ግስ><ተሳቢ>
20	<ባለቤት><ግስ>
21	<ግስ><ባለቤት>
22	<ግስ><ስም>

2. Group Three POS Patterns

Sentences that constitutes thee Geez words along with thee tags fall under this category. There are 59 POS patterns in group thee. Example is the sentence እግዚአብሔር<ባለቤት> ኢይሜንን<ማሰ.ዐን> ፍጥረተ<ስመ.ተጸውዖ> with POS pattern <ባለቤት><ማሰ.ዐን><ስመ.ተጸውዖ>. Table 4.2 lists all POS patterns of this group.

Table 4.2: *Group Three Sentence POS Patterns*

No.	POS Patterns
1	<ባለቤት> <ግስ> <ተሳቢ.ስም>
2	<ባለቤት> <ግስ> <ተሳቢ.መራህያን>
3	<ግስ> <ባለቤት> <ተሳቢ>
4	<መራሕያን> <ባለቤት> <ግስ>
5	<ማሰ.ዐን> <ስመ.ባሕርይ> <ተሳቢ>
6	<ስመ.ባሕርይ> <ማሰ.ዐን> <ተሳቢ>
7	<ተሳቢ> <ማሰ.ዐን> <ስመ.ባሕርይ>
8	<ማሰ.ዐን> <ስመ.ባሕርይ> <ማሰ.ዐን>
9	<ማሰ.ዐን> <ማሰ.ዐን> <ስመ.ባሕርይ>
10	<ማሰ.ዐን> <ተሳቢ> <ስመ.ባሕርይ>
11	<ባለቤት> <ማሰ.ዐን> <ስመ.ባሕርይ>
12	<ማሰ.ዐን> <ባለቤት> <ስመ.ባሕርይ>
13	<ስመ.ባሕርይ><ማሰ.ዐን><ባለቤት>
14	<ስመተጸውዖ><ማሰ.ዐን><ተሳቢ>
15	<ማሰ.ዐን><ባለቤት><ተሳቢ>
16	<ተሳቢ><ማሰ.ዐን><ስመተጸውዖ>
17	<ስመተጸውዖ><ማሰ.ዐን> <ተሳቢ>
18	<ስመ.ተጸውዖ><ማሰ.ዐን> <ተሳቢ>
19	<ማሰ.ዐን><ስመተጸውዖ><ተሳቢ>
20	<ባለቤት><ማሰ.ዐን><ተሳቢ>
21	<ማሰ.ዐን><ስም><ተሳቢ>
22	<ተሳቢ><ማሰ.ዐን><ባለቤት>
23	<ስም><ማሰ.ዐን><ተሳቢ>
24	<ተሳቢ><ማሰ.ዐን><ስም>
25	<መድበል.ስመ.ተጸውዖ><ማሰ.ዐን><ተሳቢ>
26	<ባለቤት><ማሰ.ዐን><ተሳቢ.ስመ.ተጸውዖ>
27	<ማሰ.ዐን><ባለቤት><ተሳቢ.ስመ.ተጸውዖ>
28	<ተሳቢ.ስመ.ተጸውዖ><ማሰ.ዐን><ባለቤት>

No.	POS Patterns
29	<ባለቤት><ማሰ.ዐን><ገቢር.ስመ.ተጸውዖ>
30	<ገቢር.ስመ.ተጸውዖ><ማሰ.ዐን><ባለቤት>
31	<ባለቤት><ማሰ.ዐን><ስመ.ተጸውዖ>
32	<ሳድስ.ውስ.ቅጽል><ስም><ማሰ.ዐን>
33	<ስም><ሳድስ.ውስ.ቅጽል><ማሰ.ዐን>
34	<መድብል.አሐዝ.ቅጽል><ስም><ማሰ.ዐን>
35	<ስም><መድብል.አሐዝ.ቅጽል><ማሰ.ዐን>
36	<ደቂቅ.ቅጽል><ስም><ማሰ.ዐን>
37	<ስም><ዐንቀጽ><ማሰ.ዐን>
38	<ስም><ዐንቀጽ><ተሳቢ>
39	<ግልጽ.ዘ.ዐንቀጽ><ስም><ማሰ.ዐን>
40	<ቅጽል><ስም><ማሰ.ዐን>
41	<በቂ.ውስጠ.ዘ><ማሰ.ዐን><ተሳቢ>
42	<ማሰ.ዐን><በቂ.ውስጠ.ዘ><ስም>
43	<በቂ.ውስጠ.ዘ><ተሳቢ><ማሰ.ዐን>
44	<አድራጊ.ስም><ግስ><ተደራጊ.ስም>
45	<ግስ><አድራጊ.ስም><ተደራጊ.ስም>
46	<አድራጊ.ስም><ገቢር.ግስ><ተደራጊ.ስም>
47	<ገቢር.ግስ><አድራጊ.ስም><ተደራጊ.ስም>
48	<ባለቤት><ገቢር.ግስ><ተሳቢ>
49	<ገቢር.ስም><ባለቤት><ተሳቢ>
50	<ባለቤት><ግስ><ተሳቢ>
51	<ግስ><ባለቤት><ተሳቢ>
52	<ባለቤት><ነባር.ዐንቀጽ><ተሳቢ>
53	<ስም><ግስ><ስምና.ዝርዝር>
54	<ባለቤት><ንኡስ.አግባብ.ዝርዝር><ተጠቃሽ>
55	<ንኡስ.አግባብ><ነባር.ማሰ.አንቀጽ><ባለቤት>
56	<ባለቤት><ግስ><ደቂቅ.አግባብ>
57	<አርእስተ.ግስ><ባለቤት><ተሳቢ>

No.	POS Patterns
58	<ባለቤት><አስራውና.አናቅጽ><ተሳቢ>
59	<ባለቤት><አስራውና.ግስ><ተሳቢ>

3. Group Four POS Patterns

The fourth group of sentences is made up of four words and corresponding four POS tags. For this group there are 30 POS patterns. For example ኒቆዲሞስ<ስም> ድሑን<ሳድስ.ውስ.ቅጽል> ሰብሐ<ማሰ.ዐን> ፈጣሪሁ<ተሳቢ> has a pattern <ስም><ሳድስ.ውስ.ቅጽል><ማሰ.ዐን><ተሳቢ>. Table 4.3 lists POS patterns belonging to this group.

Table 4.3*: Group Four Sentence POS Patterns*

No.	POS Patterns
1	<መራሕያን> <ባለቤት> <ግስ> <ተሳቢ.ስም>
2	<ማሰሪያ.አንቀጽ> <ባለቤት> <ተሳቢ.መራሕያን> <ተሳቢ>
3	<ሳድስ.ውስ.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>
4	<ስም><ሳድስ.ውስ.ቅጽል><ማሰ.ዐን><ተሳቢ>
5	<ሳድስ.ውስ.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን>
6	<ሣልስ.ው.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>
7	<ስም><ሣልስ.ው.ቅጽል><ማሰ.ዐን><ተሳቢ>
8	<መስም.ው.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>
9	<አሐዝ.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>
10	<ስም><መድብል.አሐዝ.ቅጽል><ማሰ.ዐን><ተሳቢ>
11	<ዘርእ.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን>
12	<መድበል.ው.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>
13	<ስም><መተርጉም.ው.ቅጽል><ማሰ.ዐን><ተሳቢ>
14	<ንኡስቅጽል><ስም><ማሰ.ዐን><ተሳቢ>
15	<ደቂቅ.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>
16	<ስም><ዐንቀጽ><ማሰ.ዐን><ተሳቢ>
17	<ግልጽ.ዘ.ዐንቀጽ><ስም><ማሰ.ዐን><ተሳቢ>
18	<ውስጠ.ዘ.ቅ><ስም><ግልጽ.ዘ.ዐንቀጽ><ማሰ.ዐን>

No.	POS Patterns
19	<ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>
20	<ስም><ቅጽል><ማሰ.ዐን><ተሳቢ>
21	<ማሰ.ዐን><በቂ.ውስጠ.ዘ><ስም><ተሳቢ>
22	<በቂ.ውስጠ.ዘ><ተሳቢ><ማሰ.ዐን><ተሳቢ>
23	<ውስጠ.ዘ.ስም><ውስጠ.ቅጽል><ማሰ.ዐን><ተሳቢ>
24	<ውስጠ.ቅጽል> <ውስጠ.ዘ.ስም> <ማሰ.ዐን> <ተሳቢ>
25	<ውስጠ.ዘ.ስም><ግልጽ.ዘ.ቅጽል><አንቀጽ><ማሰ.ዐን>
26	<ግልጽ.ዘ.ቅጽል><አንቀጽ><ውስጠ.ዘ.ስም><ማሰ.ዐን>
27	<ተሳቢ.ቅጽል><ተሳቢ.ስም><ማሰ.ዐን><ባለቤት>
28	<ባለቤት><ማሰ.ዐን><ተሳቢ.ስም><ተሳቢ.ቅጽል>
29	<ባለቤት><ንኡስ.አግባብ.ዝርዝር><ነባር.ማሰ.ዐን><ተጠቃሽ>
30	<መድበል.አሐዝ.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>

4. Group Five POS Patterns

The last group contains sentences made up of five Geez words and five POS tags. Under this category there are 10 POS patterns. For example the sentence ቅቱለ<ሳድስ.ውስ.ቅጽል> ቃኤል<ዘርፍ> አቤል<ስም> ኮነ<ማሰ.ዐን> ሰማዕተ<ተሳቢ> has POS pattern <ሳድስ.ውስ.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>. Table 4.4 lists POS patterns of this group.

Table 4.4*: Group Five Sentence POS Patterns*

No.	POS Patterns
1	<ሳድስ.ውስ.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>
2	<ሣልስ.ው.ቅጽል><ተሳቢ><ስም><ማሰ.ዐን><ተሳቢ>
3	<ስም><ተሳቢ><ስም><ማሰ.ዐን><ተሳቢ>
4	<ስም><ሣልስ.ው.ቅጽል><ተሳቢ><ማሰ.ዐን><ማሰ.ዐን>
5	<መስም.ው.ቅጽል><ተሳቢ><ስም><ማሰ.ዐን><ተሳቢ>
6	<ዘርእ.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>
7	<መድበል.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>
8	<ነባር.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>

No.	POS Patterns
9	<ስም><ካባር.ው.ቅጽል><ዘርፍ><ማሰ.ዐን><ተሳቢ>
10	<ውስጠ.ዘ.ቅ><ስም><ግልጽ.ዘ.ዐንቀጽ><ማሰ.ዐን><ተሳቢ>

4.2.3 Reordering Rules

Owing to the difference in the rules of sentence formation between Geez and Amharic, a reordering rule must be applied on the Geez corpus being processed to align the right source word (Geez) to the right target word (Amharic). As explained in section 4.2.2 for ease of implementation POS Patterns are grouped into four categories. In the same manner this subcomponent implements corresponding reordering rules for POS Patterns belonging to Group Two to Group Five. For this study 126 reordering rules that are taken from [25] that are also checked by a person with Geez knowledge have been implemented by this subcomponent.

1. Group Two Reordering Rules

This Group contains corresponding reordering rules for Group Two POS Patterns. There are 22 POS based patterns and corresponding 22 reordering rules. For example the sentence ንግበር<ማሰ.ዐን> ሰብአ<ስመ.ባሕርይ> with pattern <ማሰ.ዐን><ስመ.ባሕርይ> will have a reordering rule <ስመ.ባሕርይ><ማሰ.ዐን> since the corresponding Amharic sentence is ሰውን<ስመ.ባሕርይ> ንፍጠር<ማሰ.ዐን>. Table 4.5 lists all reordering rules applied.

Table 4.5*: Group Two Reordering Rules*

No.	POS Patterns	Reordering rules
1	<ባለቤት> <ተሳቢ.ስም>	<ባለቤት> <ተሳቢ.ስም>
2	<ማሰ.ዐን> <ስመ.ባሕርይ>	<ስመ.ባሕርይ><ማሰ.ዐን>
3	<ስመ.ባሕርይ><ማሰ.ዐን>	<ስመ.ባሕርይ><ማሰ.ዐን>
4	<ስም><ማሰ.ዐን>	<ስም><ማሰ.ዐን>
5	<ማሰ.ዐን><ስም>	<ስም><ማሰ.ዐን>
6	<ስመ.ተጸውዖ><ማሰ.ዐን>	<ስመ.ተጸውዖ><ማሰ.ዐን>
7	<መድበል.ስመ.ተጸውዖ><ማሰ.ዐን>	<መድበል.ስመ.ተጸውዖ><ማሰ.ዐን>
8	<ባለቤት><ማሰ.ዐን>	<ባለቤት><ማሰ.ዐን>

No.	POS Patterns	Reordering rules
9	<ግልጽ.ዘ.ቅጽል><ስም>	<ግልጽ.ዘ.ቅጽል><ስም>
10	<ንኡስ.አንቀጽ><ስም>	<ንኡስ.አንቀጽ><ስም>
11	<በቂ.ውስጠ.ዘ><ማሰ.ዐን>	<በቂ.ውስጠ.ዘ><ማሰ.ዐን>
12	<አድራጊ.ስም><ግስ>	<አድራጊ.ስም><ግስ>
13	<ግስ> <አድራጊ.ስም>	<አድራጊ.ስም><ግስ>
14	<ተደራጊ.ስም><ተገብሮ.ግስ>	<ተደራጊ.ስም><ተገብሮ.ግስ>
15	<ተደራጊ.ስም><ተገብሮ.ስም>	<ተደራጊ.ስም><ተገብሮ.ስም>
16	<ተገብሮ.ግስ><ተደራጊ.ስም>	<ተደራጊ.ስም><ተገብሮ.ግስ>
17	<ባለቤት><ተገብሮ.ግስ>	<ባለቤት><ተገብሮ.ግስ>
18	<ተገብሮ.ግስ><ባለቤት>	<ባለቤት><ተገብሮ.ግስ>
19	<ተገብሮ.ግስ><ተሳቢ>	<ተገብሮ.ግስ><ተሳቢ>
20	<ባለቤት><ግስ>	<ባለቤት><ግስ>
21	<ግስ><ባለቤት>	<ባለቤት><ግስ>
22	<ግስ><ስም>	<ስም><ግስ>

2. Group Three Reordering Rules

Reordering rules applied to POS Patterns of Group Three fall under this category. There are 59 POS patterns and corresponding 59 reordering rules in group three. Consider the sentence እግዚአብሔር<ባለቤት> ኢይሜንን<ማሰ.ዐን> ፍጥረተ<ስመ.ተጸውዖ> with POS pattern <ባለቤት><ማሰ.ዐን><ስመ.ተጸውዖ> of length three having corresponding reordering rule written as <ባለቤት><ስመ.ተጸውዖ><ማሰ.ዐን> because the Amharic translation is እግዚአብሔር<ባለቤት> ፍጥረትን<ስመ.ተጸውዖ> አይንቅም<ማሰ.ዐን> which follows POS pattern <ባለቤት><ስመ.ተጸውዖ><ማሰ.ዐን>. Table 4.6 lists all reordering rules of this group.

Table 4.6*: Group Three Reordering Rules*

No.	POS Patterns	Reordering rules
1	<ባለቤት> <ግስ> <ተሳቢ.ስም>	<ባለቤት> <ተሳቢ.ስም> <ግስ>
2	<ባለቤት> <ግስ> <ተሳቢ.መራህያን>	<ባለቤት><ተሳቢ.መራህያን><ግስ>
3	<ግስ> <ባለቤት> <ተሳቢ>	<ባለቤት> <ተሳቢ><ግስ>

No.	POS Patterns	Reordering rules
4	<መራሕያን> <ባለቤት> <ግስ>	<መራሕያን> <ባለቤት> <ግስ>
5	<ማሰ.ዐን> <ስመ.ባሕርይ> <ተሳቢ>	<ስመ.ባሕርይ> <ተሳቢ> <ማሰ.ዐን>
6	<ስመ.ባሕርይ> <ማሰ.ዐን> <ተሳቢ>	<ስመ.ባሕርይ><ተሳቢ><ማሰ.ዐን>
7	<ተሳቢ> <ማሰ.ዐን> <ስመ.ባሕርይ>	<ስመ.ባሕርይ><ተሳቢ><ማሰ.ዐን>
8	<ማሰ.ዐን> <ስመ.ባሕርይ> <ማሰ.ዐን>	<ስመ.ባሕርይ><ማሰ.ዐን><ማሰ.ዐን>
9	<ማሰ.ዐን> <ማሰ.ዐን> <ስመ.ባሕርይ>	<ስመ.ባሕርይ><ማሰ.ዐን><ማሰ.ዐን>
10	<ማሰ.ዐን> <ተሳቢ> <ስመ.ባሕርይ>	<ስመ.ባሕርይ><ማሰ.ዐን> <ተሳቢ>
11	<ባለቤት> <ማሰ.ዐን> <ስመ.ባሕርይ>	<ባለቤት><ስመ.ባሕርይ><ማሰ.ዐን>
12	<ማሰ.ዐን> <ባለቤት> <ስመ.ባሕርይ>	<ባለቤት> <ስመ.ባሕርይ><ማሰ.ዐን>
13	<ስመ.ባሕርይ><ማሰ.ዐን><ባለቤት>	<ባለቤት><ስመ.ባሕርይ><ማሰ.ዐን>
14	<ስመተጸውዖ><ማሰ.ዐን><ተሳቢ>	<ስመተጸውዖ><ተሳቢ><ማሰ.ዐን>
15	<ማሰ.ዐን><ባለቤት><ተሳቢ>	<ባለቤት><ተሳቢ><ማሰ.ዐን>
16	<ተሳቢ><ማሰ.ዐን><ስመተጸውዖ>	<ስመተጸውዖ><ተሳቢ><ማሰ.ዐን>
17	<ስመተጸውዖ><ማሰ.ዐን> <ተሳቢ>	<ስመተጸውዖ><ተሳቢ><ማሰ.ዐን>
18	<ስመ.ተጸውዖ><ማሰ.ዐን> <ተሳቢ>	<ስመተጸውዖ><ተሳቢ><ማሰ.ዐን>
19	<ማሰ.ዐን><ስመተጸውዖ><ተሳቢ>	<ስመተጸውዖ><ተሳቢ><ማሰ.ዐን>
20	<ባለቤት><ማሰ.ዐን><ተሳቢ>	<ባለቤት><ተሳቢ><ማሰ.ዐን>
21	<ማሰ.ዐን><ስም><ተሳቢ>	<ስም><ተሳቢ><ማሰ.ዐን>
22	<ተሳቢ><ማሰ.ዐን><ባለቤት>	<ባለቤት><ተሳቢ><ማሰ.ዐን>
23	<ስም><ማሰ.ዐን><ተሳቢ>	<ስም><ተሳቢ><ማሰ.ዐን>
24	<ተሳቢ><ማሰ.ዐን><ስም>	<ስም><ተሳቢ><ማሰ.ዐን>
25	<መድበል.ስመ.ተጸውዖ><ማሰ.ዐን><ተሳቢ>	<መድበል.ስመ.ተጸውዖ><ተሳቢ><ማሰ.ዐን>
26	<ባለቤት><ማሰ.ዐን><ተሳቢ.ስመ.ተጸውዖ>	<ባለቤት><ተሳቢ.ስመ.ተጸውዖ><ማሰ.ዐን>
27	<ማሰ.ዐን><ባለቤት><ተሳቢ.ስመ.ተጸውዖ>	<ባለቤት><ተሳቢ.ስመ.ተጸውዖ><ማሰ.ዐን>
28	<ተሳቢ.ስመ.ተጸውዖ><ማሰ.ዐን><ባለቤት>	<ባለቤት><ተሳቢ.ስመ.ተጸውዖ><ማሰ.ዐን>
29	<ባለቤት><ማሰ.ዐን><ገቢር.ስመ.ተጸውዖ>	<ባለቤት><ገቢር.ስመ.ተጸውዖ><ማሰ.ዐን>
30	<ገቢር.ስመ.ተጸውዖ><ማሰ.ዐን><ባለቤት>	<ባለቤት><ገቢር.ስመ.ተጸውዖ><ማሰ.ዐን>
31	<ባለቤት><ማሰ.ዐን><ስመ.ተጸውዖ>	<ባለቤት><ስመ.ተጸውዖ><ማሰ.ዐን>
32	<ሳድስ.ውስ.ቅጽል><ስም><ማሰ.ዐን>	<ሳድስ.ውስ.ቅጽል><ስም><ማሰ.ዐን>

No.	POS Patterns	Reordering rules
33	<ስም><ሳድስ.ውስ.ቅጽል><ማሰ.ዐን>	<ሳድስ.ውስ.ቅጽል><ስም><ማሰ.ዐን>
34	<መድብል.አሐዝ.ቅጽል><ስም><ማሰ.ዐን>	<መድብል.አሐዝ.ቅጽል><ስም><ማሰ.ዐን>
35	<ስም><መድብል.አሐዝ.ቅጽል><ማሰ.ዐን>	<መድብል.አሐዝ.ቅጽል><ስም><ማሰ.ዐን>
36	<ደቂቅ.ቅጽል><ስም><ማሰ.ዐን>	<ደቂቅ.ቅጽል><ስም><ማሰ.ዐን>
37	<ስም><ዐንቀጽ><ማሰ.ዐን>	<ዐንቀጽ><ስም><ማሰ.ዐን>
38	<ስም><ዐንቀጽ><ተሳቢ>	<ዐንቀጽ><ስም><ተሳቢ>
39	<ግልጽ.ዘ.ዐንቀጽ><ስም><ማሰ.ዐን>	<ግልጽ.ዘ.ዐንቀጽ><ስም><ማሰ.ዐን>
40	<ቅጽል><ስም><ማሰ.ዐን>	<ቅጽል><ስም><ማሰ.ዐን>
41	<በቂ.ውስጠ.ዘ><ማሰ.ዐን><ተሳቢ>	<በቂ.ውስጠ.ዘ><ተሳቢ><ማሰ.ዐን>
42	<ማሰ.ዐን><በቂ.ውስጠ.ዘ><ስም>	<በቂ.ውስጠ.ዘ><ማሰ.ዐን><ስም>
43	<በቂ.ውስጠ.ዘ><ተሳቢ><ማሰ.ዐን>	<ተሳቢ><በቂ.ውስጠ.ዘ><ማሰ.ዐን>
44	<አድራጊ.ስም><ግስ><ተደራጊ.ስም>	<አድራጊ.ስም><ተደራጊ.ስም><ግስ>
45	<ግስ><አድራጊ.ስም><ተደራጊ.ስም>	<አድራጊ.ስም><ተደራጊ.ስም><ግስ>
46	<አድራጊ.ስም><ገቢር.ግስ><ተደራጊ.ስም>	<አድራጊ.ስም><ተደራጊ.ስም><ገቢር.ግስ>
47	<ገቢር.ግስ><አድራጊ.ስም><ተደራጊ.ስም>	<አድራጊ.ስም><ተደራጊ.ስም><ገቢር.ግስ>
48	<ባለቤት><ገቢር.ግስ><ተሳቢ>	<ባለቤት><ተሳቢ><ገቢር.ግስ>
49	<ገቢር.ስም><ባለቤት><ተሳቢ>	<ባለቤት><ተሳቢ><ገቢር.ስም>
50	<ባለቤት><ግስ><ተሳቢ>	<ባለቤት><ተሳቢ><ግስ>
51	<ግስ><ባለቤት><ተሳቢ>	<ባለቤት><ግስ><ተሳቢ>
52	<ባለቤት><ነባር.ዐንቀጽ><ተሳቢ>	<ባለቤት><ነባር.ዐንቀጽ><ተሳቢ>
53	<ስም><ግስ><ስምና.ዝርዝር>	<ስም><ስምና.ዝርዝር><ግስ>
54	<ባለቤት><ንኡስ.አግባብ.ዝርዝር><ተጠቃሽ>	<ባለቤት><ተጠቃሽ><ንኡስ.አግባብ.ዝርዝር>
55	<ንኡስ.አግባብ><ነባር.ማሰ.አንቀጽ><ባለቤት>	<ባለቤት><ንኡስ.አግባብ><ነባር.ማሰ.አንቀጽ>
56	<ባለቤት><ግስ><ደቂቅ.አግባብ>	<ባለቤት><ደቂቅ.አግባብ><ግስ>
57	<አርእስተ.ግስ><ባለቤት><ተሳቢ>	<ባለቤት><ተሳቢ><አርእስተ.ግስ>
58	<ባለቤት><አስራውና.አናቅጽ><ተሳቢ>	<ባለቤት><ተሳቢ><አስራውና.አናቅጽ>
59	<ባለቤት><አስራውና.ግስ><ተሳቢ>	<ባለቤት><ተሳቢ><አስራውና.ግስ>

3. Group Four Reordering Rules

The fourth group of reordering rules is applied on sentences made up of four words with POS Patterns belonging to Group Four. There are 30 reordering rules in this group. For example, ኒቆዲሞስ<ስም> ድሑን<ሳድስ.ውስ.ቅጽል> ሰብሐ<ማሰ.ዐን> ፈጣሪሁ<ተሳቢ> has a POS pattern <ስም><ሳድስ.ውስ.ቅጽል><ማሰ.ዐን><ተሳቢ> and reordering rule <ሳድስ.ውስ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን> because the Amharic translation is የዳነ<ሳድስ.ውስ.ቅጽል> ኒቆዲሞስ<ስም> ፈጣሪውን<ተሳቢ> አመሰገነ<ማሰ.ዐን> which follows <ሳድስ.ውስ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን> POS pattern. Table 4.7 lists reordering rules applied to sentences with POS Patterns belonging to group four.

Table 4.7*: Group Four Reordering Rules*

No.	POS Patterns	Reordering rules
1	<መራሕያን> <ባለቤት> <ግስ> <ተሳቢ.ስም>	<መራሕያን><ባለቤት><ተሳቢ.ስም><ግስ>
2	<ማሰሪያ.አንቀጽ> <ባለቤት> <ተሳቢ.መራሕያን> <ተሳቢ>	<ባለቤት> <ተሳቢ.መራሕያን><ተሳቢ><ማሰሪያ.አንቀጽ>
3	<ሳድስ.ውስ.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<ሳድስ.ውስ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
4	<ስም><ሳድስ.ውስ.ቅጽል><ማሰ.ዐን><ተሳቢ>	<ሳድስ.ውስ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
5	<ሳድስ.ውስ.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን>	<ዘርፍ><ሳድስ.ውስ.ቅጽል><ስም><ማሰ.ዐን>
6	<ሣልስ.ው.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<ሣልስ.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
7	<ስም><ሣልስ.ው.ቅጽል><ማሰ.ዐን><ተሳቢ>	<ሣልስ.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
8	<መስም.ው.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<መስም.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
9	<አሐዝ.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<አሐዝ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
10	<ስም><መድብል.አሐዝ.ቅጽል><ማሰ.ዐን><ተሳቢ>	<ስም><መድብል.አሐዝ.ቅጽል><ተሳቢ><ማሰ.ዐን>
11	<ዘርእ.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን>	<ዘርፍ><ዘርእ.ው.ቅጽል><ስም><ማሰ.ዐን>
12	<መድበል.ው.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<መድበል.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
13	<ስም><መተርጉም.ው.ቅጽል><ማሰ.ዐን><ተሳቢ>	<መተርጉም.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
14	<ንኡስቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<ንኡስቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
15	<ደቂቅ.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<ደቂቅ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
16	<ስም><ዐንቀጽ><ማሰ.ዐን><ተሳቢ>	<ዐንቀጽ><ስም><ተሳቢ><ማሰ.ዐን>
17	<ግልጽ.ዘ.ዐንቀጽ><ስም><ማሰ.ዐን><ተሳቢ>	<ግልጽ.ዘ.ዐንቀጽ><ስም><ተሳቢ><ማሰ.ዐን>
18	<ውስጠ.ዘ.ቅ><ስም><ግልጽ.ዘ.ዐንቀጽ><ማሰ.ዐን>	<ስም><ግልጽ.ዘ.ዐንቀጽ><ውስጠ.ዘ.ቅ><ማሰ.ዐን>

No.	POS Patterns	Reordering rules
19	<ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
20	<ስም><ቅጽል><ማሰ.ዐን><ተሳቢ>	<ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
21	<ማሰ.ዐን><በቂ.ውስጠ.ዘ><ስም><ተሳቢ>	<በቂ.ውስጠ.ዘ><ተሳቢ><ስም><ማሰ.ዐን>
22	<በቂ.ውስጠ.ዘ><ተሳቢ><ማሰ.ዐን><ተሳቢ>	<ተሳቢ><በቂ.ውስጠ.ዘ><ማሰ.ዐን><ተሳቢ>
23	<ውስጠ.ዘ.ስም><ውስጠ.ቅጽል><ማሰ.ዐን><ተሳቢ>	<ውስጠ.ቅጽል><ውስጠ.ዘ.ስም><ተሳቢ><ማሰ.ዐን>
24	<ውስጠ.ቅጽል> <ውስጠ.ዘ.ስም> <ማሰ.ዐን> <ተሳቢ>	<ውስጠ.ቅጽል><ውስጠ.ዘ.ስም><ተሳቢ><ማሰ.ዐን>
25	<ውስጠ.ዘ.ስም><ግልጽ.ዘ.ቅጽል><አንቀጽ><ማሰ.ዐን>	<ግልጽ.ዘ.ቅጽል><አንቀጽ><ውስጠ.ዘ.ስም><ማሰ.ዐን>
26	<ግልጽ.ዘ.ቅጽል><አንቀጽ><ውስጠ.ዘ.ስም><ማሰ.ዐን>	<ግልጽ.ዘ.ቅጽል><አንቀጽ><ውስጠ.ዘ.ስም><ማሰ.ዐን>
27	<ተሳቢ.ቅጽል><ተሳቢ.ስም><ማሰ.ዐን><ባለቤት>	<ባለቤት><ተሳቢ.ቅጽል><ተሳቢ.ስም><ማሰ.ዐን>
28	<ባለቤት><ማሰ.ዐን><ተሳቢ.ስም><ተሳቢ.ቅጽል>	<ባለቤት><ተሳቢ.ቅጽል><ተሳቢ.ስም><ማሰ.ዐን>
29	<ባለቤት><ንኡስ.አግባብ.ዝርዝር><ነባር.ማሰ.ዐን><ተጠቃሽ>	<ባለቤት><ተጠቃሽ><ንኡስ.አግባብ.ዝርዝር><ነባር.ማሰ.ዐን>
30	<መድብል.አሐዝ.ቅጽል><ስም><ማሰ.ዐን><ተሳቢ>	<ስም><መድብል.አሐዝ.ቅጽል><ተሳቢ><ማሰ.ዐን>

4. Group Five Reordering Rules

The last group contains reordering rules for POS Patterns in Group Five. Under this category there are 10 reordering rules. For example the sentence ቅቱለ<ሳድስ.ውስ.ቅጽል> ቃኤል<ዘርፍ> አቤል<ስም> ኮነ<ማሰ.ዐን> ሰማዕተ<ተሳቢ> has POS pattern <ሳድስ.ውስ.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ> a reordering rule <ዘርፍ><ሳድስ.ውስ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን> because the Amharic translation sentence ቃኤል <ዘርፍ> የገደለው<ሳድስ.ውስ.ቅጽል> አቤል<ስም> ሰማዕተ<ተሳቢ> ሆነ<ማሰ.ዐን> follows <ዘርፍ><ሳድስ.ውስ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን> POS pattern. Table 4.8 lists reordering rules for this group.

Table 4.8*: Group Five Reordering Rules*

No.	POS Patterns	Reordering rules
1	<ሳድስ.ውስ.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>	<ዘርፍ><ሳድስ.ውስ.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
2	<ሣልስ.ው.ቅጽል><ተሳቢ><ስም><ማሰ.ዐን><ተሳቢ>	<ተሳቢ><ሣልስ.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
3	<ስም><ተሳቢ><ስም><ማሰ.ዐን><ተሳቢ>	<ስም> <ተሳቢ><ስም><ተሳቢ><ማሰ.ዐን>
4	<ስም><ሣልስ.ው.ቅጽል><ተሳቢ><ማሰ.ዐን><ማሰ.ዐን>	<ተሳቢ><ሣልስ.ው.ቅጽል><ስም><ማሰ.ዐን><ማሰ.ዐን>
5	<መስም.ው.ቅጽል><ተሳቢ><ስም><ማሰ.ዐን><ተሳቢ>	<ተሳቢ><መስም.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
6	<ዘርእ.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>	<ዘርፍ><ዘርእ.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
7	<መድበል.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>	<ዘርፍ><መድበል.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
8	<ነባር.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>	<ዘርፍ><ነባር.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
9	<ስም><ነባር.ው.ቅጽል><ዘርፍ><ማሰ.ዐን><ተሳቢ>	<ዘርፍ><ነባር.ው.ቅጽል><ስም><ተሳቢ><ማሰ.ዐን>
10	<ውስጠ.ዘ.ቅ><ስም><ግልጽ.ዘ.ዐንቀጽ><ማሰ.ዐን><ተሳቢ>	<ግልጽ.ዘ.ዐንቀጽ><ውስጠ.ዘ.ቅ><ስም><ተሳቢ><ማሰ.ዐን>

4.2.4 Corpus Preprocessing

The subcomponent of the Rule Based Geez Corpus Preprocessor implements set of activities that takes sentences from POS Tagged Geez Corpus and determines the sentence's pattern from POS Patterns subcomponent subsequently applies the appropriate reordering rule from the Reordering Rules subcomponent and finally writes the reordered Geez sentence without POS information onto the Reordered Geez Corpus.

It first reads all sentences from POS Tagged Geez Corpus and stores all sentences in Array variable called Sentence_Array[] it then declares a variable called Array_Index to be used as a counter that holds the current value of the Array index. Its value is set to zero initially. Then checks if Array_Index is less than that of the total number of sentences initially loaded into Sentence_Array[] to avoid index out of bound error. If Array_Index is greater than or equal to that of total number of sentences in Sentence_Array[] the processing stops indicating that it has finished preprocessing the corpus. If it Array_Index is less than that of the total number of sentences in Sentence_Array[] it means that there are still unprocessed sentences.

Then the sentence at the position of the current counter, i.e., Array_Index is retrieved and the number of words will be counted to determine the group in which the sentence belongs in as described in Section 4.2.2. It then it determines POS Pattern and subsequently determine the corresponding reordering rule and apply it on the sentence being processed. Then remove POS information from the sentence and writes it to the Reordered Geez Corpus and increments Array_Index by one and the processing continues until all sentences have been processed.

Figure 4.2 depicts the flowchart of this subcomponent; activities performed by this subcomponent are indicated with a shaded background while those with white background are subcomponents of the Rule Based Geez Corpus Preprocessor not the Corpus Preprocessing subcomponent. They are included in the diagram to merely indicate that the Corpus Preprocessing interacts with them in the course of performing its set of activities.

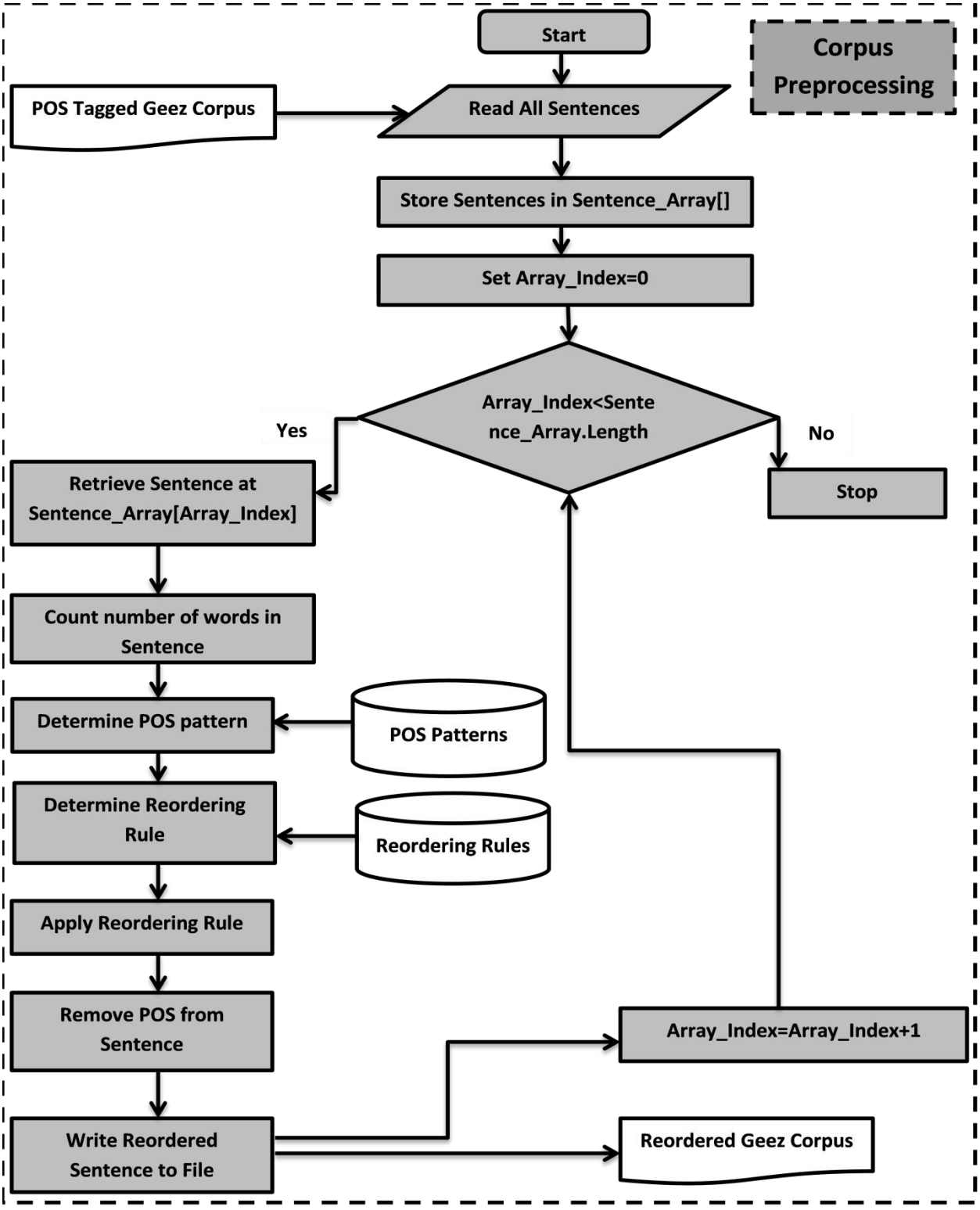

Figure 4.2: *Flowchart for Corpus Preprocessing Subcomponent*

For the sake of simplicity the entire algotihm that implements Coprus Preprocessing component is broken down into five pseudocode segements. Algorithm 4.1 is the implementation of the core function of this subcomponent that accepts POSTaggedCorpus as input reads all senetnces from the input and call respective functions that perform POS pattern and reordering rule determination in the four sentence Groups from GroupTwo to GroupFour and finally write the reordered sentence to the ReorderedGeezCoprus.

```
Input: POSTaggedGeezCorpus
Output: ReorderedGeezCorpus
IF IS_TEXT_FILE(POSTaggedGeezCorpus)//check if input file is text file
      OPEN_FOR_READING(POSTaggedGeezCorpus)
      IF CONTAINS_TEXT(POSTaggedGeezCorpus)//check for text
          Sentence_Array[]=LOAD_SENTENCES(POSTaggedGeezCorpus)
          Array_Index=0
          ReorderedSentence[]//no length specified
          WHILE Array_Index<LENGTH_OF(Sentence_Array[])
           S= Sentence_Array[Array_Index]
           Number_of_Words=COUNT_WORDS(S)
           IF Number_of_Words=1//GroupOne
               REMOVE_POS(S)
               ReorderedSentence[]=S
           ELSE IF Number_of_Words=2//GroupTwo
              ReorderedSentence[]=GROUPTWO(S)
           ELSE IF Number_of_Words=3//GroupThree
              ReorderedSentence[]=GROUPTHREE(S)
           ELSE IF Number_of_Words=4//GroupFour
              ReorderedSentence[]=GROUPFOUR(S)
           ELSE IF Number_of_Words=5//GroupFive
              ReorderedSentence[]=GROUPFIVE(S)
           WRITE_TEXT_TO_FILE(ReorderedSentence[],
           ReorderedGeezCorpus)
           Array_Index=Array_Index+1//increment array index
          END WHILE
    END IF//contains text
END IF//input file is text file
```

***Algorithm 4.1**: Algorithm for the Core Function of Corpus Preprocessing Component*

The implementation of Group Two POS patterns and Reordering rules is presentend in Algorithm 4.2.

```
Input: Input Sentence S to be processed

Output: Reordered sentence in array RS[]

FUNCTION GROUPTWO(S)

    RS[].Length=3

    IF POS_PATTERN(S) =<ግሰ.ዐን><ስም.ባሕርይ>//Check_pattern

         RS[0]=S[1]

         RS[1]=" "

         RS[2]=S[0]

     ELSE IF POS_PATTERN(S) =<ግሰ.ዐን><ስም>//Check for another pattern

         RS[0]=S[1]
         RS[1]=" "
         RS[2]=S[0]
    // Check for the 22 POS patterns in this Group using ELSE IF
Condition

    return RS[]

END FUNCTION
```

***Algorithm 4.2**: Algorithm for Group Two POS Patterns and Reordering Rules*

The implementation of Group Three POS patterns and Reordering rules is presentend in Algorithm 4.3.

```
Input: Input Sentence S to be processed

Output: Reordered sentence in array RS[]

FUNCTION GROUPTHREE(S)

    RS[].Length=5

    IF POS_PATTERN(S) =<ባለቤት> <ግስ> <ተሳቢ.ስም>//Check_pattern

         RS[0]=S[0]
         RS[1]=" "
         RS[2]=S[2]
         RS[3]=" "
         RS[4]=S[1]

   ELSE IF POS_PATTERN(S) =<ማስ.ዐን> <ተሳቢ> <ስመ.ባሕርይ> //Check for
another pattern

         RS[0]=S[2]
         RS[1]=" "
         RS[2]=S[0]
         RS[3]=" "
         RS[4]=S[1]

    // Check for the 59 POS patterns in this Group using ELSE IF
Condition

    return RS[]

END FUNCTION
```

***Algorithm 4.3**: Algorithm for Group Three POS Patterns and Reordering Rules*

The implementation of Group Four POS patterns and Reordering rules is presentend in Algorithm 4.4.

```
Input: Input Sentence S to be processed

Output: Reordered sentence in array RS[]

FUNCTION GROUPFOUR(S)

    RS[].Length=7

    IF POS_PATTERN(S) =<ማሰሪያ.አንቀጽ> <ባለቤት> <ተሳቢ.መራሕያን>
<ተሳቢ>//Check_pattern

          RS[0]=S[1]
          RS[1]=" "
          RS[2] =S[2]
          RS[3]=" "
          RS[4]=S[3]
          RS[5]=""
          RS[6]=S[0]
    ELSE IF POS_PATTERN(S) =<ማሰሪያ.አንቀጽ> <ባለቤት> <ተሳቢ.መራሕያን> <ተሳቢ>//Check
for another pattern

          RS[0]=S[1]
          RS[1]=" "
          RS[2] =S[2]
          RS[3]=" "
          RS[4]=S[3]
          RS[5]=""
          RS[6]=S[0]

    // Check for the 30 POS patterns in this Group using ELSE IF
Condition

    return RS[]

END FUNCTION
```

***Algorithm 4.4**: Algorithm for Group Four POS Patterns and Reordering Rules*

The implementation of Group Five POS patterns and Reordering rules is presentend in Algorithm 4.5.

```
Input: Input Sentence S to be processed

Output: Reordered sentence in array RS[]

FUNCTION GROUPFIVE(S)

    RS[].Length=9

    IF POS_PATTERN(S) =<ሳድስ.ውስ.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>//Check_pattern

         RS[0]=S[1]
         RS[1]=" "
         RS[2]=S[0]
         RS[3]=" "
         RS[4]=S[2]
         RS[5]=" "
         RS[6]=S[4]
         RS[7]=" "
         RS[8]=S[3]
    ELSE IF POS_PATTERN(S) =<ዘርእ.ው.ቅጽል><ዘርፍ><ስም><ማሰ.ዐን><ተሳቢ>//Check
for another pattern
         RS[0]=S[1]
         RS[1]=" "
         RS[2]=S[0]
         RS[3]=" "
         RS[4]=S[2]
         RS[5]=" "
         RS[6]=S[4]
         RS[7]=" "
         RS[8]=S[3]

    //Check for the 10 POS patterns in this Group using ELSE IF
Condition

    return RS[]

END FUNCTION
```

***Algorithm 4.5**: Algorithm for Group Five POS Patterns and Reordering Rules*

4.2.5 Reordered Geez Corpus

After each sentence in the input POS Tagged Geez Corpus is reordered to make the word ordering similar to that of Amharic sentence by Corpus Preprocessing, this corpus holds the reordered Geez sentence without POS information. Then it will be used in training as well as translation process by the Baseline SMT. Table 4.9 shows an excerpt from Geez corpus before and after reordering.

***Table 4.9**: Geez Sentence Before and After Reordering*

Before Reordering	**After Reordering**
አነ<ባለቤት> ዐጸውኩ<ግስ> ኖኅተ<ተሳቢ.ስም>	አነ ኖኅተ ዐጸውኩ
ንሕነ<ባለቤት> ተወከፍነ<ግስ> ነግደ<ተሳቢ.ስም>	ንሕነ ነግደ ተወከፍነ
አንተ<ባለቤት> አፍቅርከ<ግስ> ዐርከ<ተሳቢ.ስም>	አንተ ዐርከ አፍቅርከ
አንትሙ<ባለቤት> ዘንብክሙ<ግስ> ወርቀ<ተሳቢ.ስም>	አንትሙ ወርቀ ዘንብክሙ
አንቲ<ባለቤት> ተቀባእኪ<ግስ> አፈወ<ተሳቢ.ስም>	አንቲ አፈወ ተቀባእኪ
አንትን<ባለቤት> ሠራዕክን<ግስ> ማዕደ<ተሳቢ.ስም>	አንትን ማዕደ ሠራዕክን
ውእቱ<ባለቤት> አልሐቀ<ግስ> ሕጻነ<ተሳቢ.ስም>	ውእቱ ሕጻነ አልሐቀ
ውእቶሙ<ባለቤት> ሐመዩ<ግስ> ቢጸሙ<ተሳቢ.ስም>	ውእቶሙ ቢጸሙ ሐመዩ
እሙንቱ<ባለቤት> ሐመዩ<ግስ> ቢጸሙ<ተሳቢ.ስም>	እሙንቱ ቢጸሙ ሐመዩ
ይእቲ<ባለቤት> ሐመዩ<ግስ> ቢጸሙ<ተሳቢ.ስም>	ይእቲ ቢጸሙ ሐመዩ
ይእቲ<ባለቤት> ነጸረት<ግስ> ሥና<ተሳቢ.ስም>	ይእቲ ሥና ነጸረት

4.3 Amharic Corpus

Amharic Corpus contains 976 Amharic sentences taken from [25] and each is perfectly aligned with their possible Geez translations. It is directly supplied as an input to the Baseline SMT to be used in the training process and doesn't contain any POS information. The Rule Based Geez Corpus Preprocessor doesn't process this Corpus it's merely used as an input by the Baseline SMT.

4.4 Baseline SMT

This is the second component of Geez to Amharic Machine Translation System that performs the actual translation of Geez (source) sentence-which is the input- into corresponding Amharic (target) sentence-which is the output- using statistical methods. As opposed to a rule based machine translation system which heavily depends on language specific linguistic knowledge,

SMT only needs a huge parallel corpus. SMT is founded upon the assumptions of the Noisy Channel Model and Bayes Rule which helps decompose the complex probabilities model that needs to be built for estimating the probability of a sentence in the source language (G) being translated into a target language (A). To build a SMT system that translates Geez sentence to Amharic sentence what is need is a mapping from Geez Sentence (g) to an Amharic Sentence (a), i.e., $g \rightarrow a$ by building a model $p(a|g)$ that estimates the conditional probability of any Amharic sentence a give Geez sentence g. Using Noisy Channel Model the probabilistic model can be expressed using Equation 1 taken from [44].

$$argmax_a p(a|g) = argmax_a p(a) * p(g|a) \quad (1)$$

The aim of equation 1 is to search for the Amharic Sentence a that maximizes $p(a) * p(g|a)$. By decomposing the conditional probability model $p(a|g)$ the burden of accuracy expected from the model is distributed to two independent probabilities $p(a)$ and $p(g|a)$. The first probability is the language model (for Amharic language). Since the proposed translation system is unidirectional, i.e., only from Geez to Amharic one language model for Amharic is built. The second probability is the translation model for predicting Geez sentences (source sentences) g from target sentences (Amharic sentences) a.

4.4.1 Language Model

Estimation of the likelihood of a sentence is the goal of a statistical machine translation. A language model gives that probability using n-gram model. By using chain rule shown in Equation 2 the probability of a sentence $p(S)$ is broken down to the probability of each word $p(W)$ that make up the sentence given the words that precedes it according to [44]. An n-gram of size one is referred to as a unigram; size two is a bigram; three a trigram and higher order grams in general are referred to as n-gram.

$$p(S) = (W1, W2, W3, \ldots, WN)$$

$$= p(W1)p(W2|W1)p(W3|W1W2)\ldots\ldots p(WN|W1W2..WN-1) \quad (2)$$

Consider the following Amharic sentences S taken from the corpus.

ዳዊት አምላክን አመሰገነ

ሴቶች መልአክን አገኙ

አብርሃም እግዚአብሔርን ፈለገ

እግዚአብሔር መልአክን ይልካል

እግዚአብሔር ሁሉን ይገዛል

እግዚአብሔር ሁሉን

Using the sentence <እግዚአብሔር ሁሉን ይገዛል> the unigram probability is calculated as

$$p(W1) = \frac{\text{count}(W1)}{\text{Total words observed}}$$

p(እግዚአብሔር) = 3 / 17 = 0.176

The bigram probability is calculated as

$$p(W2 \mid W1) = \frac{\text{count}(W1W2)}{\text{count}(W1)}$$

p(ሁሉን | እግዚአብሔር) = 2 / 3 = 0.333

The trigram probability is calculated as

$$p(W3 \mid W1W2) = \frac{\text{count}(W1W2W3)}{\text{count}(W1W2)}$$

p(ይገዛል | እግዚአብሔር ሁሉን) = 1 / 2 = 0.5

Since our Geez to Amharic translation system is unidirectional, i.e., the system is only able to translate Geez text to Amharic text not vice versa, one language model for Amharic is needed.

4.4.2 Translation Model

A translation model assigns a probability that a given source sentence generates a target sentence in the target language which is denoted by $p(a \mid g)$ where g is sentence in source language G

and a is a string in the target language A. The formula of the probability is given in Equation 3 taken from [45].

$$p(a|g) = \text{count}(a,g)/\text{count}(g) \quad (3)$$

Since it is difficult to find enough data in the parallel corpus that satisfy the values for all sentences, the solution is to break sentences in to smaller chunks. These chunks can be phrases or words and this is given by Equation 4 taken from [45].

$$p(g|a) = \sum_h \; p(h,g|a) \quad (4)$$

The variable h represents alignments between the individual chunks in the sentence pair where the chunks in the sentence pair can be words or phrases. In word based translation the chunks are words. In phrase based translation, which is the most commonly used form of translation, the chunks are phrases, and these phrases are not linguistic phrases rather found by statistical methods from the corpora [45]. For this study phrase based translation is used.

4.4.3 Decoding

Decoding is the process of finding target translated sentence (Amharic Sentence a) for a given source sentence (Geez Sentence g) using Language Model $p(a)$ and Translation Model $p(g|a)$. It is a search problem that maximizes the probability of both language and translation models. Using these models it constructs the possible translations and looks for the most probable one.

4.5 Summary

This Chapter presented the development of Geez to Amharic machine translation system. The system uses a hybrid approach by using the serial coupling of a Rule Based Geez Preprocessor component followed by a Baseline SMT. Owing to the syntactic differences between the two languages, 126 reordering rules are applied during preprocessing on the manually POS tagged Geez corpus in order to make the syntactic order of Geez sentence similar to that of Amharic before the translation process by the Baseline SMT. The preprocessing component process the input Geez sentences one by one from POS pattern determination till application of reordering rule and writing the reordered sentence to a file.

Section 4.4 of this Chapter briefly described the components of the Baseline SMT, which is composed of the language model, translation model, and decoder. The language model ensures that every translated string of the target language - Amharic - comes in the right order. Since the translation of the proposed system is unidirectional - Geez to Amharic- no language model was developed for Geez. The translation model assigns the probability that given a Geez sentence generates an Amharic text. The decoder searches for the best translation of a Geez string from a phrase translation table, since phrases are the minimum blocks of translation used in this study, and produces the Amharic meaning.

Chapter Five: System Evaluation and Results

5.1 Introduction

This Chapter discusses the results of the experiment conducted to evaluate the proposed Geez to Amharic hybrid machine translation system. It is divided into five major sections that entails, the tools used for developing the system, what are the major activities conducted during corpus preparation, brief explanation about Bilingual evaluation understudy (BLEU) evaluation metrics, experiments carried out to test the proposed system and finally discussion section that compares BLEU scores obtained by the pure statistical machine translation system with the hybrid machine translation system.

5.2 Tools Used for Development

Hybrid Geez to Amharic Machine Translation System is composed of two major components as explained in Section 4.1 that operate on two different operating system platforms. The Rule Based Geez Corpus Preprocessor operates on Windows operating system while the Baseline SMT operates on Ubuntu Linux operating system. Development tools used for each are explained hereunder.

5.2.1 Tools Used for Rule Based Geez Corpus Preprocessor

The Rule Based Geez Corpus Preprocessor component is a Windows based application that is developed using Microsoft Visual Studio 2013, .Net Framework Version 4.0, and C# Programming language. Since the preprocessed Geez corpus will be used by the Baseline SMT- which operates on another operating system, i.e., Ubuntu Linux- the file must be saved with a format that is compatible with other operating systems as well. Owing to this reason the corpus file is saved with Unicode Transformation Format (UTF-8) format. A character in UTF-8 can be from 1 to 4 bytes long. It can represent any character in the Unicode standard and is backward compatible with ASCII. Currently, it is the preferred encoding for e-mail and web pages [46].

5.2.2 Tools Used for Baseline SMT

The Baseline SMT which performs the actual translation from Geez string to Amharic string is composed of three components: Language Model, Translation Model, and Decoder. This component operates on Ubuntu Linux operating system. Thus the tools used for developing this component are all installed and configured in Ubuntu Linux, each are explained hereunder.

a. VMware 10.0

The computer used to develop the hybrid machine translation system has Windows operating system. However, the tools needed to develop a standard statistical machine translation system are compatible on Linux like operating system. We were presented with two choices to choose from, the first one is to install Cygwin on the existing Windows operating system to simulate Linux like environment. The second one is to install a Virtual machine on the existing Windows operating system then install Linux on top of the Virtual machine. At first we choose Cygwin but since it needs a lot of plugin components that took much time to thus we finally choose the second option. VMware 10.0 is the specific Virtual machine used for this study. Virtual machine software enables users to setup virtual machine on a single physical machine. Each virtual machine can execute its own operating system that is different from the operating system installed on the physical machine.

b. Ubuntu Linux 14.04

The operating system that is compatible with the tools used to develop the Baseline SMT is Ubuntu Linux operating system (version 14.04 of Ubuntu is used for this work) that is installed on top of the VMware. It is an open source operating system based on Debian architecture [47].

c. Moses

The decoder of the Baseline SMT is implemented using Moses. It is an open source statistical machine translation toolkit developed at the University of Edinburgh. Using a parallel corpus which is sentence aligned and a trained model, Moses uses a search algorithm to find the best translation of a source string. The training process in Moses takes in the parallel data and uses co-occurrences of words and phrases to infer translation correspondences between the source and target languages. In phrase-based machine translation, these correspondences are simply between continuous sequences of words. It is composed of various components that are written in C++ and Perl programming languages [48]. Moses is installed using ‘create-1.43’ script file of Moses for Mere Mortals.

d. IRSTLM

The Amharic language model is developed using IRSTLM tool which is compatible with Moses. It contains algorithms and data structures that are used to estimate, store, and access very large n-gram language models [49]. IRSTLM along with other SMT tools is installed when executing create 'create-1.43' script of Moses for Mere Mortals.

e. MGIZA

The translation model of the Baseline SMT is developed using MGIZA which is a multi-threaded word alignment tool based on GIZA. It extends GIZA in multiple ways. It provides the concept of multi-threading and memory optimization. It can resume training from any stage [44].

5.3 Corpus Preparation

Two sets of corpora were used to test the proposed Hybrid Geez to Amharic Machine Translation System and the Baseline SMT.The first set contains Geez and Amharic corpus without having any POS information that will be used as an input to the Baseline SMT without being fed to the hybrid one. The second set contains POS tagged Geez corpus and an Amharic corpus without any POS information. The POS tagged Geez corpus will first be preprocessed via the Rule Based Geez Corpus Preprocessor before being supplied as an input to the Baseline SMT. In either case the Amharic corpus doesn't contain any POS information. On the other hand the Geez corpus that is used to test the Baseline SMT has no POS information. However, the Geez corpus that is used to test the hybrid system has POS information which latter is removed by the Rule Based Geez Corpus Preprocessor after reordering rule is applied on each sentence. Regardless of POS information both sets of corpora contain the same number of sentences and words.

Each corpus was prepared manually as there is no available corpus for these languages that can be used for the purpose of this study. The Geez corpus contains 976 sentences with 3010 words. On the other hand the Amharic corpus contains the same number of sentences with 3174 words. Every sentence in each corpus was entirely taken from [25]. To make the prepared corpuses readable by Moses, UTF 8 format was used while saving the text files.

The following two procedures have been applied on the collected corpus to make it ready for training and testing the translation system. Since both are local languages the process of true-casing is not important for this study.

Tokenization: This is a procedure that inserts a space between words and punctuation.

Cleaning: This procedure helps to remove long sentences and empty sentences as they can cause problems with the training pipeline. This also helps to remove misaligned sentences.

5.4 BLEU Evaluation Metrics

BLEU is the most popular and commonly used precession (that is, it considers the number of *n*-gram matches as a fraction of the number of total *n*-grams in the output sentence) oriented metric for measuring the translation quality of a machine translation system. It considers not only single word matches between the output and the reference sentence, but also *n*-gram matches, up to some maximum *n*. This allows it to reward sentences where local word order is closer to the local word order in the reference. It is the most commonly used form of evaluation in statistical machine translation [1].

Translation quality is the correspondence between a machine translation with that of a human translator. A high quality translation is the one which is closer to a professional human translation and BLEU's main idea is the measurements of this closeness. BLEU score value falls in the range between 0 and 1, the higher the BLEU score (those close to one) the more the translation resembles with the translation of a human translation [45].

5.5 Experiment

This section details the experiments conducted to test the translation quality of the Baseline SMT and the proposed Hybrid Geez to Amharic Machine Translation System. It first discusses the experiment environment that the experiments were conducted. To easily distinguish between the two experiments, Table 5.1 summarizes the characteristics of each.

Table 5.1*: Experiments Characteristics*

	Geez Corpus POS tagged	Amharic Corpus POS tagged	Preprocessing of Geez Corpus	Use of Baseline SMT
Experiment to Test Hybrid Geez to Amharic MT	Yes	No	Yes	Yes
Experiment to Test Baseline SMT	No	No	No	Yes

5.5.1 Experimentation Environment

Both experiments were conducted on HP laptop having Core i5 processor, 8GB RAM with Windows 8 operating system. VMware 10.0 was installed on Windows 8 operating system. On top of the VMware Ubuntu Linux 14.04 was installed on which the Baseline SMT executes on. The Rule Based Geez Corpus Preprocessor operates on the Windows platform.

5.5.2 Experiment to Test Hybrid Geez to Amharic MT

The purpose of testing the proposed hybrid system is to compare the translation quality against that of the Baseline SMT. Both the Rule Based Geez Corpus Preprocessor component which is installed on Windows 8 operating system and Baseline SMT which is installed on Ubuntu 14.04 Linux operating system play role in this test. The steps of the experiment are as follows.

Step 1: Save the POS tagged Geez corpus text file with .txt extension in a UTF-8 format in Windows 8 operating system. Notepad program was used for this purpose.

Step 2: Supply the corpus file to the Rule Based Geez Corpus Preprocessor component by browsing the file for testing using the browse button on the application as shown in Figure 5.1.

Step 3: After the preprocessing is finished copy the processed Geez and Amharic corpus files and paste them under the folder Moses for Mere mortal installation folder ‘MMM’ in Ubuntu Linux operating system’s desktop.

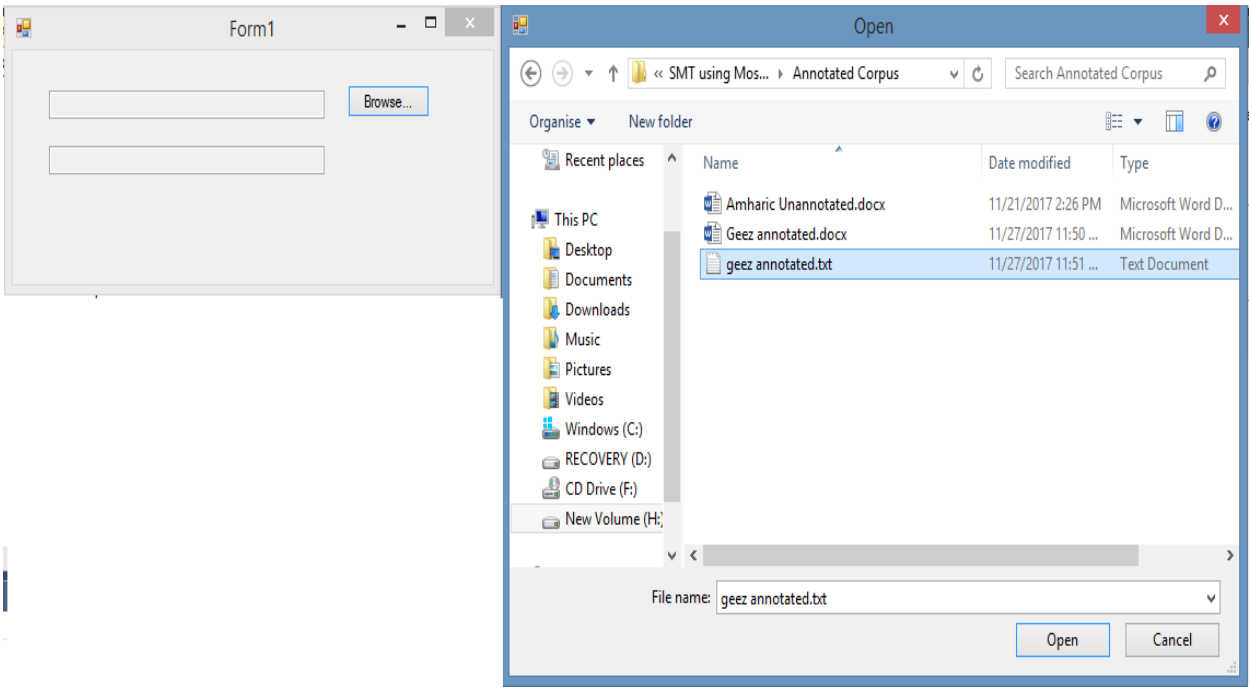

Figure 5.1*: Corpus Preprocessor User Interface*

Step 4: Before starting the actual translation by the Baseline SMT, both corpus files must have an extension that allows Moses to differentiate the languages using two characters that represent the language in which the file is written. However, both Geez and Amharic languages don't have a standard two character language codes. Thus, Portugal with code 'pt' and English with code 'en' were randomly selected as the language extensions for Geez and English respectively. So Geez corpus is named as '976Proposed.pt' and '976Proposed.en' for Amharic corpus. At first we have tried using the correct extensions for each language 'am' for Amharic and 'gez' for Geez but Moses decoder displays error message so we were forced to change the extension to the aforementioned language extensions.

Step 5: Run the script file 'make-test-filestest' under the scripts folder of Moses for Mere Mortals installation folder by opening the terminal window. After changing the base name and source, target language attributes and this will create training and testing files with the names '976Proposed.for_train.pt' and '976Proposed.for_train.en' used for training '976Proposed.for_test.pt' and '976Proposed.for_test.en' for testing.

Step 6: Run 'traintest-1.22' script file in the scripts folder under Moses for Mere Mortals installation folder to start the training process. Figure 5.2 shows screenshot after the training process completion. Sample corpus used to test the proposed system is presented in Appendex I.

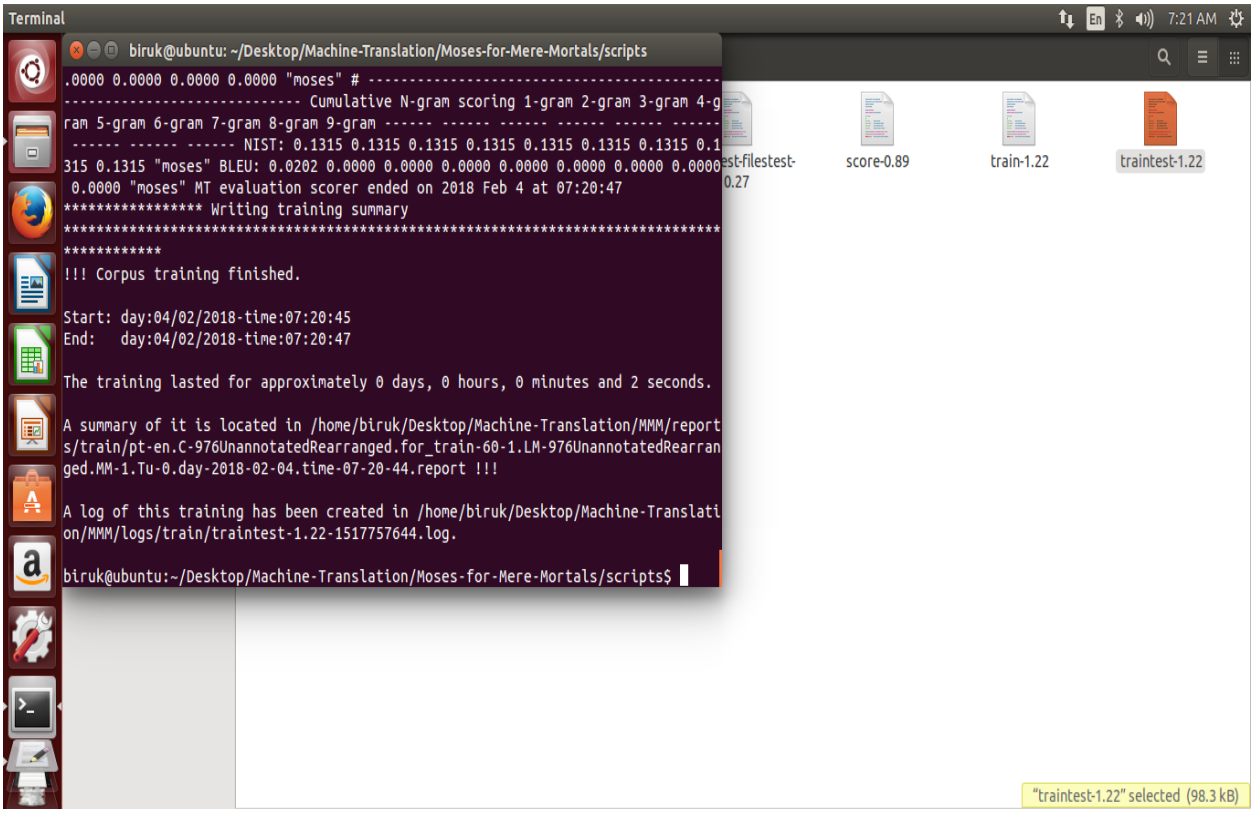

Figure 5. 2: *Training Screenshot*

Step 7: Run 'translatetest-1.38' script file in the scripts folder under Moses for Mere Mortals installation folder to start the translation process. Now the translation output is produced for the test Geez corpus file. Figure 5.3 shows a screenshot after the translate process completion. The translation input file '976Proposed.for_test.pt' that is used to test the system contains Geez sentences that are reordered to make them similar the word ordering of Amharic sentences. A sample of corpus used for testing the proposed solution is presented in Appendix III.

Step 8: Finally scoring of the translation will be done by executing 'score-0.89' script after copying a reference translation output file in folder 'translation_reference' with the name '976Proposed.for_test.en.ref' that holds the Amharic translation of '976Proposed.for_test.pt' by

a human translator. The scorer script compares the translation output '976Proposed.for_test.pt.en.moses' with '976Proposed.for_test.en.ref' and calculates BLEU score. It then creates a score report in folder 'MMM/reports/scorer/' according to the report a BLEU score of 0.7630 was achieved.

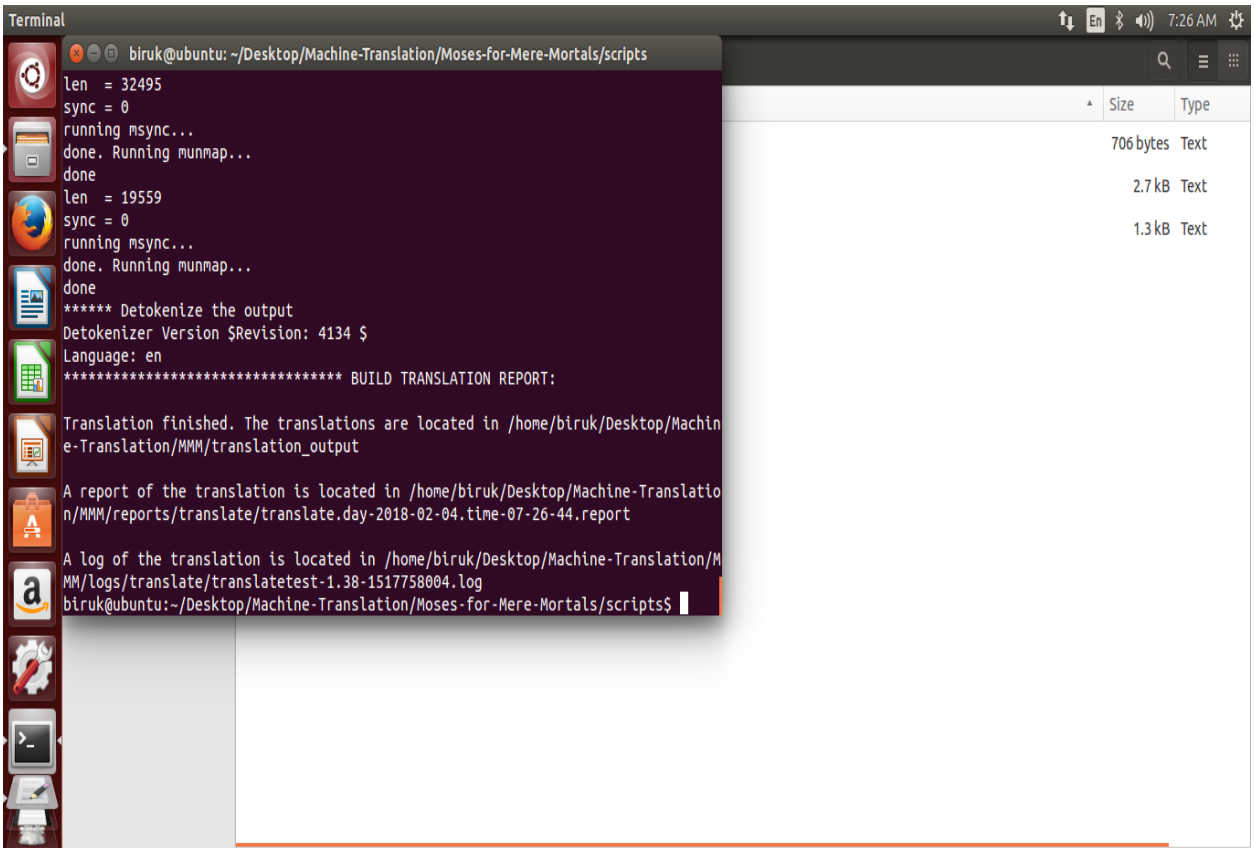

Figure 5.3: *Translation Screenshot*

An excerpt of the translation input and output of the proposed Geez to Amharic Hybrid Machine Transaltion System is presented in Table 5.2.

Table 5.2: *Translation Input and Output*

Translation Input	**Translation output**
እግዚያብሔር ፈጠረ ኪያነ ሕዝቦ	እግዚያብሔር እኛን ወገኖቹን ፈጥሯል
ውእቶን	እነሷ
ሰብእት ተፈጥሩ	ሰዎች ተፈጠሩ
ውእቶን ሜላተ ፈተላ	እነሷ ሀርን ፈትለዋል
ጠቢብት ቤተ የሐንጹ	ጥበበኞች ቤት ይሰራሉ
ሙሴ ባሕረ ከፈለ	ሙሴ ባሕርን ከፈለ

Translation Input	Translation output
ዳዊት አምላክ ሰብሐ	ዳዊት አምላክን አመሰገነ
ሐሜት ሕይወተ ታረኩስ	ሐሜት ሕይወትን ታረክሳለች
ጤሜዎስ ዕውር ብርሃነ ርእየ	ዕውር ጤሜዎስ ብርሃንነ አየ
ዜናዊ ጥበበ ገብርኤል ብስራተ ዜነወ	ጥበብን የሚናገር ገብርኤል ብስራትን አበሰረ
ቃኤል ቅቱለ አቤል ሰማዕተ ኮነ	ቃኤል የገደለው አቤል ሰማዕተ ሆነ
ማርቆስ ግብፀ ፆደ	ማርቆስ ግብፅን ዞረ
መዓትም ብእሲ ሕጻናተ ቀተለ	ቁጡ ሰው ሕጻናትን ገደለ
ደብር እግረ ቢታንያ አውጽአት ሰርጸ	የተራራ እግር ቢታንያ ቡቃያን አወጣች
ዘጌሳ ሰሎሜወማርያም መልአከ ረከባ	የገሰገሱ ማርያምና ሰሎሜ መልአክን አገኙ
ዮሐንስ ሐዋርያ ወንጌለ ጸሐፈ	ዮሐንስ የተባለ ሐዋርያ ወንጌልን ጻፈ
እግዚአብሔር ዐማፄ ኤሳውሀ ጸልአ	እግዚአብሔር ዓመፀኛው ኤሰሳውን ጠላ
አይሁድ መድሕነ ወልደ ሰቀሉ	አይሁድ አዳኝ ወልድን በቀለ
ለነፈሱ ዘረከባ ለይግድፉ	ነፈሱን ያገኛት ይጠላታል
ዳዊት ዝኰ ብእሴ ቀተለ	ዳዊት ይህን ሰውይ ገደለ
መልአክ ለእንስት ረከባሆን	መልአክ ሴቶችን አገኛቸው
ዕርገተ ዐርገ	ዕርገትን ዐረገ
ክብረ ክብረ	ክብርን ክብርን
አልአዛር ሐይወ	አልአዛር ሕያው ሆነ
ፈጣሪ ብርሃናተ ገብረ	ፈጣሪ ብርሃናትን ፈጠረ
አብርሃምወይስሐቅ ነግደ ተወከፍክሙ	አብርሃምና ይስሐቅ እንግዳን ተቀበላችሁ
ንሕነ ምዕራፈ ንጉበር	እኛ ማረፊያን እንሰራለን

5.5.3 Experiment to Test Baseline SMT

In order to test the Baseline SMT, a corpus file with no annotation was used and no preprocessing on the Geez corpus was performed, i.e., both Geez (976Baseline.pt) and Amharic (976Baseline.en) corpus files were directly supplied to the Baseline SMT system with no prior processing. Sample corpus used to test the Baseline SMT is presented in Appendix II. All the above steps were followed and a BLEU score of 0.7219 was achieved. Both the translation and testing files used in this experiment are not reordered.

5.6 Discussion

Two experiments were conducted to evaluate the translation quality of the proposed Hybrid Geez to Amharic Machine Translation system against the Baseline SMT system. The first experiment

was to test the translation quality of the hybrid machine translation. The second was on the Baseline SMT. The Amharic corpus used in both is without POS information. However, a POS tagged Geez corpus was used while testing the hybrid system. BLEU score was used as the evaluation metrics. Except for the tagging information applied on the Geez corpus the sentences used were identical. The proposed system applies 126 reordering rules to make the order of Geez words similar to that of Amharic in order to improve the translation quality of the Baseline SMT.

While testing the Baseline SMT a BLEU score of 0.7219 in the scale of 0 to 1 was obtained. On the other hand a score of 0.7630 in the scale of 0 to 1 was achieved while testing the hybrid machine translation system. From the BLEU scores obtained the proposed Hybrid Machine Translation system outperforms the Baseline SMT system by 4%. Thus, to improve the translation quality of a SMT system for languages that follow different word reordering such as the case between Geez and Amharic application of reordering rules that makes word order of source sentences similar to that of target sentences plays a big role.

The following points were the challenges faced during experimentation.

- Due to the unavailability of Geez and Amharic corpus at least for a research purpose we were forced to prepare the corpus manually which is very time consuming.
- Unavailability of automated POS tagger for Geez language again forced us to prepare the POS manually.
- The installation of the tools for developing the Baseline SMT requires a lot of time and a high speed Internet connection which forced us to spend more time on the prototype development as compared to the time allotted to complete the overall thesis.

Chapter Six: Conclusion and Future work

The purpose of this thesis is to design and develop a hybrid Geez to Amharic machine translation system. What makes the proposed system a hybrid one is its design; a rule based preprocessing of a manually POS tagged Geez corpus is followed by a baseline statistical machine translation system. Thus, a serial coupling of rule based preprocessing followed by SMT system was followed while designing. The preprocessing applies reordering rules in the input POS tagged Geez corpus in order to make the order of Geez words similar to that of Amharic words.

During the course of the work the main activates performed was corpus preparation. As stated in Section 5.3 two sets of corpora were prepared, the first set that contains Geez and Amharic corpuses with no POS tags. The second set contains Geez corpus with POS information and an Amharic corpus with no POS tags. Both are saved in UTF-8 format to be readable by Moses decoder, development of the Rule Based POS Geez Corpus Preprocessor using Visual Studio 2013, C# programming language, and Dot Net framework. This component runs on Windows operating system. Development of Baseline SMT using Moses for Mere Mortal, IRSTLM for language model, MGIZA for translation model, and Moses for decoding, the Baseline system runs on Ubuntu 14.04 Linux operating system.

To test the translation quality of the developed system, two experiments were conducted. The first experiment tested the translation of the Hybrid Machine Translation System POS tagged Geez corpus and Amharic corpus without POS was used and 76% BLEU score was achieved. Baseline SMT system using Geez and Amharic corpuses without POS information was tested in the second experiment and 72% BLEU score was achieved. The hybrid system achieved a 4% improvement in the BLEU score when compared to the baseline SMT. This resulted from the reordering applied on the POS tagged Geez corpus to make the word of Geez sentences similar to that of an Amharic sentence.

6.1 Contribution of the work

The contribution of this thesis is as follows.

- The work implemented a novel preprocessing rule based system that process a POS tagged Geez corpus.
- The study showed that making the word order of source and target languages similar has a promising improvement in translation quality of SMT system.
- Preparation of Geez and Amharic corpus files that can be used as parallel corpus for other similar researches.

6.2 Future work

This work aimed at developing a hybrid machine translation that translates Geez to Amharic. The following are future works that can be conducted in areas of NLP on these languages especially Geez.

- Even if the achieved BLEU score with a small corpus used in the work is promising, increasing corpus size will have a significant improvement in BLEU score especially in the development of a real world translation system.
- The preprocessing module can further be expanded to support more complex sentences and apply more reordering rules to further improve the translation quality.
- To aid the translation process various NLP applications shall be developed like morphological analyzer for Geez and POS tagger for Geez.

References

[1] Adam Lopez, “Statistical Machine Translation”, ACM Computing Surveys, Vol. 40, No. 3, August 2008.

[2] Rosna P. Haroon and Shaharban T. A., “Malayalam Machine Translation using Hybrid Approach”, In Proceedings of IEEE International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016.

[3] መምህር ደሴ ቀለብ, ትንሳኤ ግእዝ (The revival of geez) 2ኛ እትም: በኢትዮጲያ አርቶዶክስ ተዋሕዶ ቤተ ክርስቲያን ማኅበረ ቅዱሳን: 2007.

[4] Stefan Weninger, The Semitic Languages An International Handbook, 2011.

[5] C. H. Dawkins, The Fundamentals of Amharic, 1969.

[6] Anais Wion, Collecting manuscripts and scrolls in Ethiopia: The missions of Johannes Flemming (1905) and Enno Littmann (1906), 2012.

[7] መርኖ ሰዋስው ዘልሳነ ግእዝ /በአዲስ ጥበብ/, ዘርአ-ዳዊት አድሐና, 2008 ዓ.ም

[8] Allen, Introduction to Natural Language Understanding, 1995.

[9] Peter Jackson and Isabelle Moulinier, Natural Language Processing for Online Applications Text Retrival, Extraction and Categorization, 2002.

[10] Robert Dale, Hermann Moisl, Harold Somers, Handbook of Natural Language Processing, 2000.

[11] Niti Indurkhya and Fread J. Damerau, Hand book of natural language processing. 2nd edition, 2010.

[12] James Pustejovsky and Amber Stubbs, Natural language annotation for machine learning, 2011.

[13] Daniel Jurafsky and James H. Martin, Speech and language processing, 2000.

[14] Mark Y. Liberman and Kenneth W. Church, “Text analysis and word pronunciation in text-to-speech synthesis”, Advances in speech signal processing, pp. 791-831, Dekker, New York, 1992.

[15] Abdullah H. Homiedan, Machine translation, 2010

[16] Elizabeth D. Liddy, Natural Language Processing, 2001

[17] Eiichiro SUMITA, Yasuhiro AKIBA, Takao DOI, Andrew FINCH, Kenji IMAMURA, Hideo OKUMA, Michael PAUL, Mitsuo SHIMOHATA, Taro WATANABE, "EBMT, SMT, Hybrid and More: ATR Spoken Language Translation System", International Workshop on Spoken Language Translation (IWSLT), 2004.

[18] Christopher D. Manning, Foundations of Statistical Natural Language Processing Cambridge, 1999.

[19] Eiichiro SUMITA, Hitoshi IIDA, and Hideo KOHYAMA, "Translating with examples: A New Approach to Machine Translation", 2005.

[20] Chunyu Kit, Haihua Pan and Jonathan J. Webster, "Example Based Machine Translation: A New Paradigm", 2002.

[21] Mouiad Fadiel Alawneh and Tengku Mohd Sembok, "Rule Based and Example based Machine Translation from English to Arabic", Sixth International Conference on Bio-Inspired Computing: Theories and Applications, 2011.

[22] Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn, "Rule-based Translation With Statistical Phrase-based Post-editing", Proceedings of the Second Workshop on Statistical Machine Translation, pp.203-206, Prague, Czech Republic, 2007.

[23] Wolf Leslau, Comparative Dictionary of Geez, 1991.

[24] ሊቀ ጎሩያን በላይ መኮንን, ሕያው ልሳን ሦስተኛ እትም, 2005 ዓ.ም.

[25] ዘርአ-ዳዊት አድሐና, መርኖ ሰዋስው ዘልሳነ ግእዝ በአዲስ ጥበብ, 2008 ዓ.ም.

[26] መ/ት ኑኅሚን ዋቅጅራ, ማኅቶተ ጥበብ ዘልሳነ ግእዝ 2ኛ እትም: በኢትዮጲያ ኦርቶዶክስ ተዋሕዶ ቤተ ክርስቲያን ማኅበረ ቅዱሳን, 2008.

[27] ተባባሪ ፕሮፌሰር ጌታሁን አማረ, የአማርኛ ሰዋስው በቀላል አቀራረብ, 1989.

[28] Marta R.Costa-Jussa and Jose A.R.Fonollosa, "Latest trends in hybrid machine translation and its applications", Computer Speech & Language, Volume 32, Issue 1, pp. 3-10, 2015.

[29] Thurmair, G., “Comparing different architectures of hybrid machine translation system”, In proceedings of the MT Summit XII, pp. 340-347, 2009.

[30] A.-L. Lagarda, V. Alabau, F. Casacuberta, R. Silva, and E. Dıaz-de-Liano, “Statistical post-editing of a rule based MT”, Proceedings of NAACL HLT: Short Papers, pp. 217–220, 2009.

[31] Almut Silja and Stephan Vogal, “Combination of Ma-chine Translation Systems via Hypothesis Selection from Combined N-Best Lists”, Eighth AMTA Conference, p. 21-25, Hawaii, 2008.

[32] Evi Yulianti, Indra Budi, Achmad N. Hidayanto, Hisar M. Manurung, and Mirna Adriani, “Developing Indonesian-English Hybrid Machine Translation”, in Proceedings of IEEE International Conference on Asian Language Processing (IALP), pp. 173-176, 2012.

[33] Arafat Ahsan, Prasanth Kolachina, Sudheer Kolachina, Dipti Misra Sharma, Rajeev Sangal, “Coupling Statistical Machine Translation with Rule-based Transfer and Generation”, In AMTA-The Ninth Conference of the Association for Machine, Translation in the Americas, Denver, Colorado, 2010.

[34] Michel Simard, Nicola Uef_ng, Pierre Isabelle and Roland Kuhn, “Rule based MT Combined with Statistical Post Editor for Japanese to English Patent Translation”, Proceedings of the Second Workshop on Statistical Machine Translation, pp. 203–206, 2007.

[35] Omkar Dhariya, Shrikant Malviya and Uma Shanker Tiwary, “A Hybrid Approach For Hindi-English Machine Translation”, 31st International Conference on Information Networking (ICOIN), 2017.

[36] Kenji Yamada and Kevin Knight, “Syntax-based Multi-system Machine Translation”, Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 523-530, Toulouse, France, 2001.

[37] Jaime G. Carbonell, Steve Klein, David Miller, Mike Steinbaum, and Tomer Grassiany, “Context-Based Machine Translation”, In Proceeding of the 7th Conference of The

Association for Machine Translation in the Americas “Visions for the Future of Machine Translation”, pp. 19-28, Cambridge, MA, USA, 2006.

[38] Anitha T Nair and Sumam Mary Idicula, “Syntactic based MT from English to Malayalam”, In Proceedings of IEEE International Conference on Data Science & Engineering (ICDSE), 2012.

[39] Fangli Liang, Lei Chen, and Miao Li, “A rule based source side reordering on phrase structure sub trees”, In Proceedings of IEEE International Conference on Asian Language Processing (IALP), 2011.

[40] Chung-chi Huang, Wei-teh Chen, and Jason S. Chang, “Source sentence reordering for Phrase-based MT”, In Proceedings of IEEE 4th International Conference on Innovative Computing, Information and Control (ICICIC), 2009.

[41] Rahul.C, Dinunath.K, Remya Ravindran, K.P.Soman, “Rule Based Reordering and Morphological Processing for English-Malayalam Statistical machine Translation”, International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009.

[42] M.Kasthuri and S.Britto Ramesh Kumar, “Rule Based Machine Translation System from English to Tamil”, In Proceedings of IEEE World Congress on Computing and Communication Technologies, 2014.

[43] Michael Gasser, “Toward a rule-based system for English-Amharic Translation”, Proceedings of the Workshop on Language Technology for Normalization of Less-Resourced Languages (SaLTMiL 8 - AfLaT2012), 2012.

[44] Sankaravelayuthan, Rajendran & Vasuki, G., “English to Tamil Machine Translation System Using Parallel Corpus”, Unpublished PhD Thesis, Amrita University, Coimbatore, India, 2013.

[45] Philipp Koehn, Statistical Machine Translation, 2010.

[46] https://www.w3schools.com/charsets/ref_html_utf8.asp Last Accessed on March 27, 2018.

[47] https://www.ubuntu.com/ Last Accessed on March 27, 2018.

[48] Moses for Mere Mortals user manual, Retrieved from http://www.cs.cmu.edu/afs/cs/project/cmt-55/lti/Courses/731/homework/mosesdecoder/scripts/moses-for-mere-mortals/ Last Accessed on August 5, 2017.

[49] https://hlt-mt.fbk.eu/technologies/irstlm Last Accessed on February 4, 2018.

Appendix I: Sample Training Corpus for the Proposed System

Geez Sentences	Amharic Sentences
አንተ ዐርከ አፍቅርከ	አንተ ጓደኛነ ወደሃል
አንትሙ ወርቀ ዘገብክሙ	እናንተ ወርቅን ሰብስባችኋል
አንቲ አፈወ ተቀባእኪ	አንቺ ሽንቶ ተቀብተሻል
አንትን ማዕደ ሠራዕክን	እናንቺ ማዕድን ሠርታችኋል
ውእቱ ሕጻነ አልሐቀ	እሱ ሕጻንን አሳድጓል
ውእቶሙ ቢጾሙ ሐመዩ	እነሱ ጓደኞቸውን አምተዋል
እሙንቱ ቢጾሙ ሐመዩ	እነሱ ጓደኞቸውን አምተዋል
ይእቲ ቢጾሙ ሐመዩ	እነሱ ጓደኞቸውን አምተዋል
ይእቲ ሥና ነጸረት	እሷ ውበቷን ተመልክታለች
እማንቱ ሜላተ ፈተላ	እነሷ ሀርን ፈትለዋል
አኀዊነ ተመሰሉ ኪያየ	ወንድሞቻችን እኔን ምሰሉ
ኪያከ ተወከልኩ	አንተን ታምኛለሁ
ኪያክሙ ይሰስድክሙ	እናንተንም ያሳድዷችኋል
ውእቱ ኪያየ ተወከፈ	እሱ እኔን ተቀበለ
አሕዛብ ኪያነ ይሰደዱ	አሕዛብ እኛን ያሳድዳሉ
ሕግ ኪያከ ያገርር	ሕግ አንተን ይገዛል
ወልድ ኪያክሙ ፈነወ	ወልድ እናንተን ላከ
ወሬዛ ኪያኪ ሐደገ	ጎልማሳ አንቺን ተወ
ጴጥሮስ ኪያክን በደረ	ጴጥሮስ እናንቺን ቀደመ

Appendix II: Sample Training Corpus for the Baseline SMT

Geez Sentence	Amharic Sentence
አንተ አፍቅርከ ዐርከ	አንተ ጓደኛነ ወደሃል
አንትሙ ዘገብክሙ ወርቀ	እናንተ ወርቅን ሰብስባችኋል
አንቲ ተቀባእኪ አፈወ	አንቺ ሽንቶ ተቀብተሻል
አንትን ሠራዕክን ማዕደ	እናንቺ ማዕድን ሠርታችኋል
ውእቱ አልሐቀ ሕጻነ	እሱ ሕጻንን አሳድጓል
ውእቶሙ ሐመዩ ቢጾሙ	እነሱ ጓደኞቸውን አምተዋል
እሙንቱ ሐመዩ ቢጾሙ	እነሱ ጓደኞቸውን አምተዋል
ይእቲ ሐመዩ ቢጾሙ	እነሱ ጓደኞቸውን አምተዋል
ይእቲ ነጸረት ሥና	እሷ ውበቷን ተመልክታለች
እማንቱ ፈተላ ሜላተ	እነሷ ሀርን ፈትለዋል
አኀዊነ ኪያየ ተመሰሉ	ወንድሞቻችን እኔን ምሰሉ
ኪያከ ተወከልኩ	አንተን ታምኛለሁ
ኪያክሙ ይሰስድክሙ	እናንተንም ያሳድዷችኋል
ውእቱ ተወከፈ ኪያየ	እሱ እኔን ተቀበለ
አሕዛብ ይሰደዱ ኪያነ	አሕዛብ እኛን ያሳድዳሉ
ሕግ ያገርር ኪያከ	ሕግ አንተን ይገዛል
ወልድ ፈነወ ኪያክሙ	ወልድ እናንተን ላከ
ወሬዛ ሐደገ ኪያኪ	ጎልማሳ አንቺን ተወ
ጴጥሮስ በደረ ኪያክን	ጴጥሮስ እናንቺን ቀደመ

Appendix III: Sample Testing Data used for the Proposed and Baseline System

Geez sentence for the proposed system	Geez sentence for the Baseline SMT
እግዚያብሔር ፈጠረ ኪያነ ሕዝቦ	ፈጠረ እግዚያብሔር ኪያነ ሕዝቦ
ውእቶን	ውእቶን
ሰብአት ተፈጥሩ	ሰብአት ተፈጥሩ
ውእቶን ሜላተ ፈተላ	ውእቶን ፈተላ ሜላተ
ጠበብት ቤተ የሐንጹ	ጠበብት የሐንጹ ቤተ
ሙሴ ባሕረ ከፈለ	ከፈለ ሙሴ ባሕረ
ዳዊት አምላክ ሰብሐ	ዳዊት ሰብሐ አምላክ
ሐሜት ሕይወተ ታረኩሰ	ሐሜት ታረኩሰ ሕይወተ
ጤሜዎስ ዕውር ብርሃነ ርእየ	ጤሜዎስ ዕውር ርእየ ብርሃነ
ዜናዊ ጥበበ ገብርኤል ብስራተ ዜነወ	ዜናዊ ጥበበ ገብርኤል ዜነወ ብስራተ
ቃኤል ቅቱለ አቤል ሰማዕተ ኮነ	ቅቱለ ቃኤል አቤል ኮነ ሰማዕተ
ማርቆስ ግብፀ ዖደ	ዖደ ማርቆስ ግብፀ
መዓትም ብእሲ ሕጻናተ ቀተለ	መዓትም ብእሲ ቀተለ ሕጻናተ
ደብር እግረ ቢታንያ አውጽአት ሰርጸ	እግረ ደብር ቢታንያ አውጽአት ሰርጸ
ዘጌሳ ሰሎሜወማርያም መልአከ ረከባ	ዘጌሳ ሰሎሜወማርያም ረከባ መልአከ
ዮሐንስ ሐዋርያ ወንጌለ ጸሐፈ	ሐዋርያ ዮሐንስ ጸሐፈ ወንጌለ
እግዚአብሔር ዐማፄ ኤሳውሀ ጸልአ	ዐማፄ ኤሳውሀ ጸልአ እግዚአብሔር

Declaration

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

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _____

Signature: _____

Date: _____