ADDIS ABABA UNIVERSITY

College of Natural and Computional Sciences

School of Information Science

A Bidirectional Tigrigna – English Statistical Machine Translation

A Thesis Submitted to the School of Information Science in Partial Fulfillment for the Degree of Master of Science in Information Science

By: *Mulubrhan Hailegebreal*

Advisor: *Martha Yifiru (PhD)*

Addis Ababa, Ethiopia
June, 2017

Addis Ababa University
School of Graduate Studies
College of Natural Sciences
School of Information Science

Mulubrhan Hailegebreal

Advisor: *Martha Yifiru (PhD)*

This is to certify that the thesis prepared by *Mulubrhan Hailegebreal*, titled: *Bidirectional Tigrigna – English Statistical Machine Translation* and submitted in partial fulfillment of the requirement for the Degree of Master of Science in Information Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor :	Dr. Martha Yifiru		June, 2017
Examiner :	Dr. Dereje Teferi		June, 2017
Examiner :	Dr. Million Meshesha		June, 2017

Dedication

I would like to dedicate this work to my family for their unconditional love.

Especially to: *Hailegebreal G/cherkos (Father) and*

Azmera Berhe (Mother)

Tsehay Hailegebreal (Sister)

Acknowledgements

This research work would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

First and foremost, I would like to express my gratitude to my advisor Dr. Martha Yifiru for showing the faith and confidence in me, and for her invaluable guidance, monitoring and support during the course of my work.

I owe my thanks to school of Information Science, Addis Ababa University, each staff member contributed in their own way towards the completion of this work. A special mention to Dr. Solomon Teferra who inspired me a lot and gave me due direction with his experience in NLP research.

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

I am grateful to my classmates for creating such a wonderful working environment and for all the fun during the last two years.

All my friends deserve special thanks. They are the ones who are always there to spend time with, and share my joys and troubles.

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

Abstract

Machine Translation (MT) is one task in Natural Language Processing (NLP), where automatic systems are used to translate text from one (source) language to another (target) while preserving the meaning of source language. Since there is a need for translation of documents between Tigrigna and English languages, there needs to be a mechanism to do so. Hence, this study explored the possibility of developing Tigrigna – English statistical machine translation and improving the translation quality by applying linguistic information. In this work, experimental quantitative research method is used.

In order to achieve the objective of this research work, a corpora are collected from different domain and classified into five sets of corpora, and prepared in a format suitable for use in the development process. In order to realize the goal, three sets of experiments are conducted: baseline (phrase based machine translation system), morph-based (based on morphemes obtained using unsupervised method) and post processed segmented systems (based on morphemes obtained by post-processing the output of the unsupervised segmenter). We work on MOSES which is a free statistical machine translation framework, which allows automatically training translation model using parallel corpus. Since the system is bidirectional, four language models are developed; one for English and the other three are for Tigrigna language includes for baseline, morph-based and the other for the post processed experiment. Translation models which assigns a probability that a given source language text generates a target language text are built and a decoder which searches for the shortest path is used. BLUE score is used to evaluate the performance of each set of experiment.

Accordingly, the result obtained from the post processed experiment using corpus II has outperformed the other, and the result obtained has a BLUE score of 53.35 % for Tigrigna – English and 22.46 % for English – Tigrigna translations. This research focuses on segmenting prepositions and conjunctions because of data scarcity . Therefore future research should focus to further improve the BLUE score by applying semi supervised segmentation to include the remaining linguistic information.

Keywords: Machine translation, Statistical Machine Translation, Segmentation, Tigrigna – English

Table of Contents

Abstract iv
List of Tables viii
List of Figures ix
List of Algorithms x
List of Abbreviations xi
CHAPTER ONE 1
Introduction 1
1.1 Background 1
1.2 Motivation 2
1.3 Statement of the Problem 3
1.4 Objectives of the Study 5
1.4.1 General Objective 5
1.4.2 Specific Objectives 5
1.5 Methodology 5
1.5.1 Research Design 5
1.5.2 Data Collection 5
1.5.3 Tools and Techniques 6
1.5.4 Evaluation techniques 6
1.6 Scope and Limitation 7
1.7 Significance of the Study 7
1.8 Thesis organization 7
CHAPTER TWO 9
Literature Review 9
2.1 An overview of Machine Translation 9
2.2 Machine Translation Approaches 10
2.2.1 Rule-Based Machine Translation (RBMT) 10
2.2.2 Example Based Machine Translation (EBMT) 13
2.2.3 Statistical Based Machine Translation (SMT) 14
2.2.4 Hybrid Based Machine Translation (HMT) 22
2.4 Text Alignment 22
2.4.1 IBM Alignment 23

2.4.2 HMM Alignment 23
2.5 Evaluation of Machine Translation 24
2.5.1 BLEU Score 24
2.6 Related Works 26
2.6.1 German – English and French – English Statistical Machine Translation 26
2.6.2 English-to-Arabic Statistical Machine Translation 27
2.6.3 Hindi to English Statistical Machine Translation 27
2.6.4 English-to-Czech Factored Machine Translation 28
2.6.5 English-to-Thai Factored Machine Translation 28
2.6.6 English – Amharic Statistical Machine Translation 29
2.6.7 English – Afaan Oromo Machine Translation 30
2.6.8 Summary 31
CHAPTER THREE 33
The Structure of Tigrigna and English languages 33
3.1 Overview of Tigrigna Language 33
3.1.1 The Writing System 33
3.1.1 Tigrigna Language Parts of Speech 34
3.1.2 Morphology of Tigrigna Language 38
3.1.3 Derivational Morphology of Tigrigna Language 39
3.1.4 Inflectional Morphology of Tigrigna Language 41
3.1.5 Tigrigna Phrases Classification 43
3.1.6 Tigrigna Sentence Structure 44
3.2 Overview of English Language 45
3.3 Challenges of Tigrigna and English Language to SMT 48
CHAPTER FOUR 49
Design of Bidirectional Tigrigna – English Machine Translation (BTEMT) 49
4.1 Architecture of BTEMT 49
4.2 The Corpora 51
4.2.1 Preliminary Preparation 51
4.2.2 Corpora Preparation 51
4.3 Preprocessing the Corpora 53
4.4 Morphological Segmentation 53
4.4.1 Tokenizer and Frequency Calculator 55
4.4.2 Segmentation Learner 56

4.4.3 Segmenter 57
4.4.4 Post-processor 58
4.5 Language Model 59
4.6 Translation Model 62
4.6.1 Alignment 62
4.7 Decoder 66
4.8 Evaluation 66
4.9 Morphological Concatenation 67
CHAPTER FIVE 68
Experiment 68
5.1 Experimental Setup 68
5.1 Parallel corpora 69
5.2 Training the System 70
5.3 Experiment I: Baseline SMT 70
5.4 Experiment II: Morph-based System 73
5.5 Experiment III: Post-Processed Segmented System 75
5.5.1 Post-Processed Segmentation Model 75
5.6 Finding 80
CHAPTER SIX 82
Conclusion and Recommendation 82
6.1 Conclusion 82
6.2 Recommendation 83
References 85
Appendices 89
Appendix I : Tigrigna Alphabet (Fidel) 89
Appendix II : Snapshot Result From Baseline System 90
Appendix III: Snapshot Result from the Morph-based System 93
Appendix IV: Snapshot Result From Post-processed segmented system 96
Appendix V: Sample Parallel Corpus for training from Corpus I 99
Appendix VI: Sample Parallel Corpus for training from Corpus II 101
Appendix VII: Sample translated output from Post Segmented System 103
Appendix VIII: Sample translated output from Post Segmented System 104

List of Tables

Table 3.1: Summary of coordinate conjunctions and interjections 36
Table 3.2: Summary of subordinate conjunctions 37
Table 3.3 Tigrigna affixes 39
Table 3.4 Inflections of perfective tense 41
Table 3.5 Inflections of imperfective tense 42
Table 3.6 Inflection of Nouns 43
Table 3.7 Inflection of Adjectives from [44] 43
Table 3.8 : Punctuation marks used in both languages 34
Table 5.1: summary of the size of the total data used in the research 68
Table 5.2: Summary of BLUE score from the Baseline System 71
Table 5.3 Sample of translation errors 72
Table 5.4 : Summary counts of token and vocabulary 73
Table 5.5 : Summary of BLUE score from the Morph-based System 74
Table 5.6 Sample of over and under segmented morphems 74
Table 5.7 : List of Tigrigna conjunctions and prepositions found in affixes form 75
Table 5.8 : Sample of the processed segmentation model 77
Table 5.9: Summary counts of token and vocabulary 79
Table 5.10 : Summary of BLUE score from the Post-processed segmented system 79
Table V.1: sample parallel data from Corpus I 100
Table VI.1: Sample Parallel data from Corpus II 102
Table VII.1: Sample English – Tigrigna Translation 103
Table VIII.1: Sample Tigrigna – English Translation 104

List of Figures

Figure 2.1: Direct based machine translation [21] 11
Figure 2.2: transfer based machine translation [22] 11
Figure 2.3: Interlingua model [22] 12
Figure 2.4: Different methods of RBMT adopted from [21] 13
Figure 2.5: Factored translation model; adopted from [33] 20
Figure 3.1 Some Fidels (alphabet) of Tigrigna languages 33
Figure 3.2 Structure of Tigrigna and English languages 45
Figure 4.1: Architecture of the proposed system 50
Figure 4.2 : Architecture of the Morphological Segmenter 55
Figure 4.3: Sample output from the Segmentation learner 57
Figure 4.4. Input to the unsupervised segmenter 58
Figure 4.5. Output from the post-segmentation learner 58
Figure I.1: Tigrigna Alphabet (Fidel) 89
Figure II.1: English – Tigrigna BLEU score result from Corpus I 90
Figure II.2: Tigrigna – English BLEU score result from Corpus I 90
Figure II.3: English – Tigrigna BLEU score result from Corpus II 90
Figure II.4: Tigrigna – English BLEU score result from Corpus II 90
Figure II.5: English – Tigrigna BLEU score result from Corpus III 91
Figure II.6 Tigrigna – English BLEU score result from Corpus III 91
Figure II.7: English – Tigrigna BLEU score result from Corpus IV 91
Figure II.8 Tigrigna – English BLEU score result from Corpus IV 91
Figure II.9: English – Tigrigna BLEU score result from Corpus V 92
Figure II.10: Tigrigna – English BLEU score result from Corpus V 92
Figure III.1: English – Tigrigna BLEU score result from Corpus I 93
Figure III.2: Tigrigna – English BLEU score result from Corpus I 93
Figure III.3: English – Tigrigna BLEU score result from Corpus II 93
Figure III.4: Tigrigna – English BLEU score result from Corpus II 93
Figure III.5: English – Tigrigna BLEU score result from Corpus III 94
Figure III.6: Tigrigna – English BLEU score result from Corpus III 94
Figure III.7: English – Tigrigna BLEU score result from Corpus IV 94
Figure III.8: Tigrigna – English BLEU score result from Corpus IV 94
Figure III.9: English – Tigrigna BLEU score result from Corpus V 95
Figure III.10: Tigrigna – English BLEU score result from Corpus V 95
Figure IV.1: English – Tigrigna BLEU score result from Corpus I 96
Figure IV.2: Tigrigna – English BLEU score result from Corpus I 96
Figure IV.3: English – Tigrigna BLEU score result from Corpus II 96
Figure IV.4: Tigrigna – English BLEU score result from Corpus II 96
Figure IV.5: English – Tigrigna BLEU score result from Corpus III 97
Figure IV.6: Tigrigna – English BLEU score result from Corpus III 97
Figure IV.7: English – Tigrigna BLEU score result from Corpus IV 97
Figure IV.8: Tigrigna – English BLEU score result from Corpus IV 97
Figure IV.9: English – Tigrigna BLEU score result from Corpus V 98
Figure IV.10: Tigrigna – English BLEU score result from Corpus V 98

List of Algorithms

Algorithm 4.1: Algorithm for word tokenizing and frequency calculator 56
Algorithm 4.2 : Algorithm for Post processing the segmented Tigrigna sentences 59
Algorithm 4.3 : Algorithm for concatenation the Tigrigna target sentences 67
Algorithm 5.1: Algorithm for processing the segmentation model 76
Algorithm 5.2 : Algorithm for concatenating the Tigrigna target sentences 78

List of Abbreviations

BLUE: BiLingual Understudy Evaluation

BTEMTS: Bidirectional Tigrigna – English Machine Translation

CAT : Computer Aided Translation

EBMT: Example Based Machine Translation

EM: Expected Maximization

FDRE: Federal Democratic Republic of Ethiopia

HAMT: Human Aided Machine Translation

HMM: Hidden Markov Model

IRSTLM : IRST Language Model

LM: Language Model

MAHT : Machine Aided Human Translation

MERT: Minimum Error Rate Training

MT: Machine Translation

NLP: Natural Language Processing

POS: Part of Speech

RBMT: Rule Based Machine Translation

RTF: Rich Text Format

SMT: Statistical Machine Translation

SRILM: Stanford Research Institute Language Model

TM: Translation Model

VM: Virtual Machine

CHAPTER ONE

Introduction

1.1 Background

Natural Language Processing (NLP) is a branch of artificial intelligence that deals with analyzing, understanding and generating the languages that humans use naturally in order to interact with computers in both written and spoken contexts. One of the challenges inherent in natural language processing is teaching computers to understand the way humans learn and use language. There are theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications [1].

NLP is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech. Natural language processing researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks [2]. Computers do not genuinely understand ordinary language or converse naturally with us about arbitrary topics. That is a barrier separating us from the machines. The goal of research in NLP and in natural language understanding is to break down this barrier. In fact, any application that utilizes text is a candidate for NLP. The most frequent NLP applications include: Machine Translation, Text Summarization, Information Retrieval, Information Extraction, and Question-Answering [1].

Machine translation is one of the NLP applications that deals with automated translation of text or speech from source language to target language or “translation carried out by a computer”, as defined in the Oxford English dictionary (2016 edition). A wide variety of machine translation approaches have been developed in the past years such as rule based, example based, statistical machine translation and hybrid approaches [2]. Rule based machine translation is a general term that denotes machine translation systems based on linguistic information about source and target languages basically retrieved from (bilingual) dictionaries and grammars. Example based machine translation is characterized by its use of bilingual corpus with parallel texts as its main knowledge, in which translation by analogy is the main idea. Statistical machine translation (SMT) is generated

on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. By taking the advantage rule based, example based, statistical based translation methodologies, a new approach was developed, called hybrid based approach, which has proven to have better efficiency in the area of MT systems.

In recent years there has been an enormous boom in MT research. There has been not only an increase in the number of research groups in the field and in the amount of funding, but there is now also optimism for the future of the field and for achieving even better quality. The major reason for this change has been a paradigm shift away from linguistic/rule-based methods towards empirical/data-driven methods in MT. This has been made possible by the availability of large amounts of training data and large computational resources. This paradigm shift towards empirical methods has fundamentally changed the way MT research is done. The field faces new challenges. For achieving optimal MT quality, we want to train models on as much data as possible, ideally language models trained on hundreds of billions of words and translation models trained on hundreds of millions of words. Doing that requires very large computational resources, a corresponding software infrastructure, and a focus on systems building and engineering [3].

This study discuss, automatic translation between English and Tigrigna languages using statistical approach. The two languages are from different language families and this makes the machine translation task very difficult. English is from Indo-European language family [4] and a language that is mainly spoken on different parts of the world, for example it is the foreign language in Ethiopia and Eritrea as a medium of instruction and communication in governmental and non-governmental organization. And Tigrigna is from Semitic branch of the Afro-Asiatic family spoken in northern part of Ethiopia and Eritrea, which is the official language of Tigray regional state and one of the official languages of Eritrea [5].

1.2 Motivation

Machine Translation has widespread commercial, military and political applications [6]. For example, increasingly, the Web is accessed by non-English speakers, reading non-English pages . The ability to find relevant information should not be bounded by our language-speaking capabilities. Basically researcher have been motivated to discuss and propose direction to problems of machine translation community. The problems are [7]:

- Translation into morphologically rich languages: Most MT systems will not generate word forms that they have not observed, languages like Tigrigna.
- Translation of low-resource language pairs: The most straightforward example of unresourced languages like Ethiopic languages.
- Translation across domains: Translation systems are not robust across different types of data, performing poorly on text whose underlying properties differ from those of the system's training data.

1.3 Statement of the Problem

Machine translation for different languages has been done so far by different researchers, but many researches and applications have done for foreign languages using different methodologies and approaches. Most of the works have been done on language pair of English and other foreign languages, such as Arabic [10],Japanese [11], and French [12] etc. This is because English languages is one of the dominant languages spoken in the world and the language of international communication.

Tigrigna is the official and working language of Tigray region of Ethiopia and one of the official languages of Eritrea. The language is spoken approximately by 7 million people in the northern part of Ethiopia (Tigray region) and central part of Eritrea [8]. And English language is a language that was first spoken in Anglo-Saxon England in the early middle Ages. It is now the most widely used language in the world. It is spoken in many countries around the world. It is the first language of the United Kingdom, the United States, Canada, Australia, Ireland, New Zealand and a number of Caribbean nations [9]. So the Tigrigna speakers faces the following challenges:

- Lack of information: As English is one of the dominant language spoken in the world, world news, journals, books, international documents and others would be written in English language so that the speakers other than English would not be aware of the information.
- Communication gap: Speakers of Tigrigna who couldn't speak and understand English cannot communicate with English speakers.

However, only little works have been done on language pair of English and Ethiopian languages. Some of the studies are carried out on English-Amharic language pair [13], [14] and English – Affan Oromo language pair [15] [16].

But, to the best knowledge of the researchers, there is no prior investigation on the development of machine translation system on the pair of English-Tigrigna or Tigrigna-English. As a result, the way of developing a machine translation system (for these language pairs) with a reasonable performance is not known. Because of this, people use human translation and they tend to be slower as compared to machines. Sometimes it can be hard to get a precise translation that reveals what the text is about without everything being translated word-by-word, so everyone needs a well-organized and proficient translation. Those who can speak both languages need it for confirmation purposes and those who only speak one of the languages need it for grasping knowledge, if the translation is bidirectional. In addition, it can be more important to get the result without delay which is hard to accomplish with a human translator, and MT allows documents to be translated that would otherwise remain untranslated due to lack of resources.

The purpose of this research is, therefore, to investigate the ways of developing bidirectional Tigrigna – English machine translation systems that has a reasonable performance. This will enable the developers to easily adopt the best way and develop machine translation systems for these two language pairs. According to Online Computer Library Center (OCLC) there are abundant documents in English, and there is also a need to use the ever increasing amount of useful information produced throughout the world available in English language. However, lack of NLP resources and English language knowledge creates a problem of fully utilizing these documents. The researcher believes that investigating the ways of developing bidirectional Tigrigna – English machine translation makes a direction to make available these documents in local language (Tigrigna) and is vital in addressing the language barrier thereby reducing the effect of digital divide. The research answers the following research questions:

- Does the type and size of corpus has impact on the performance of the machine translation system of these language pairs?
- Does morphological segmentation improve the quality of machine translation system of the Tigrigna – English language pairs?

- Which kind of morpheme improves the performance of the translation system for the Tigrigna – English language pairs?

1.4 Objectives of the Study

1.4.1 General Objective

The general objective of this research is to design a bidirectional Tigrigna – English machine translation system using statistical approach.

1.4.2 Specific Objectives

The following specific objectives are identified in order to accomplish the above general objective.

- To review related systems and literature,
- To understand the linguistic behaviors of both languages,
- To develop parallel bilingual corpus for English and Tigrigna languages,
- To prepare monolingual corpora for the language modelling purpose,
- To construct bidirectional Tigrigna – English machine translation model and
- To test and evaluate the performance of the model.

1.5 Methodology

To accomplish the research a lot of methods, tools, and techniques are applied.

1.5.1 Research Design

The research design for this study is Experimental quantitative research method. The methodology of a quantitative research maintains the assumption of an empirical knowledge based on data, in this case the text data called corpora is employed. For this study, statistical machine translation or corpus based approach was employed. The reason why the researchers prefer statistical approach is; this approach does not require linguistic preprocessing of documents and it is also data driven i.e. it learns from the data distribution itself.

1.5.2 Data Preparation

In order to develop translation models for Bidirectional Tigrigna – English machine translation systems, a detailed literature review has been done on machine translation on different languages pairs in general and statistical approach of machine translation in particular with regard to techniques used in the approach. And a parallel corpora are prepared from different sources such

as Holy Bible, Constitution of the FDRE, and simple sentences customized from [12]. Since Tigrigna is under-resourced language, the corpora are limited to these three different sources. And the collected data are grouped in to five; this is used to recognize the impact of corpus size and types on performance of machine translation. The first group of corpus is named as Corpus I and contained 6106 relatively long sentences pairs which is collected from Holy Bible. The second group of corpus named as Corpus II and contains 1160 simple sentences from Amharic corpus

The third group of corpus named as Corpus III and contained 1982 sentence pairs which are collected from two sources; 822 sentences from Constitution of FDRE and 1060 simple sentences customized from Amharic corpus, the reason why we decide to combine is because of the size. From this Corpus four which is named as Corpus IV is prepared by combining Corpus I and II.

And the fifth group of corpus is named as Corpus V and contained 3000 sentence pairs which is prepared by combining all sentences from Corpus III and 1018 sentence pairs randomly selected from Corpus I. different preprocessing techniques are implemented such as : sentence level alignment using Hunalign, splitting punctuation marks by using Perl script [17] and true-casing is also implemented for English language.

1.5.3 Tools and Techniques

Since we have developed the bidirectional Tigrigna - English machine translation system by using statistical machine translation approach, we have used freely available software such as : the widely-used language modeling tool IRST Language Model (IRSTLM) toolkit is used for language modeling because the Moses MT system has a support for IRSTLM as a language modelling tool, and it requires less memory than other tools [17]. For alignment, the state of- the-art tool is GIZA++, which implements the word alignment methods IBM1 to IBM5. Decoding has been done using Moses, which is an SMT system that is also used to train translation models to produce phrase tables., Ubuntu operating system which is suitable for the Moses environment, and VMware workstation, a workstation that enables us to set up multiple virtual machines (VMs) and use them simultaneously along with the actual machine have been used. In addition, unsupervised Morpheme segmentation tool; Morfessor 1.0 is used for segmenting Tigrigna sentence.

1.5.4 Evaluation techniques

Machine translation systems can be evaluated by using human evaluation method or automatic evaluation method. Since human evaluation is time consuming compared with automatic

evaluation, we have used BLEU (Bilingual Evaluation Understudy) score to evaluate the performance of the system, which is an automatic evaluation technique.

1.6 Scope and Limitation

The bidirectional Tigrigna – English machine translation system was designed using a statistical machine translation approach which enables to translate an English text into Tigrigna text and vice versa. Researchers try to show that integrating linguistic information to the statistical machine translation improves the translation quality of morphological rich languages [18]. The number of the linguistic features to be applied to improve the quality of the translation to be studied depends on the time and resource availability. The time and resource that we have are limited to apply all the linguistic features that can improve SMT systems. However, we tried to use morpheme segmentation (Prepositions and Conjunctions) in the SMT systems since Tigrigna is morphologically rich language. And the other challenge is the scarcity of Tigrigna – English parallel corpus and this makes the systems limited to some specific domains: Holy Bible, Constitution of FDRE, and some simple sentences. Translation performance depends on every component of the system, so every component need to be checked. But the Language model and segmentation components are not evaluated because of time.

1.7 Significance of the Study

Machine translation plays a great role in exchanging information among different languages in the globe. This thesis work is believed to propose an effective model for Tigrigna to English and English to Tigrigna machine translation. Therefore, it will have a significant contribution in the following areas:

- For Natural Language applications, to integrate this research in other studies like speech to speech Tigrigna - English translation,
- The research done helps to adopt for other local languages, like Tigrigna – Amharic machine translation;
- For developing English – Tigrigna and Tigrigna – English translation systems

1.8 Thesis organization

This thesis paper is organized into six Chapters. The first Chapter describes the background of the study, statement of the problem, objective of the study, methodology, scope and limitation, and

significance of the study. Chapter two presents literature review on machine translation and its approaches, and related works on statistical machine translation for different language pairs. Chapter three presents an overview of the Tigrigna language such as; Tigrigna POS, morphology of Tigrigna, Tigrigna phrasal classification, Tigrigna sentence structure, numbering and punctuation, and the challenges of Tigrigna language to SMT. Chapter four presents approach followed, corpora preparation, and design of bidirectional Tigrigna – English machine translation using statistical approach. Chapter five presents the experimental part of the work, and finally Chapter six presents conclusion and recommendations.

CHAPTER TWO

Literature Review

In this Chapter, a brief overview of machine translation (MT) and review of related works are provided. The Chapter briefly describes machine translation approaches, which are statistical machine translation (SMT), rule based machine translation (RBMT), example based machine translation (EBMT) and hybrid machine translation (HMT), the alignment techniques and the machine translation evaluation metrics.

2.1 An overview of Machine Translation

Machine translation is automated translation system. It is a computer software which is responsible for the production of translations with or without human assistance. The ideal aim of machine translation systems is to produce the best possible translation with or without human assistance. To process any translation, human or automated, the meaning of a text in the original (source) language must be fully restored in the target language. While on the surface, this seems straightforward, it is far more complex. Translation is not a mere word-for-word substitution. A translator must interpret and analyze all the elements in the text and know how each word may influence another. This requires extensive expertise in grammar, syntax (sentence structure), semantics (meanings), etc., in the source and target languages, as well as familiarity with each local region [19].

Machine translation systems are designed either for two languages (bilingual systems) or for more than a single pair of languages (multilingual systems). Bilingual systems may be designed to operate either in only one direction (unidirectional), e.g. from English into Tigrigna, or in both directions (bidirectional). Multilingual systems are usually intended to be bidirectional; most bilingual systems are unidirectional. A translation can be human-aided machine translation (HAMT) or it can be machine-aided human translation (MAHT). In case of machine-aided human translation, the translation is performed by human translators with the help of computer-based translation tools. In the case of human-aided machine translation, the translation process is performed by computer with the help of humans. Humans are involved before the translation process which is called pre-editing or it can be after the translation process which is called post-editing. The boundaries between MAHT and HAMT are often uncertain and the term computer-

aided translation (CAT) can cover both, but the central core of MT itself is the automation of the full translation process [19].

2.2 Machine Translation Approaches

Machine translation is one of the research areas under computational linguistics with different approaches. The basic approaches of machine translation are [20]: RBMT, SMT, EBMT, and HMT.

2.2.1 Rule-Based Machine Translation (RBMT)

Rule based machine translation also known as Knowledge-Based Machine Translation, which is the classical approach of MT, is a general term that denotes machine translation systems based on linguistic information about source and target languages. It has much to do with the morphological, syntactic and semantic information about the languages. RBMT is able to deal with the needs of wide variety of linguistic phenomena and is extensible and maintainable [20]. In a rule-based machine translation system the original text is first analyzed morphologically and syntactically in order to obtain a syntactic representation. This representation can then be refined to a more abstract level, putting emphasis on the parts relevant for translation and ignoring other types of information. The transfer process, then converts this final representation (still in the original language) to a representation of the same level of abstraction in the target language. From the target language representation, the stages are then applied in reverse. And translation in rule based machine translation system is done by pattern matching of the rules. The success lies in avoiding the pattern matching of unfruitful rules. Knowledge and reasoning are used for language understanding.

The advantage of rule based machine translation approach is that it can deeply analyze at syntax and semantic levels. But there are also drawbacks such as requirement of huge linguistic knowledge and very large number of rules to cover all the features of a language. The rule based approach can be classified as direct, transfer-based and Interlingua [20].

Direct (dictionary based) transfer, which uses a large bilingual dictionary and the translation is done word by word. Words of source language are translated without passing through an additional information of the language. Figure 2.1 describes the detail process and the major components of direct methods are [20]:

- ✓ Shallow morphological analysis

- ✓ Lexical transfer using bilingual dictionary
- ✓ Local reordering
- ✓ Morphological generation

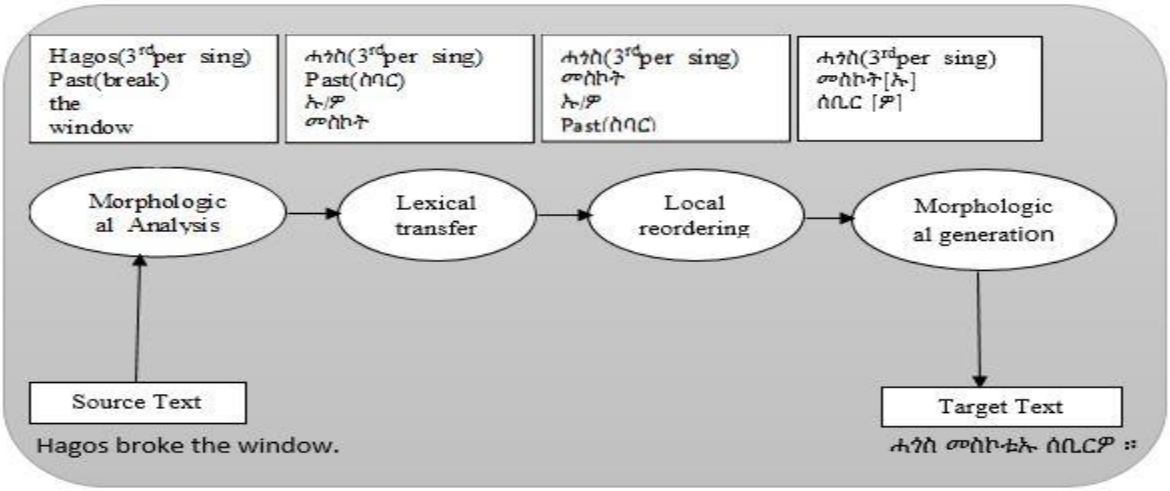


Figure 2.1: Direct based machine translation [21]

Transfer based: uses an intermediate representation that captures the structure of the original text in order to generate the correct translation. In transfer based approach the source language is transformed into abstract, less language-specific representation. The process of transfer-based translation involves analysis, structural transfer, and generation. Analysis of the source text is done based on linguistic information such as morphology, syntax, and semantics. The semantic/ syntactic structure of source language is then transferred into the semantic/ syntactic structure of the target language. The generation phase replaces the constitutes in the source language to the target language equivalent [20]. Figure 2.2 shows the process.

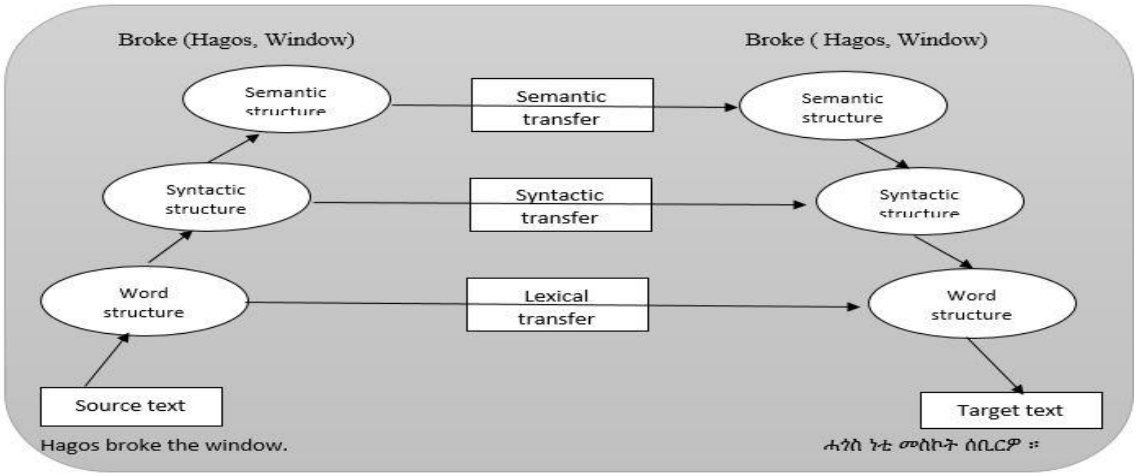


Figure 2.2: transfer based machine translation [22]

Interlingua: in interlingua, the source language is transformed into an intermediary language (representation) which is independent of any of the languages involved in the translation. The translated verse for the target language is then derived through this intermediary representation. The representation is thus a prediction from the source text and at the same time acts as the basis for the generation of the target text; it is an abstract representation of the target text as well as a representation of the source text. The system requires analysis and synthesis modules. Because of its language independency on the language pair for translation, this system has much relevance in multilingual machine translation [20].

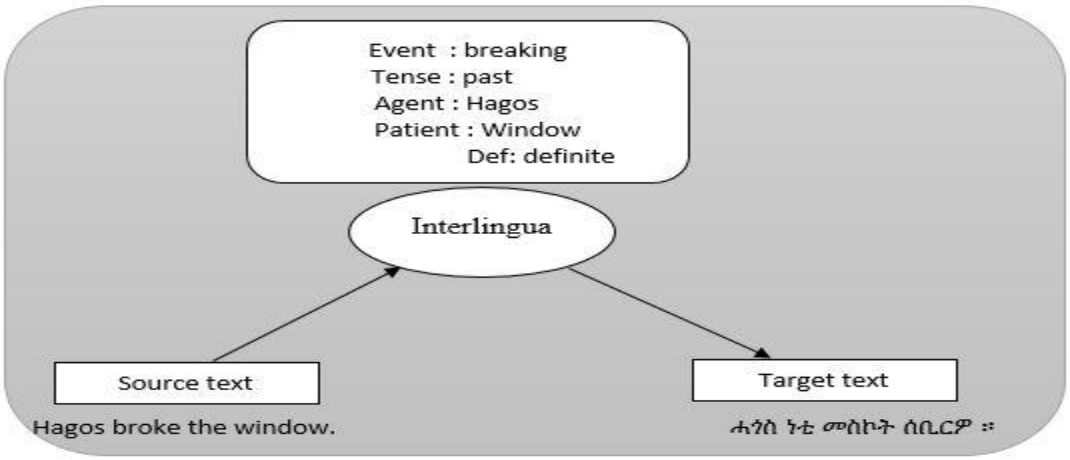


Figure 2.3: Interlingua model [22]

Figure 2.4 illustrates the general architecture of RBMT. Starting with the shallowest level at the bottom, direct transfer is made at the word level. Moving upward through syntactic and semantic transfer approaches, the translation occurs on representations of the source sentence structure and meaning respectively. Finally, at the Interlingua level, the notion of transfer is replaced with a single underlying representation the Interlingua that represents both the source and target texts simultaneously. Moving up the triangle reduces the amount of work required to traverse the gap between languages, at the cost of increasing the required amount of analysis (to convert the source input into a suitable pre-transfer representation) and synthesis (to convert the post-transfer representation into the final target surface form). For example, at the base of the triangle, languages can differ significantly in word order, requiring many permutations to achieve a good translation. However, a syntactic dependency structure expressing the source text may be converted more easily into a dependency structure for the target equivalent because the grammatical relations (subject, object, modifier) may be shared despite word order differences. Going further, a

semantic representation (Interlingua) for the source language may totally abstract away from the syntax of the language, so that it can be used as the basis for the target language sentence without change.

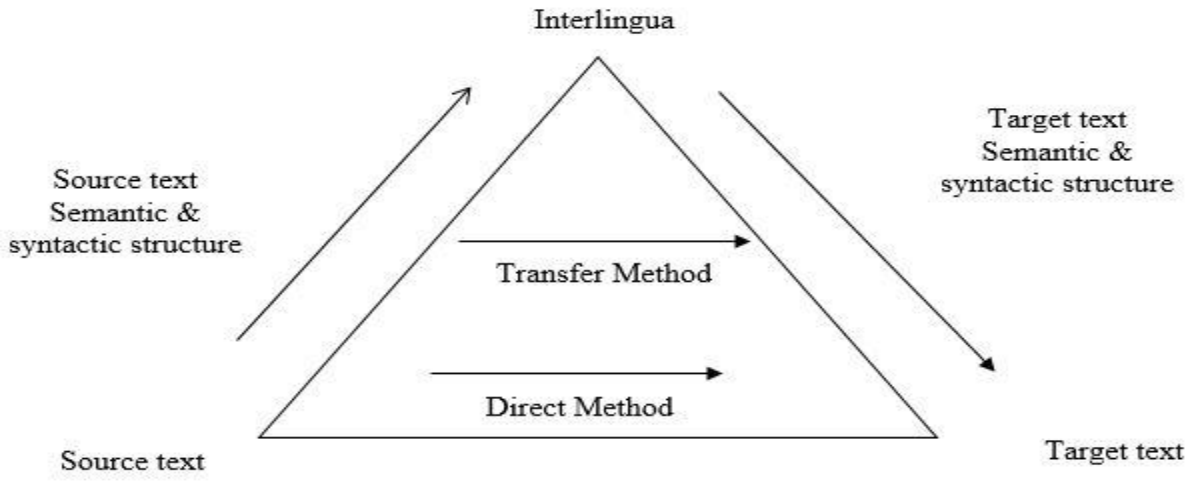


Figure 2.4: Different methods of RBMT adopted from [21]

2.2.2 Example Based Machine Translation (EBMT)

Example-based machine translation is characterized by its use of a bilingual corpus with parallel texts as its main knowledge, in which translation by analogy is the main idea. And the basic idea of this approach is to reuse examples of already existing translations as the basis for new translation [23]. The basic processes of EBMT are broken down into the following stages: alignment of texts which is the matching of input sentences against phrases (examples) in the corpus, selection and extraction of equivalent target language phrases, and the adaptation and combination of target language phrases as an acceptable output sentences. Alignment of text is used to identify which parts of the corresponding translation are to be reused. Alignment is done by using bilingual dictionary or comparing with other examples.

The matching stage in example based machine translation finds examples that contribute to the translation based on their similarity with the input. The way matching should be implemented is based on how the examples are stored, there are options in the processes of matching (character based, word based, structure based), measures of similarity (e.g. statistical and/or by reference to thesauri), the adaptation of extracted examples and their 'recombination' to produce target sentences. Recombining the reusable parts in example identified during alignment are putting

together in a genuine way, and the design of recombination strategy depends on previous matching and alignment strategy [20].

2.2.3 Statistical Based Machine Translation (SMT)

The idea of statistical machine translation was introduced by Warren Weaver in 1949 [24] . It is a machine translation process, sometimes referred to as Natural Language Processing which uses a bilingual data set and other language assets to build language and translation models used to translate text. Statistical machine translation deals with automatically mapping sentences in one human language (the source language) into another human language (the target language) [20]. Since statistical machine translation is corpus based, translation can be described as finding the most likely target sentence (such as Tigrigna) for some source sentence (such as English).

The initial model of SMT based on Bayes theorem, proposed by Peter et al. [25] takes the view that every sentence in one language is a possible translation of any sentence in the other and the most appropriate is the translation that is assigned the highest probability by the system. The idea behind SMT comes from information theory. A document is translated according to the probability distribution function indicated by p(t|e), which is the probability of translating a sentence e in the source language E (for example, English) to a sentence t in the target language T (for example, Tigrigna). The translation accuracy of these systems mainly depends on the domain, quantity and quality of a parallel corpus. So, in order to have a good translation quality, the data must be preprocessed consistently.

In statistical MT, every sentence in source language (such as English e) is a possible translation of a sentence in target language (such as Tigrigna t). Each Tigrigna sentence t is assigned a probability P(t|e). The task of translating source language (such as English sentence, e) is then to choose the target language such as (Tigrigna sentence, t) for which P(t̂|e) is the greatest, and this is based on noisy channel model (T for target sentence and E for source sentence) [24]:

$$\hat{T} = argmax\ P(T|E)$$

$$\hat{T} = argmax\ \frac{P(E|T)\ P(T)}{P(E)}$$

$$\hat{T} = argmax\ P(E|T)P(T) \quad \ldots\ldots\ldots\ldots \quad (2.1)$$

Where P(T|E) the translation model for T (such as Tigrigna) to E (such as English) translation

P(E|T) the translation model for E (such as English) to T (such as Tigrigna) translation

P(T) language model for T (such as Tigrigna)

The denominator P(E) can be ignored inside the argmax, because the aim is to choose the best Tigrigna sentence for a fixed English sentence E, and hence P(E) is a constant. The resulting noisy channel equation shows that two components are needed. These are a translation model P(E|T), and a language model P(T) [24]. The noisy channel model of statistical MT requires three components to translate a source sentence E to target sentence T:

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

2.2.3.1 Language Model

A statistical language model is a probabilistic way to capture regularities of a language, in the form of word-order constraints. In other words, a statistical language model expresses the likelihood that a sequence of words is a fluent sequence in a language. Plausible sequences of words are given high probabilities whereas nonsensical ones are given low probabilities. Statistical language modeling is important for different language technology applications such as speech recognition, machine translation, document classification, optical character recognition, information retrieval, handwriting recognition [26]. Axelrod and Amittai [27] describes different techniques of statistical language models, such as n-gram language model, factored language model, and syntactic language model. From this techniques n-gram model is the most widely used statistical language, and the factored based is an extension of this technique and their detail is presented below.

I. The N-gram Language Model

This is the leading technique in language modeling and the n-gram language models are based on statistics of how likely words are to follow each other. The n-gram model assumes that the nth word w depends only on the history h, which consists of the n−1 preceding words. The task of predicting the next word can be stated as attempting to estimate the probability function P [28]. By neglecting the leading terms, it models language as a Markov chain method:

$$P(wi|hi) = P\ (wi\ |w1, w2, \ldots, wi-1)$$

$$= P\ (wi|wi-N+1, \ldots, wi-1) \quad \ldots\ldots\ldots\ldots\ (2.2)$$

N-grams are essential in any task in which we have to identify words in noisy, ambiguous input. N-gram models are also essential in SMT. An N-gram will choose the most fluent translation sentence, i.e. the one that has the highest probability. The language model component is monolingual, and so acquiring training data is relatively easy [27]. SMT systems are based on the same N-gram language models as speech recognition and other applications [22].

II. Factored Language Model

Factored Language Model (FLM) is a newer language modeling technique more sophisticated than n-grams. Factored language models do not reduce the complexity of an n-gram model, but they use a richer set of conditioning possibilities to improve performance. The application of factored language models for machine translation is to layer linguistic information in the corpus, when available, on top of the power of statistical n-gram language models [29]. This model considers a word as a collection of features or factors, one of which may be the actual surface form of the word. In a factored language model, a word is viewed as a vector of k factors.

$$w_t = \{f_t^1, f_t^2, \ldots, f_t^k\}$$

Factors can be anything, including morphological classes, stems, roots, and others, such features found in highly inflected languages [29]. The probabilistic language model of the word can depend on the word itself and its factors [27]. For example, if POS is an additional factor to the language model, then the representation of words in "the black cat" would look like:

the = ("the", article)

black = ("black", adjective)

cat = ("cat", noun)

Therefore, the probability of the next factored feature being ("cat", noun) depends on its factored history: (("the", article) ("black", adjective)). In case of n-gram the trigram probability of the next word being "cat" depends only on its word history: ("the", "black").

The challenge of factored language model is to determine appropriate set of factors to represent the words for the betterment of the model. The simplest factored language model has only one factor per word, the word itself, which is equivalent to an n-gram language model. A two-factor factored language model generalizes standard class-based language models, where one factor is the word class and the second is the word's surface form [29].

2.3.4.2 Translation Model

Translation model deals with how likely a source sentence is translated into a target sentence. Most of state-of-the-art translation models used for regular text translation can be grouped into four categories [30] :

a. word-based models,
b. phrase-based models,
c. syntax-based model, and
d. factored based model

a. Word Based Translation Modeling

The translation model probability cannot be reliably calculated based on the sentences as a unit, due to sparseness. Instead, the sentences are decomposed into a sequence of words. Word-based models use words as translation units. Peter et al [25] presents five word-based translation models of increasing complexity, namely IBM Model 1 to 5 and also IBM model 6 was proposed later. In the IBM models, words in the target sentence are aligned with the word in the source sentence that generated them. These models are word-based because they model how individual words in source sentence are translated into words in target sentence. They translate the source sentence word for word into the target language and then reorder the words until they make the most sense [25].

b. Phrase-Based Statistical Machine Translation

The phrase-based models form the basis for most of state-of-the-art SMT systems. Phrase based translation models are based on the earlier word-based models. Like the word-based models, the phrase-based models are generative models that translate source language sentence into target language sentence. In phrase based models, the unit of translation is any contiguous sequence of words, which we call a phrase. The idea of phrase-based SMT is to use phrases (sequences of words) instead of single words as the fundamental units of translation. Each phrase in source

language is nonempty and translates to exactly one nonempty phrase in target language [31]. This is done using the following steps:

- ✓ The source sentence is segmented into phrases
- ✓ Each phrase is translated
- ✓ The translated phrases are permuted into final order

This set of steps govern the process contained in a phrase table, which is simply a list of all source phrases and all their translation. The phrase table is learned from the training data. The probability model for phrase-based translation relies on a translation probability and a distortion probability. The factor $\phi(T_i \mid E_j)$ is the translation probability of generating target phrase T_j from source phrase E_i. And the reordering of the target phrase is done by distortion probability. Distortion in SMT refers to a word having a different (distorted) position in the target sentence than it had in the source sentence; it is thus a measure of the distance between the positions of a phrase in the languages [31].

The limitation of phrase-based translation is it treats phrases as sequences of fully-inflected words and do not incorporate any additional linguistic information, they are limited in the following ways [18]:

- Phrase- based models are incapable to learn translations of words that do not occur in the data, because they are unable to generalize. Current approaches know nothing of morphology, and fail to connect different word forms. When a form of a word does not occur in the training data, current systems are unable to translate it. This problem is severe for languages which are highly inflective, and in cases where only small amounts of training data are available.
- They are unable to distinguish between different linguistic contexts. When current models have learned multiple possible translations for a particular word or phrase, the choice of which translation to use is guided by frequency information rather than by linguistic information.

c. Syntax-Based Model

Syntax based translation is based on the idea of translating syntactic units, rather than single words or strings of words [30]. The syntactic structure differs from language to language these different

syntactic structure require the reordering of words and the insertion and deletion of function words during translation, and this makes it difficult for machine translation. For example, re-ordering is needed when translating between languages with subject-verb-object sentence structure such as English, and languages with subject-object-verb structure such as Tigrigna. And it is possible to use the syntax information in the input language, output language or in both languages by constructing trees to model the sentence structure [30].

Syntax-based models of statistical machine translation are based on models by Yamada and Knight [32]. And the model transforms a source-language parse tree into a target-language string by applying stochastic operations at each node. These operations capture linguistic differences such as word order and case marking. The model can make use of syntactic information and performs better for language pairs with different word orders and case marking schema.

d. Factored Based Translation Model

Factored translation models, which extends traditional phrase-based statistical machine translation models to take advantage from additional annotation, especially linguistic markup. The basic idea behind factored translation models is to represent phrases not simply as sequences of fully inflected words, but instead as sequences containing multiple levels of information. A word in factored model is not a single token, but vector of factors. This enables straight-forward integration of part-of-speech tags, morphological information, and even shallow syntax. Such information has been shown to be valuable when it is integrated into pre-processing or post-processing steps. An integration of linguistic information into the translation model is desirable for two reasons [18]:

- Translation models that operate on more general representations, such as lemmas instead of surface forms of words, can draw on richer statistics and overcome data sparseness problems caused by limited training data.
- Many aspects of translation can be best explained on a morphological, syntactic, or semantic level.

The goal of factored model is to advance statistical machine translation by adding different information. These information are integrated into both training data and models. The parallel corpora used to train Factored Translation Models include factors such as parts of speech and lemmas. For example, a configuration of a factored translation model which employs translation

steps between lemmas and POS + morphology, and generation step from the POS + morphology and lemma to the fully inflected words [33]

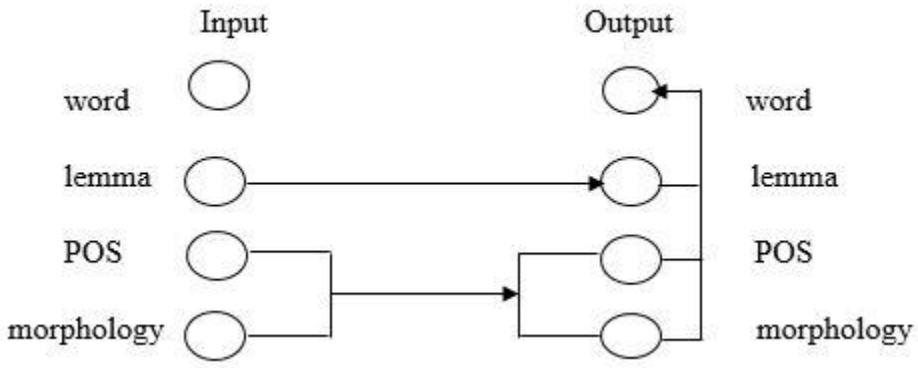


Figure 2.5: Factored translation model; adopted from [33]

Generally, the use of factors introduces several advantages over current phrase-based approaches such as: better handling of morphology, linguistic context can facilitate better decisions when selecting among translations, and linguistic markup of the training data allows for many new modeling possibilities [18].

2.3.4.3 Decoder

A good decoding algorithm is critical to the success of any statistical machine translation system. The decoder's job is to find the translation that is most likely according to set of previously learned parameters (product of the translation and language model). Decoders in MT are based on best-first search, a kind of heuristic or informed search; these are search algorithms that are informed by knowledge from the problem domain. Best-first search algorithms select a node n in the search space to explore based on an evaluation function f(n). MT decoders are variants of a specific kind of best-first search called A* search [22].

A* decoding algorithm is a kind of best-first search which was first introduced in the domain of speech recognition. The basic intuition is to maintain a priority queue which is traditionally referred to as a A* with all the partial translation hypotheses together with their scores. The job of a decoder is to find the highest scoring sentence in the target language (according to the translation model) corresponding to a given source sentence. It is also possible for the decoder to output a

ranked list of the translation candidates, and to supply various types of information about how it came to its decision [22].

A decoder searches for the best sequence of transformations that translates source sentence to the corresponding target sentence, decoding is the main part of the translation system. It looks up all translations of every source phrase, using phrase translation table and recombine the target language phrases that maximizes the translation model probability multiplied by the language model probability. At the end, sentence which covers all of the source words and which has the highest probability is chosen. For example if we want to translate a sentence **t** in the source language **T** (Tigrigna) to a sentence **e** in the target language **E** (English), the noisy channel model describes the situation in the following way [34]:

Suppose that the sentence **t** to be translated was initially conceived in language **E** as some sentence **e**. During communication **e** was corrupted by the channel to **t**. Now, assuming that each sentence in **E** is a translation of a with some probability, and the sentence that we choose as the translation ($\overline{e}$) is the one that has the highest probability. In mathematical terms [34],

$$\overline{e} = \underset{e}{\operatorname{argmax}}\, p(e|t) \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.3)$$

Intuitively, $p(e|t)$ should depend on two factors:

- ✓ The kind of sentences that is likely in the language **E** which is known as the language model, $p(e)$.
- ✓ The way sentences in **E** get converted to sentences in **T** which is called the translation model, $p(t|e)$.

Baye's rule states the following [22]:

$$p(e|t) = \frac{p(t|e) * p(e)}{p(t)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.4)$$

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

$$\operatorname{argmax} p(\mathrm{e}|\mathrm{t}) = \underset{e}{\operatorname{argmax}} \frac{p(t|e) * p(e)}{(t)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.5)$$

Since t is fixed, we ignore it from the equation then we get the noisy channel equation, which is

$$\overline{t} = \underset{e}{\operatorname{argmax}}\, p(t|e) * p(e)$$

- taking Tigrigna as input and display English as an output.

where $p(t|e)$is the translation model and $p(e)$ is the language model**.**

$$\overline{e} = \underset{t}{\operatorname{argmax}}\, p(e|t) * p(t)$$

- taking English as input and display Tigrigna as an output.

where **e** is the source text and **t** is the target

$p(e|t)$ is the translation model and $p(t)$ is the language model**.**

2.2.4 Hybrid Based Machine Translation (HMT)

Hybrid systems have been developed by combining the positive sides of each of the approaches, it takes the synergy effect of rule based, statistical and example-based MT. At present, several governmental and private based MT sectors use this hybrid based approach to develop translation from source to target language, which is based on both rules and statistics. The hybrid approach can be used in several different ways. In some cases, translations are performed in the first stage using a rule based approach followed by adjusting or correcting the output using statistical or example based information. In the other way, rules are used to pre-process the input data as well as post-process the statistical output of a statistical-based translation system or the matching output of the example based system [20].

2.4 Text Alignment

Alignment is the arrangement of something in an orderly manner in relation to something else. It can be performed at different levels, from paragraphs, sentences, segments, words and characters. Alignment models are used in statistical machine translation to determine translational correspondences between the words and phrases in a sentence of one language with the words and phrases in a sentence with the same meaning in a different language. They form an important part of the translation process, as they are used to produce word-aligned parallel text which is used to initialize machine translation systems. Improving the quality of alignment leads to systems which model translation more accurately and an improved quality of output.

The objective of alignment modeling is to define a correspondence between the words of source sentence and target sentence, given a source sentence $e = e_1^i$ and target sentence $t = t_1^j$. The number of possible alignments between t and s is $2^{|t||e|}$ and increases exponentially as the sentence lengths increase. This poses a problem for modelling due to the complexity of the models involved. The quality of alignment is crucial to the quality of the phrases extracted from that alignment and overall quality of the translation system [35].

All statistical translation models are based on the idea of a word alignment. A word alignment is a mapping between the source words and the target words in a set of parallel sentences. For phrase-based SMT, the alignment algorithms are used just to find the best alignment for a sentence pair (S, T), in order to help extract a set of phrases [22]. There are different alignment models used in SMT such as IBM alignment and HMM alignment. The following is a brief account of some of them.:

2.4.1 IBM Alignment

IBM alignment models are a sequence of increasingly complex models used in statistical machine translation to train a translation model and an alignment model, starting with lexical translation probabilities and moving to reordering. Peter et al. [25] develops a series of five translation models together with the algorithms necessary to estimate their parameters, numbered from 1 to 5 in increasing order of complexity and corresponding number of parameters. Training is carried out using a corpus of sentence-aligned parallel text. The models are trained in order, using Model 1 to provide an initial estimate of the parameters of Model 2, using Model 2 to provide an initial estimate of the parameters of Model 3, and using Model 3 to provide an initial estimate of the parameters of Model 4 and so on. Model 1 and Model 2 operate by selecting, for each target word, a source word to which it aligns. The remainder of the models, which are model 3, 4, 5 are fertility-based models and Model 6 is by combining Model 4 with HMM alignment model in a log linear way [35].

2.4.2 HMM Alignment

HMM alignment model is based on the HMM model. The idea of the model is to make the alignment probabilities dependent on the differences in the alignment positions rather than on the absolute positions. Vogal et al. [36] proposes Hidden Markov Model (HMM) word alignment in statistical translation. The characteristic feature of this approach is to make the alignment

probabilities explicitly dependent on the alignment position of the previous word. HMM model for word-to-word alignment of sentences of parallel text, where the words of the source sentence are viewed as states of the HMM and produce target words according to a probability distribution. The sentence-level alignment model is effectively split into two components: the alignment component models the transitions between the states of the HMM and the translation component controls the emission of target words from those states [35].

2.5 Evaluation of Machine Translation

The evaluation of machine translation systems is a vital field of research, both for determining the effectiveness of existing MT systems and for optimizing the performance of MT systems. Traditionally, machine translation evaluation paradigms are: Glass Box and Black Box evaluation [37]. Glass box focuses upon an examination of the system's linguistic aspects, which measures the quality of a system based upon internal system properties. Black Box evaluation, on the other hand measures the quality of a system based solely upon its output, regardless of the internal mechanisms of the translation system.

Machine translation systems can be evaluated by using human evaluation method or automatic evaluation method. Automatic machine translation evaluation metrics were developed due to the high costs, lack of repeatability, subjectivity, and slowness of human evaluation. These automatic measures generally judge the quality of machine translation output by comparing the system output, also referred to as candidates or hypotheses, against a set of reference translations of the source data. And the evaluation metrics are Word Error Rate, METEOR (Metric for Evaluation of Translation with Explicit ORdering), and BLEU which is the current standard for automatic machine translation evaluation. BLEU is a precision-oriented metric in that it measures how much of the system output is correct, rather than measuring whether the references are fully reproduced in the system output [37]. And these metrics differ on what counts as translation closeness. The quality of these evaluation metrics is usually measured by determining the correlation of the scores assigned by the evaluation metrics to scores assigned by a human evaluation metric, most commonly fluency and adequacy.

2.5.1 BLEU Score

The BiLingual Evaluation Understudy (BLEU) score is a metric widely used for automatic evaluation of machine translation output. The basic premise is that a translation of a piece of text

is better if it is close to a high-quality translation produced by a professional translator. The translation hypothesis is compared to the reference translation, or multiple reference translations, by counting how many of the n-grams in the hypothesis appear in the reference sentences, better translations will have a larger number of matches. The ranking of sentences using BLEU score has been found to closely approximate human judgement in assessing translation quality, and this is computed based on modified n-gram precision. For a candidate sentence e, unigram precision is calculated by finding the number of words in the candidate sentence that occur in any reference transcription and dividing by the total number of words in the candidate sentence [35].

$$\text{unigram precision(e)} = \frac{\sum_{w \in e} c(e,w) 1\{w \text{ occurs in some reference } e_r\}}{\sum_{w \in e} c(e,w)} \quad \ldots\ldots\ldots\ldots \quad (2.6)$$

However, this is not an accurate measure of translation quality as the system can generate many words that occur in the reference but not output grammatical or meaningful sentences. To overcome this, clip the count of the number of times a word occurs so that it is not counted if it occurs more than the maximum number of times it occurs in any reference sentence. For words in the candidate sentence e, define:

$$c_{max}(w) = \max_{e_r \in R(e)} c(e_r, w) \quad \ldots\ldots\ldots\ldots\ldots \quad (2.7)$$

$$c_{clip}(e,w) = \min(c_{max}(w), c(e,w)) \quad \ldots\ldots\ldots\ldots \quad (2.8)$$

The modified unigram precision for a sentence e is given by:

$$p1(e) = \frac{\sum_{w \in e} c_{clip}(e,w)}{\sum_{w \in e} c\,(e,w)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.9)$$

Similarly, for an n-gram $w_1^n = w_1 \ldots\ldots w_n$, define $c(e, w_1^n)$ to be the number of times the n-gram appears in a sentence and define

$$c_{max}(w_1^n) = \max_{e_r \in R(e)} c(e_r, w_1^n)$$

$$c_{clip}(e, w_1^n) = \min(c_{max}(w_1^n), c(e, w_1^n))$$

The modified n-gram precision is:

$$p_n(e) = \frac{\sum_{n-grams\ w_1^n \in e} c_{clip}(e, w_1^n)}{\sum_{n-grams\ w_1^n \in e} c\,(e, w_1^n)} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.10)$$

And the modified n-gram precision for the whole corpus is C is given by:

$$p_n(C) = \frac{\sum_{e \in C} \sum_{n-grams\ w_1^n \in e} c_{clip}(e, w_1^n)}{\sum_{e \in C} \sum_{n-grams\ w_1^n \in e} c(e, w_1^n)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.11)$$

2.6 Related Works

Many researches have been conducted so far for different languages to simplify the barrier among the languages. Some of the studies are given bellow:

2.6.1 German – English and French – English Statistical Machine Translation

Sereewattana [38], reported unsupervised approach to segment German – English and French - English parallel corpora for statistical machine translation. The segmentation of two parallel corpora, German -English and French – English, has shown to successfully improve the overall translation quality. Moreover, the unsupervised approach presented requires no language nor domain-specific knowledge in order to achieve higher translation accuracy. The segmentation helps to effectively reduce the number of unknown words and singletons in the corpora which helps improve the translation model. As a result, word error rates are lowered by 0.37% and 2.15% in the translation of German to English and French to English respectively. The benefits of segmentation to statistical machine translation are more pronounced when the training data size is small.

In the German – English translation, most German compounds are distinct and do not appear frequently enough to be learned correctly by a translation model. By breaking up compounds into regular words that are consistent with the data, the number of single compounds is lessened. Moreover, the segmented German data also promote a one-to-one relationship between German and English words in the alignment model which ultimately results in a better translation quality. And also, segmentation has proved to lower the French – English translation word error rate especially when the training corpus size is limited. When more training data becomes available, however, segmentation has shown to hurt the overall quality.

Finally the paper recommends, the application of segmentation to other translation pairs is expected to yield a similar result. Languages with rich compounding are expected to most benefit from segmentation because the segmented data improves the alignment of words in the source and target sentences. And without losing its language and domain independence, the unsupervised

segmentation for statistical machine translation presented in the paper can be enhanced in several ways. Such as: developing a synthesizer that is more context sensitive, improving the translation model by dividing the translation into a morphological and a lexical translation subtask, and using a named entity recognizer to exclude words to segment and translate.

2.6.2 English-to-Arabic Statistical Machine Translation

The study English to Arabic statistical machine translation is conducted by [39] using segmentation technique. The paper discuss morphological decomposition of Arabic language and Factored Translation Models for English to Arabic translation. The languages are from different family, Arabic has a complex morphology compared to English. Words are inflected for gender, number, and sometimes grammatical case, and various clitics can attach to word stems. An Arabic corpus will therefore have more surface forms than an English corpus of the same size, and will also be more sparsely populated. These factors adversely affect the performance of Arabic to English Statistical Machine Translation (SMT). The data used for this experiment is from two domain text news, trained on a large corpus, and spoken travel conversation, trained on a significantly smaller corpus. For the news language domain, 4-grams for the baseline system and 6-grams for the segmented Arabic are implemented. The average sentence length is 33 for English, 25 for non-segmented Arabic and 36 for segmented Arabic.

For the spoken language domain, trigrams for the baseline system and a 4-grams for segmented Arabic. The average sentence length is 9 for English, 8 for Arabic, and 10 for segmented Arabic. The scores generated by BLUE score are generally lower than those of comparable Arabic-to-English systems, this is because of only one reference was used to evaluate translation quality. For the news corpus, the BLUE score yields: 26.44 for the baseline and 27.30 for the factored Models with tuning. Factored models perform better than the segmented approach with the large training corpus, although at a significantly higher cost in terms of time and required resources.

2.6.3 Hindi to English Statistical Machine Translation

Kunal et.al [40], describe a Hindi to English statistical machine translation system and improve over the baseline using multiple translation models. And the researchers uses multiple strategies for translation and choose one translation as the final output. The selection was done using prediction strategies which analyze the candidate translation and predict the best suited translation. They initially explored phrase based, hierarchal, and factored model, taking POS as factor, but the

result was unsatisfactory compared to the previous Hindi – English machine translation systems. And they implement the methodology by training a regression model on features that bear a high correlation with a better translation and the evaluation metric measure as the corresponding regression value. The regression model thus acquired is later utilized for guided selection of the translations from multiple models. And the system is trained over textual as well as syntactic features extracted from source and target of the aforementioned translations. The one having higher regression value i.e. correctness score, was taken to be the output of the system. The system shows significant improvement over the baseline systems for both automatic as well as human evaluations. And the proposed methodology is quite generic and can easily be extended to other language pairs as well.

2.6.4 English-to-Czech Factored Machine Translation

The study English to Czech machine translation is conducted by [41] using factored based. The paper discusses English-to-Czech multi-factor translation. The current state of the art for translation model is phrase based machine translation, but there are problems with translating to morphologically rich languages. Czech is a Slavic language with very rich morphology and relatively free word order. The researcher conducts the experiment with 55,676 pairs of sentences. And he uses two types of language model simultaneously: a 3-gram language model over word forms and a 7-gram language model over morphological tags.

The result of the experiment on English-to-Czech translation with morphological tags achieves significant improvement in BLEU score over the word forms, by explicit modelling of morphology and using a separate morphological language model to ensure the coherence. This shows that multi-factored models always outperform the baseline translation.

2.6.5 English-to-Thai Factored Machine Translation

The study English-Thai machine translation is conducted by [42]. The objective of the research is to find an appropriate combination of English factors to support English-Thai translation task. The size of the parallel corpus that used for the experiment is 319,155 sentence pairs and tools used for this experiment are: Giza++ for alignment, SRILM for language model, Moses for decoding. Since the work is factored based translation, feature factors are necessary, there are four English factors those are surface, lemma, POS, and combinatory categorical grammar (CCG). Various combinations of those factors were trained to build the most efficient model for translation, the

combinations are: surface(baseline), surface and POS, surface and CCG, lemma and POS, lemma and CCG, surface and POS and CCG, lemma and POS and CCG. And they measure the accuracy using BLEU score, the result shows that the factored model which consists of lemma and CCG produce the highest accuracy at 24.21% (1.05% superior than baseline).

2.6.6 English – Amharic Statistical Machine Translation

A preliminary experiments on English – Amharic statistical machine translation is conducted by [14]. The researchers presents preliminary experiment using statistical machine translation. Since it is statistical machine translation, bilingual corpus is prepared, the corpora collected consisted of raw English-Amharic corpus from the Parliament of the Federal Democratic Republic of Ethiopia. The researchers conduct preprocessing task in the parallel corpus. Some of the preprocesses include text conversion, trimming, sentence splitting, sentence aligning and tokenization. A total of 115 parliamentary documents have been processed out of 632 collected raw documents.

After dropping sentences with more than 200 tokens and sentences that do not have matching, 18,432 English sentences aligned with Amharic sentences have been retained and used for the experiment. The architecture proposed by the researchers includes the basic components of SMT: language model, translation model, decoder, and evaluation metrics. Tools used to implement the system are: SRILM for creating language model, GIZA++ for word alignment, Moses for decoding, and Perl scripts for pre-processing purposes. The researcher conducts the experiment with segmented and unsegmented text of the target language, and BLEU score of 35.32% for the baseline phrase-based, and 35.66 for the morpheme segmentation have been obtained. And the researcher improves the result using supervised segmented FSM-based with BLUE score of 41.07.

Eleni [13] conducted a study on bidirectional English-Amharic machine translation using constrained corpus. The main reason that motivates the researcher to conduct this research is the absence of an application that translates English texts into Amharic or the other way. And the objective of the paper is to design and develop a bidirectional English-Amharic machine translation using a constrained corpus. The research work was implemented using statistical machine translation approach. In the study, the researcher prepared two different corpora i.e. one for simple sentences and the other is for complex sentences. Since the translation is bidirectional, two language models were developed for each language. Two different experiments were

conducted for each corpus and the evaluation was performed using automatic and human evaluation methods. The testing sets are prepared from the training set.

The experiments are, one for the simple sentences and the result obtained from this using BLEU score has an average score of 82.22% and 90.59% accuracy for English to Amharic and Amharic to English translations, respectively. And using the other evaluation method, manual questionnaire method, accuracy of English to Amharic translation is 91% and accuracy of Amharic to English translation is 97%. For complex sentences the result obtained was a BLEU score of 73.38% for English to Amharic translation and 84.12% for Amharic to English translation and from questionnaire method, accuracy of English to Amharic translation is 87% and Amharic to English translation is 89%.

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

2.6.7 English – Afaan Oromo Machine Translation

Sisay [15] undertaken a study on English – Affan Oromo machine translation. The research has two main goals: one to test how far one can go with the available limited parallel corpus for the English – Afaan Oromo language pair and the applicability of existing statistical machine translation systems on this language pair. The second one is to identify the challenges that need a solution regarding the language pair. The architecture of the proposed system includes four components of SMT: language modeling, translation model, Decoding and Evaluation.

The researcher prepared parallel corpus from different domains including Affan Oromo versions of some chapters of the Bible, the United Nation's Declaration of Human Rights, the Kenyan Refugee Act, the Ethiopian Constitution, some medical documents, the proclamations and the regional constitution of Oromia Regional State. The documents were preprocessed by using different scripts which are customized to handle some special behaviors of the target language such as apostrophe, and list of the abbreviations for Affan Oromo that is used for tokenization. Sentence alignment was done manually. The researcher used 20k bilingual sentences and 62,300 monolingual sentences for training and testing of the system. Tools used for implementation of this system are: SRILM toolkit for language modeling, GIZA++ for word alignment, and decoding was done using Moses. The experiment yields a BLEU score of 17.74%.

Jabesa [16] further conducted a study on bidirectional English – Affan Oromo machine translation system using a hybrid of rule based and statistical approaches. The study has two main goals: the first one is to apply a hybrid of rule based and statistical systems on English – Afaan Oromo language pair and vice versa by using available parallel corpus and the second one is to identify the better approach from statistical, and a hybrid of rule based and statistical approaches. The researchers implemented syntactic reordering, because English and Affan Oromo have different sentence structure with the purpose of making the structure of source sentences similar to the structure of target sentences. So, reordering rules are developed for simple, interrogative and complex English and Afaan Oromo sentences. The reordering rules are applied on both the training and test sets in a preprocessing step. Since the translation is bidirectional the researcher develops two language model one for English and the other for Affan Oromo. Tools used to implement the system are: IRSTLM toolkit for language modeling, GIZA++ for alignment, Moses for decoding, and Python programming language for the rule part. Two different experiments were conducted for each approach, one for the statistical approach and the result obtained from the BLEU score is 41.5% from Affan Oromo – English translation and 32.39% from English - Affan Oromo translation. The second experiment on a hybrid of rule based and statistical, the result returns from the BLEU score is 52.02% and 37.41% for Affan Oromo - English and English - Affan Oromo translations, respectively. From the result, a hybrid of rule based and statistical approach is better than the statistical approach for the language pair and a better translation is acquired when Afaan Oromo is used as a source language and English is used as a target language.

2.6.8 Summary

Generally, almost all of the studies have been done using statistical machine translation approach. In most of the studies, English is used as a source language and statistical approach is used as a base translation system. For example, in English to Thai and English to Czech factored Based Statistical Machine Translation, they have used linguistic information in order to improve the performance of the translation done by using statistical approach. In most of the studies, different NLP tools such as IRSTLM, SRILM, GIZA++ and Moses are used for training and decoding purpose and BLEU score is used for the evaluation purpose. In this research work, we have used parallel corpus collected from different domains, statistical approach, and morphological segmentation have been used. To the best knowledge of the researchers, there is no prior work on Tigrigna – English or English – Tigrigna machine translation. The research is different from the ones that are done on local languages not only because of the language pairs, but also in the experiments. Most of the works focused on phrase based translation, but this research includes two

different techniques in addition to the phrase based translation, the techniques are; first experiment based on morphemes obtained using unsupervised method and then post-editing the output of the unsupervised segmenter. These makes our research different from the other ones.

CHAPTER THREE

The Structure of Tigrigna and English languages

3.1 Overview of Tigrigna Language

Tigrigna language is one of the Ethio-Semitic languages which belong to Afro-Asiatic super family. Although it is mainly spoken in Tigray region of Ethiopia and central Eritrea, there are also immigrant communities who speak Tigrigna language in the world such as: Sudan, Saudi Arabia, USA, Canada, Germany, Italy, UK, Sweden as well as in other countries [5]. According to the 2007 G.C population and housing census of Ethiopia, there are over 4.3 million Tigrigna speakers in Tigray regional state [43] and according to Ethnologue [44] there are 2.4 million Tigrigna speakers in Eritrea. As a result, there are around seven million Tigrigna language speakers worldwide. Tigrigna language is written using Ge'ez script, which is called Ethiopic and it is originally developed for Geez language. In writing system, each symbol represents a consonant and vowel combination and the symbols are organized in groups of similar symbols on the basis of both the consonant and the vowel [45].

3.1.1 The Writing System

A writing system is any conventional method of visually representing verbal communication. While both writing and speech are useful in conveying messages, writing differs in also being reliable form of information storage and transfer. The Tigrigna writing system uses Ge'ez syllable or alphabet called Fidel(ፊደል) meaning letter, which was adapted from Ge'ez, the extinct classical language of Ethiopia. Unlike Latin language, the script is far more complex because every first letter has six suffixes. And it has thirty-five base symbols with seven orders which represent seven vowels for each base symbol [46]. Some of the characters are described below:

ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ

Figure 3.1 Some Fidels (alphabet) of Tigrigna languages

And the full list of characters are given in Appendix I.

3.1.1.1 Numbering and Punctuation Mark

Tigrigna number system uses Ge'ez numbering systems. It has twenty characters. They represent numbers from one to ten (፩-፲), twenty to ninety (፳-፺), hundred (፻) and thousand (፼). However, these are not suitable for arithmetic computation purposes because there is no representation for zero (0), decimal points.

In Tigrigna there are words representing numbers that are called Numerals. The numerals can be classified as ordinal numbers and cardinal numbers [46]. The cardinal numbers are numbers like ሓደ(one), ክልተ (two), ሰለስተ (three), ዓሰርተ (ten), etc. and the corresponding ordinal numbers are ቀዳማይ (first), ካልኣይ (second), ሳልሳይ(third), ዓስራይ (tenth), etc…. There are also special numerals in Tigrigna that correspond to the English like: ፍርቂ (half), ርብዒ(quarter) ሲሶ (one third) etc. Punctuation marks are symbols that are used to organize and clarify the meaning of writing. Identifying punctuation marks is vital to know word demarcation for natural language processing.

Tigrigna Symbol	English Symbol	Purpose
፡	White space	To separate words
።	.	To show the end of the sentence
፣	,	To separate list
፤	;	To indicate a pause longer than a comma
?	?	end of question
!	!	sentence to symbolize the anger, surprise or excitement
-	-	to link the parts of a compound word or phrase.
“ “ or ‘ ‘	“ “	Used around direct speeches, quotations or to give emphasis to a word or phrase

Table 3.1 : Punctuation marks used in both languages

3.1.1 Tigrigna Language Parts of Speech

Tigrigna language has eight parts of speech (POS) these are: nouns, adjectives, verbs, adverbs, prepositions, pronouns, conjunctions and interjection. And these are further grouped as, nouns, adjectives, verbs, adverbs and prepositions due to the role and behavioral similarity between nouns and pronouns, conjunctions and prepositions, and because Interjection (exclamation) cannot stand independently [45]. Therefore, although all the POS is the characteristics of Tigrigna language, we

do not need to discuss the nouns, verbs, and adjective here (presented on the morphology section), and prepositions and conjunctions are discussed below.

3.1.1.1 Tigrigna Conjunction

Conjunctions are used to connect words, phrases or clauses. Conjunctions in Tigrigna are coordinating or subordinating. A coordinate conjunction is a word that (like "and") connects two words, phrase, or clause that are grammatically or logically parallel. By way of comparison, a subordinate conjunction is a word like "since" which connects two clauses but represents one of them as an explanation or cause of the other. The main clause is stated and the other clause is subordinated to it [46]. Here are some common coordinate conjunctions with their examples.

- ***ን "and"*** is the most common coordinate conjunction. It is placed at the end of each word that is to be joined. Thus in any use it must occur two or more times. For example:
  - ጠረቢዛን ወንበርን ኣምጽእ ። Bring a table and chair.
  - መጽሓፍካን ጥራዝካን ዉሰድ ። Take your book and your exercise book.
- ***ውን "and" (እውን*** after a word ending in a vowel) is another common coordinate conjunction.
  a. ውን and is used to bind two clauses together, and is placed after the word which is to be emphasized.
     - ናብ ቤት ትምህርቲ ከይዱ ኣብኡ እውን ኩዕሶ ተጻዊቱ ። He went to school and there he played ball.
     - ና ብ ቤት ትምህርቲ ከይዱ ኩዕሶ እውን ተጻዊቱ ። He went to school and played ball.
  b. It is sometimes used to mean "then"
     - ንጽበሒቱ እውን ናብ ገዛ ዋና እተን ኩበቲ ከይዱ ። Then, the next day, he went to the house of the owner of the cattle.
  c. It is used to mean "also"
     - ፈረሰይ'ውን ሞይቱ ። My horse also is dead.

The remaining coordinate conjunctions are listed in the following Table 3.1:

No.	Tigrigna Conjunctions	English equivalent
1	ን ን	And

2	ውን	And (እውን after a word ending in a vowel)
3	ከኣ	And, also
4	ድማ	And, also
5	ስጋብ	Until (also ክሳዕ) only with simple imperfect tense
6	ወይ	or
7	ኰነ	either
8	ኣሽንኳይ’ ዶ … ውን	Not only … but also
9	ግን	but
10	ደኣ	Better, rather, in preference to, sooner
11	እሞ	Therefore, in consequence
12	እኳ	Not even, even
13	ደጊም	Therefore
14	ብዝተረፈ	finally
15	ማለት	That is to say that means

Table 3.2: Summary of coordinate conjunctions and interjections

Subordinate conjunctions join two elements, but not in parallel. The two elements that are joined may differ grammatically to a greater or lesser extent, along with the logical and syntactic subordination.

No.	Tigrigna conjunctions	English equivalent	Construction
1	ስጋብ	until	ዝ + s.impf
2	ካብ or እንካብ	than	ዝ + s.impf
3	ካብ	Since, because	ዝ + s.impf
4	ክ	while	ክ with verb and ኣሎ
5	ክይ	Without, before	s.pf
6	ምእንቲ	In order to	ክ + s.impf
7	ስለ	because	ዝ + s.impf
8	ከም	that	ዝ + s.impf
9	ምንም እኳ … እንተ / ሽሕ እኳ… እንተ	Even if	s.pf
10	ኣብ ክንዲ	Instead of	ዝ + s.impf

11	ምስ	with	s.pf
12	እና	while	s.pf

Table 3.3: Summary of subordinate conjunctions

3.1.1.2 Tigrigna Prepositions

There are two kinds of prepositions, simple and compound [46]. Simple prepositions of one radical, e.g. ***ብ "by", ን "to"***, are prefixed to nouns, pronouns, and adjectives; those of two or more radicals, e.g. ***ካብ "from", ብዘይ "without"***, are written as separate words. Compound prepositions consist of simple prepositions (usually ኣብ) plus another word, e.g. ***ኣብ ጥቃ "nearby"***, ***ኣብ ቅድሚ "in front"***, ***ኣብ ውሽጢ "inside"***, ***ኣብ ልዕሊ "above"***. For example:

Examples of simple prepositions with one radical:

- ብ "by" (mainly instrumental, but notice the examples)
  - ብበቅሊ መጺኡ ። He came by mule.
  - ብክልተ ስዓት ክወድኦ እየ ። I will finish it in two hours.
  - ብሞኾኒ ኣቢሎም ናብ መቀለ አትዮም ። They came to Mekelle by way of Mekoni.
- ን "for" (to the advantage or disadvantage of something or someone)
  - እቲ ምግቢ ንሽልቢ እዩ ። The food is for the dog.
  - ጴጥሮስ ንሓው ሃሪሙዎ ። Peter struck his brother.

Examples of simple prepositions with two or more radicals:

- ምስ "with" (signifies association)
  - ዳዊት ምስ ዮሃንስ መጺኡ ። Dawit came with Yohannes.
- ቅድሚ "ago", "before" (temporal)
  - ቅድሚ ክልተ ስዓት ሓወይ መጺኡ ። My brother came two hours ago.
  - ጴጥሮስ ናብ መቃብር ቅድሚ ዮሃንስ አተወ ። Peter entered the tomb before John.
- ካብ "from"
  - ካብ ከተማ መጺኣ ። She came from the town
- ናይ "of" (possessive)
  - እቲ መጽሓፍ ናይ መምህር እዩ ። It is the teacher's book.
- ኣብ "in", "at"
  - ኣብ ቤት ጽሕፈት ኣሎ ። He's at the office.
- ብዘይ "without"

- ብዘይ ገንዘብ ዕዳጋ ከይዶም ። They went to the market without money.

➢ ናብ “to”, “toward”

- ናብ ሩባ ከይዱ ። He went to the river.

Examples of compound prepositions:

➢ ኣብ ቅድሚ “in front of”

- ኣብቲ ቅድሚ ገዛ ኦም ኣሎ ። In front of the house is a tree.
- ኣብ ቅድሚ ባንክ ንራኽብ ። Let’s meet in front of the bank.

➢ ኣብ ውሽጢ “inside”

- እቲ ወንበር ኣብ ውሽጢ ገዛ ኣሎ ። The chair is inside the house.

➢ ኣብ ጥቃ “beside”, “near”

- እቲ ብርዒ ኣብ ጥቃ እቲ መጽሓፍ ኣሎ ። The pen is beside the book

➢ ኣብ ልዕሊ “above”, “over”, “on”, “upon”

- በራድ ኣብ ልዕሊ እቶን ኣሎ ። The kettle is on the stove

➢ ኣብ ትሕቲ “under”, “beneath”, “below”

- እቲ ቆልዓ ኣብ ትሕቲ ዓራት ኣሎ ። The child is under the bed

➢ ኣብ ድሕሪ “behind”, “outside”

- እቲ ዒላ ኣብ ድሕሪ ገዛ ኣሎ ። The well is behind the house

3.1.2 Morphology of Tigrigna Language

As it is clearly stated by [44] [47], Tigrigna exhibits a root pattern morphological phenomenon like other Semitic languages such as Arabic and Amharic. Tigrigna words are created in different forms such as: from a root of three radicals (which are called tri-lateral words), quad-literal, pen-literal, and hexa-literal words. Words in Tigrigna are built from the roots by means of a variety of morphological operations such as compounding, affixation, and reduplication.

As every language has its own affixes and these are placed in different positions to produce the variants of the language, Tigrigna affixes are morphemes that can be added at the beginning, middle or end of the root to create the inflectional or derivational Tigrigna words. These words may be new in meaning and structure from their respective roots. Since Tigrigna affixes can also concatenate with each other, this increases the number of affixes in the language [45] Tigrigna affixes can be classified in to four categories:

Category	Description	Example
Prefixes	precede the base form	ን፣ ዝ፣ እንተ፣ ም፣ ብም
Suffixes	follow the base form	ና፣ ታት፣ ት፣ ነት፣ ን
Infixes	inside the base form	ባ ፣ላ ፣ ታ
Circumfixes	attached before and after the base form at the same time.	ን and ታት in ንክተማታት

Table 3.4 Tigrigna affixes

3.1.3 Derivational Morphology of Tigrigna Language

Derivational Morphology is a morphology concerned with the way in which words are derived from morphemes through processes such as affixation or compounding. This derivation process usually changes the part-of-speech category. Derivational morphology is the characteristics of Tigrigna language, nouns, adjectives and verbs [45]. The derivation of verbs, nouns, and adjectives are presented below [44].

Derivation of Verbs: Almost all Tigrigna verbs are derived from root consonants. Traditionally a distinction is made between simple and derived verbs. Simple verbs are those verbs derived from roots by intercalating vowel patterns whereas derived verbs are considered as derivatives of simple verbs. Simple verbs are those verbs derived from roots by intercalating vowel patterns whereas derived verbs are considered as derivatives of simple verbs. The derivation process can be internal or external, and this is found in the form of causative, passive, repetitive and reciprocal verbs.

Causative verbs are derived by adding the derivational morphemes ‘a- and to the verb stem as in the examples በፅሐ /’beSHe’/ ‘arrive’ - ኣበፅሐ /’abSHe’/ ’cause to arrive’ and ወሰደ /’wesed’/ ’take’ ኣውሰደ /’awsede’/ ’cause to take’. In most cases the ‘a- morpheme is used to form causative of intransitive verbs, transitive ones and verbs of state. Some exceptions are the verbs that begin with ‘a’, always take the morpheme ‘a’ but add the morpheme ‘I’ after the morpheme ‘a’ to form causative e.g. ኣሰረ /’asere’/ ‘arrest’, ኣእሰረ /’aIsere’/ ‘cause to arrest’.

Passive verbs are derived using the derivational morpheme ተ /te/. This derivational morpheme is realized as ተ-/te-/ before consonants and as ት-/t-/ before vowels. Moreover, in the imperfect, jussive and in derived nominals like verbal noun, the derivational morpheme ት/t-/ is used. In this case, it assimilates to the first consonant of the verb stem, and as a result, the first radical of the

verb geminates. Some exceptions are intransitive verbs like ፈሊሑ /faliHu/ 'it boiled' that form their passive forms using the prefix ተ- /te-/ as in ተፈሊሑ /tafeliHu/ 'it was boiled'. Such kind of verbs can derive their passive from their causative form (ኣፍሊሑ/afliHu/'he boiled').

Repetitive verbs indicate an action which is performed repeatedly. For tri-radical verbs, such verbs are formed by duplicating the second consonant of the root and using the ኣ-/'a-‘/ after the duplicated consonant as in ሰባበረ /seba-bere/ 'he broke repeatedly' derived from the root ሰብር/sbr/ break. All verb types have the same reduplicative/ repetitive forms.

Reciprocal verbs are derived by prefixing the derivational morpheme ተ- /te-/ either to the derived form (that use the vowel a after the first radical) or to the reduplicative stem. For example, reciprocal forms of ተቃተሉ /teqatelu/ 'killed each other' and ተቀታተሉ /teqetatelu/ 'killed one another' are derived from the derived stem ቃተሉ /'qatelu'/- and reduplicative stem ቀታተሉ /'qetatelu'/, respectively.

Derivation of Nouns: Tigrigna nouns can be either primary or derived. They are derived if they are related in their root consonants and/or meaning to verbs, adjectives, or other nouns. Otherwise, they are primary. Nouns are derived from other nouns, adjectives, roots, stems, and the infinitive form of a verb by affixation and intercalation. The morphemes -ነት/-net/, -ኢት/-`it/, -ኣት/-`at/,-ኡት/ut/, -ትኦ/-to/, -ኦ/o/, -ኢ/i/, -ኣ/a/, -ኣን/an/, -ኣኛ/-‘eNa/, -ኛ/-Na/, -ኣት/-et/, ኣዊ/-awi/, -ተኛ/-teNa/, -ና/-na/ and the prefix መ-/me-/ are used to derive nouns from other nouns. From the adjectives, nouns can be derived using the suffixes /net/ and /-et/ as in the examples መርዛማነት /'merzamanet'/ 'generosity' which is derived from the adjective መርዛም /'merzam'/ 'poison' and ፍልጠት /'flTet'/ 'knowledge' from the adjective ፉሉጥ /'fluT'/ 'known'.

In Tigrigna, nouns can also be formed through compounding. For example, ቤትብልዒ /' betbl`i '/ ‘restuarant’ is derived from the nouns ቤት /bet/ ‘house’ and ብልዒ /bl`i/ 'food'. As it can be seen, no morpheme is used to bind the two nouns. But, there are also compound nouns whose components came together by inserting the compounding morpheme ኣ/ `e/ as in ቤተክርስትያን /betekrstyan/ 'church' which is formed from ቤት/bet/ 'house' and ክርስትያን/krstyan/ 'Christian'.

Derivation of Adjective: Adjectives in Tigrigna include all the words that modify nouns. As it is true for nouns, adjectives can also be primary (such as ለዋህ (kind)) or derived, although the number of primary adjectives is very small. Adjectives are derived from nouns, stems or verbal roots by

adding a suffix and by intercalation. For example, it is possible to derive ሃፍታም/haftam/'rich, wealthy', ዘበናዊ/zebenawi/'modern' and ማእኸላይ/maKelay/ 'central' from the nouns ሃፍቲ/hafti/ 'wealth', ዘበን/zeben/ 'period' and ማእኸል /maKel/ 'center', respectively.

3.1.4 Inflectional Morphology of Tigrigna Language

Inflectional Morphology is a morphology that deals with the combination of a word with a morpheme, usually resulting in a word of the same class as the original stem, and serving same syntactic function. They do not change the part-of-speech category but the grammatical function. Since Tigrigna language is a highly inflectional language, a given root of a Tigrigna word can be found in different forms. The following is the description of the inflectional morphology feature of Tigrigna verbs, nouns and adjective since prepositions and adverbs are unproductive as it is stated in [47]

3.1.4.1 Inflection of Verbs

Tigrigna verbs have two tenses: perfect and imperfect. Perfect tense denotes actions completed, while imperfect denotes uncompleted actions. The imperfect tense has four moods: indicative, subjective, jussive, and imperative. Tigrigna verbs in perfect tense consist of a stem and a subject marker. The subject marker indicates the person, gender, and number of the subject. The form of a verb in perfect tense can have subject marker and pronoun suffix. The form of a subject-marker is determined together by the person, gender, and number of the subject. This is illustrated in Table 3.3 to indicate the subject marker and the pronoun.

Verb Variations	Gender	Person	Pronoun	Number
በሊ*ዐ*	Both	First	I - ኣነ	Singular
በሊዕና	Both	First	We -ንሕና	Plural
በሊ*ዐ*፡	Male	Third	He - ንሱ	Singular
በሊ*ያ*ም	Male	Third	They - ንሳቶም	Plural
በሊ*ዐ*ን	Female	Third	They - ንሳተን	Plural
በሊዒኩም	Male	Second	You - ንስኻትኩም	Plural
በሊዒኪ	Female	Second	You - ንስኺ	Singular
በሊዒካ	Male	Second	You - ንስኻ	Singular

Table 3.5 Inflections of perfective tense

As it can be seen from the above table, the suffixes attached are ኡ/u/, ኦም/om/, አን/en/, ኩም/kum/, ኪ/ki/, ካ/ka/, ና/na/, አ/ae/, ክን/kn/, ኣ/a/ and indicate person, gender and number of the subject and the pronoun that indicates the person.

Imperfect tense includes indicative, subjective, jussive and imperative forms and are inflected by adding affixes to indicate gender, person and number to the imperfective verb stem.

Person	Gender	Singular	Plural
Third	Female	ትበልዕ	ይበልዓ
Third	Male	ይበልዕ	ይበልዑ
First	Both	እበልዕ	ንበልዕ
Second	Male	ትበልዕ	ትበልዑ
Second	Female	ትሰርሒ	ትሰርሓ

Table 3.6 Inflections of imperfective tense

In the above table, ት(t), ይ(y), እ(i), and ን(n) are prefixes and ኢ(e), ኡ(u), and ኣ(a) are suffixes. Tigrigna verbs are also found in gerundive, jussive and imperative by employing affixes in addition to the root. The gerundive form is inflected by adding suffixes at the end of the gerundive verb to indicate person, gender and number. For example, ሰሪሐ፣, ሰሪሕኻ፣ ሰሪሕኺ፣ ሰሪሑ፣ ሰሪሓ፣ ሰሪሕና፣ ሰሪሕኹም፣ ሰሪሕኽን፣ ሰሪሖም፣ ሰሪሐን can show how the gerundive verb ሰሪሕ morphologically varies. Jussive and imperative verbs are sometimes called mood and jussive verbs are used to express a command for first and third persons whereas imperative verb is used to express second person in the singular and plural form [45].

3.1.4.2 Inflection of Nouns

Tigrigna nouns are also inflecting to show gender, person and number by adding affixes to the noun stem. Tigrigna language specifies two types of genders: Feminine and masculine. So, Tigrigna nouns are either male or female by nature. Therefore, in order nouns to express possession, pluralism, tribe and gender, affixes such as -ኣ/a/, ታት/-tat/, ኣት/-at/, ኣን/an/, ውቲ/-wti/, ቲ/-ti/, ዊ/wi/, ና/na/, etc. are used.. Example from [47]:

Noun Stem	Affix used	Noun after affixation	Remark
ላሕሚ	ኣ	ኣላሕም	Pluralism

ፊደል	ኣት	ፊደላት	Pluralism
ቐፅሊ	ታት	ቐፅልታት	Pluralism
መምህር	ኣን	መምህራን	Pluralism
ገዛ	ውቲ	ገዛውቲ	Pluralism
ኣድጊ	ና	ኣድጊና	Possession
ስራሕ	ቲ	ስራሕቲ	Pluralism
ኢትዮጵያ	ዊ	ኢትዮጵያዊ	Tribe
ሓሙ	ኣት	ሓማት	Gender

Table 3.7 Inflection of Nouns

3.1.4.3 Inflection of Adjectives

Tigrigna adjectives inflect for number and gender. This is to mean that Tigrigna adjectives can have singular masculine, singular feminine and the same adjective to express both masculine and feminine genders in the plural form [45]. The affixes that are used for the inflections of the adjectives are: ት/t/, ቲ/ti/, ኣዊት /awit/, ኣዊ/awi/, ኣት/at/, ኣዊያን/awyan/, etc.

Masculine	Feminine	Pluralization
ፅቡቕ	ፅብቕቲ	ፅቡቓት
ሰራሒ	ሰራሒት	ሰራሕቲ
መዐረይ	መዐረይት	መዐረይቲ
ቀታሊ	ቀታሊት	ቀተልቲ
ሰፋይ	ሰፋይት	ሰፈይቲ

Table 3.8 Inflection of Adjectives from [47]

3.1.5 Tigrigna Phrases Classification

A phrase is a sequence of two or more words arranged in a grammatical construction and acting as a unit in a sentence. In Tigrigna languages, phrases are categorized into five categories namely noun phrase, verb phrase, adjectival phrase, adverbial phrase and prepositional phrase [45]. Each phrase type can be categorized into simple and complex.

Noun Phrase (NP): A noun phrase is a syntactic unit in which the head (H) is a noun or a pronoun. It can be simple or complex. The simplest NP consists of single noun (e.g. Dawit) or pronoun such as he (ንሱ), she (ንሳ), and a complex NP can consist of a noun and other constituents (like complements, specifies, adverbial and adjectival modifiers) that modify the noun from different

aspects [45]. For example, ናይቲ ዝሓለፈ ቅነ ቆንጆ መኪና (that last week's beautiful car), is a complex NP composed of ; ናይቲ (that) – specifier, ዝሓለፈ ቅነ (last week's) – adverbial modifier, ቆንጆ (beautiful) –adjectival modifier, and መኪና (car) – noun.

Verb Phrase (VP): Verb is one classification of words in Tigrigna languages which describes different actions of the subject. A VP is a syntactic unit composed of at least one verb and its dependents objects, complements and other modifiers but not always including the subject. For example, 'ናብ ስራሕ ከይዳ'(she went to work) is a VP composed of; ናብ ስራሕ (to work) is prepositional phrase (PP) modifying the verb ከይዳ (went).

Adjectival Phrase (AdjP): The adjectival phrase of Tigrigna language is composed of an adjective (head), and other constituents such as complements, modifiers and specifies, similar to the NP and VP construction [45]. Its main objective is describing nouns, but also it can describe verbs. For example: እቲ ብጣዓሚ ቆንጆ (That very handsome (boy)) is AdjP composed of; እቲ (that) is a specifier, ብጣዓሚ (very) is a modofer modifying the head of the AdjP, ቆንጆ(handsome).

Prepositional Phrase (PP): Prepositional phrases are groups of words containing prepositions. PP is constructed from a preposition (P) head and other constituents such as nouns, noun phrases, verbs, verb phrases, etc., unlike the others the head word of the phrase cannot stand alone; at least there must be complements with it. For example, ዳዊት ምስ ሓው ነይሩ (Dawit was with his brother), is a PP constructed from ; ዳዊት (Dawit), is NP, ምስ (with) is preposition, ሓው (his brother) is NN.

Adverbial Phrase (AdvP): An Adverbial phrase is constructed from one or more adverbs in the language. For example ብጽኑዕ ሓሚማ (she is severely ill).

3.1.6 Tigrigna Sentence Structure

Language is a structured system of communication. Segments and features of Tigrigna form syllables which in turn form words. Tigrigna sentence, as in any other language, can be divided into two largest immediate constituents namely subject and predicate. The subject can be thought as a noun phrase which is used to mention some objects or persons and the predicate is used to say something true or false about the subject. These two elements of sentences are known as noun phrase (NP) and verb phrase (VP) which are headed as noun and verb respectively [45].

The sentence structure for Tigrigna language is a Subject-Object-Verb structure unlike English with a Subject-Verb-Object combination. Figure 3.1 shows an Example.

Figure 3.2 Structure of Tigrigna and English languages

Sentences in Tigrigna can be classified as simple sentence and complex/compound sentences, based on the number of verbs they contain [48]. Simple sentences have only one verb while complex sentences have more than one. For example:

- ✓ Simple sentence:

  ሓውካ መጺኡ። (your brother came)

- ✓ Complex sentence:

  እቲ ስልኪ ዝሰረቀ ስብኣይ ተኣሲሩ። (the man who stole a phone is jailed.)

The first sentence contains one noun which is "ሓውካ" (your brother) and one verb which is "መጺኡ" (came). And the second sentence contains two verbs: "ዝሰረቀ" (stole) and "ተኣሲሩ" (jailed), in other words, it contains a subordinate and main clause which have one verb each. A subordinate clause is a clause which is dependent to another independent clause called the main clause [48]. In this case the main clause is "እቲ ሰብኣይ ተኣሲሩ" (The man is jailed) and the subordinate clause is "ስልኪ ዝሰረቀ" (who stole phone).

3.2 Overview of English Language

English language is a language that was first spoken in Anglo-Saxon England in the early middle Ages. It is now the most widely used language in the world [9]. English language distinguishes seven major word classes: verbs, nouns, adjectives, adverbs, determiners (i.e. articles), prepositions, and conjunctions. Some analyses add pronouns as a class separate from nouns, and subdivide conjunctions into subordinators and coordinators, and add the class of interjections [49].

English forms new words from existing words or roots in its vocabulary through a variety of processes. One of the most productive processes in English is conversion using a word with a different grammatical role, for example using a noun as a verb or a verb as a noun. Another productive word-formation process is nominal compounding producing compound words such as babysitter or ice cream or homesick [50].

Nouns and Noun Phrases : English nouns are only inflected for number and possession. New nouns can be formed through derivation or compounding. They are semantically divided into proper nouns (names) and common nouns. Common nouns are in turn divided into concrete and abstract nouns, and grammatically into count nouns and mass nouns [49]. Most count nouns are inflected for plural number through the use of the plural suffix ***-s***, but a few nouns have irregular plural forms. Mass nouns can only be pluralized through the use of a count noun classifier, e.g. one loaf of bread, two loaves of bread.

Nouns can form noun phrases (NPs) where they are the syntactic head of the words that depend on them such as determiners, quantifiers, conjunctions or adjectives [51]. Noun phrases can be short, such as "the man", composed only of a determiner and a noun. They can also include modifiers such as adjectives (e.g. red, tall, all) and specifiers such as determiners (e.g. the, that). But they can also tie together several nouns into a single long NP, using conjunctions such as "and", or prepositions such as "with", e.g. " The tall man with the long red trousers and his skinny wife with the spectacles" (this NP uses conjunctions, prepositions, specifiers and modifiers).

Adjectives : Adjectives modify a noun by providing additional information about their referents. In English, adjectives come before the nouns they modify and after determiners. In English, adjectives are not inflected, and they do not agree in form with the noun they modify, as adjectives in most other Indo-European languages do [51]. For example, in the phrases " The slender boy", and "Many slender girls", the adjective **slender** does not change form to agree with either the number or gender of the noun. Some adjectives are inflected for degree of comparison, with the positive degree unmarked, the suffix -er marking the comparative, and -est marking the superlative: a small boy, the boy is smaller than the girl, that boy is the smallest. Some adjectives have irregular comparative and superlative forms, such as good, better, and best.

Pronouns, Case and Person : English pronouns conserve many traits of case and gender inflection. The personal pronouns retain a difference between subjective and objective case in most persons (I/me, he/him, she/her, we/us, they/them) as well as a gender distinction in the third person singular (distinguishing he/she/it) [49].

Prepositional Phrases (PP) are phrases composed of a preposition and one or more nouns [49], e.g. with the dog, for my friend, to school, in England. Prepositions have a wide range of uses in English. They are used to describe movement, place, and other relations between different entities,

but they also have many syntactic uses such as introducing complement clauses and oblique arguments of verbs. For example, in the phrase I gave it to him, the preposition to marks the recipient, or Indirect Object of the verb to give.

Verbs and Verb Phrases : English verbs are inflected for tense and aspect, and marked for agreement with third person singular subject. Only verb *to be* is inflected for agreement with the plural, and first and second person subjects. Auxiliary verbs such as *have* and *be* are paired with verbs in the infinitive, past, or progressive forms [51]. A phrasal verb is a phrase that indicates an action such as turn down or ran into. The term applies to two or three distinct but related constructions: a verb and a particle and/or a preposition together form a single semantic unit. This semantic unit cannot be understood based upon the meanings of the individual parts, but must be taken as a whole. In other words, the meaning is non-compositional and thus unpredictable. Phrasal verbs that include a preposition are known as prepositional verbs and phrasal verbs that include a particle are also known as particle verbs [51].

Adverbs: The function of adverbs is to modify the action or event described by the verb by providing additional information about the manner in which it occurs. Many adverbs are derived from adjectives with the suffix **-ly**, but not all, and many speakers tend to omit the suffix in the most commonly used adverbs [51]. For example, in the phrase " The woman walked quickly" the adverb **quickly** derived from the adjective **quick** describes the woman's way of walking. Some commonly used adjectives have irregular adverbial forms, such as good which has the adverbial form well.

The sentence structure for English is Subject-Verb-Object combination. The subject constituent precedes the verb and the object constituent follows it. The example below demonstrates how the grammatical roles of each constituent is marked only by the position relative to the verb:

The dog bites the man
S V O
The man bites the dog
S V O

3.3 Challenges of Tigrigna and English Language to SMT

From the above discussion we understand Tigrigna is morphologically rich language. The variation in morphology, writing system, affix, sentence structure and punctuation mark of the languages pose a problem in the application of NLP tools, resources and techniques of the Tigrigna – English machine translation. For example, Tigrigna and English have difference in their syntactic structure. In Tigrigna, the sentence structure is Subject-Object-Verb, for example if we take Tigrigna sentence "ሓጎስ መስኮቱ ሰቢሩዎ።" , "ሓጎስ " is the subject , "መስኮቱ" is the object and "ሰቢሩዎ" is the verb of the sentence. In case of English, the sentence structure is Subject-Verb-Object. For example, if the above Tigrigna sentence is translated into English it will be "Hagos broke the window", "Hagos" is the subject, "broke" is the verb and "window" is the subject of the sentence. This pose a problem in the translation it requires reordering.

The other issue is in morphology for example " I love you. " can be translated into two different translations. The first translation can be " የፍቅረካ እየ " (yefikireka eye) it is for men , and the second translation is " የፍቅረኪ እየ " (yefikireki eye) for female. And also there is a challenge in the structure of preposition and conjunctions of the language pairs; some prepositions and conjunctions in Tigrigna are found in affixed form but their corresponding meaning found independently for example if we take Tigrigna sentence : "ብበቅሊ መጺኡ ። " the preposition " ብ- " is attached to the word " በቅሊ" . In case of English the sentence is translated into "He came by mule." the corresponding preposition " by " found independently. This kind of issues are also happened in case of conjunctions. And these difference also affect one-to-one alignment in the translation modeling.

In general, the nature of Tigrigna compared to English has unique characteristics. Tigrigna is rich in morphology; a given headword can be found in huge number of different forms; because a given word can be distributed in the Tigrigna vocabulary, and this leads to data sparseness and out-of-vocabulary problem. The other problem is the Tigrigna orthography; the orthographic transcription in the Tigrigna writing system is Ge'ez syllable, and also uses its own number and punctuations that are different from their counterpart in the English language. And this pose problems in applying different tools, like text tokenization, sentence splitting, segmentation tools. In addition, the syntax structure of both languages English and Tigrigna is different, and this pose problem that needs reordering.

CHAPTER FOUR

Design of Bidirectional Tigrigna – English Machine Translation (BTEMT)

This Chapter discusses the bidirectional Tigrigna – English machine translation system. The overall system architecture and its components are all discussed in detail. As it is already discussed in Chapter three, Tigrigna language is one of the Semetic languages with rich morphology. It is also under resourced language. These and other property of the language makes statistical machine translation for the language challenging. To overcome these challenges, SMT with linguistic information is important because it solves problems like data sparseness and low training data.

4.1 Architecture of BTEMT

Statistical machine translation (SMT) paradigm is implemented for this study. SMT is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. Statistical machine translation has three components: translation model, language model and decoder. The language model component tries to ensure fluent translation by guiding word choice and word order, and the role of this component is to give an estimate of how probable a sequence of words is to appear in the target language. The second component is translation model, which assigns a probability that a given source language sentence generates target language sentence. The third component is a decoder, it searches for the best sequence of alterations that translates source sentence to the corresponding target sentence. It is the main part of the translation system which takes source sentence E and produces the best target sentence translation T according to the product of the language and translation models. The system uses phrase based translation, the decoder uses the phrase translation table to search for all translations of every source words or phrases, and then it recombines the target language phrases that maximizes the product of language model and translation model probability. The tools and other necessities that are essential to accomplish using this approach were implemented and will be explained as follows. Figure 4.1 shows the Architecture of the system and the succeeding subsections explain the components of the architecture in detail.

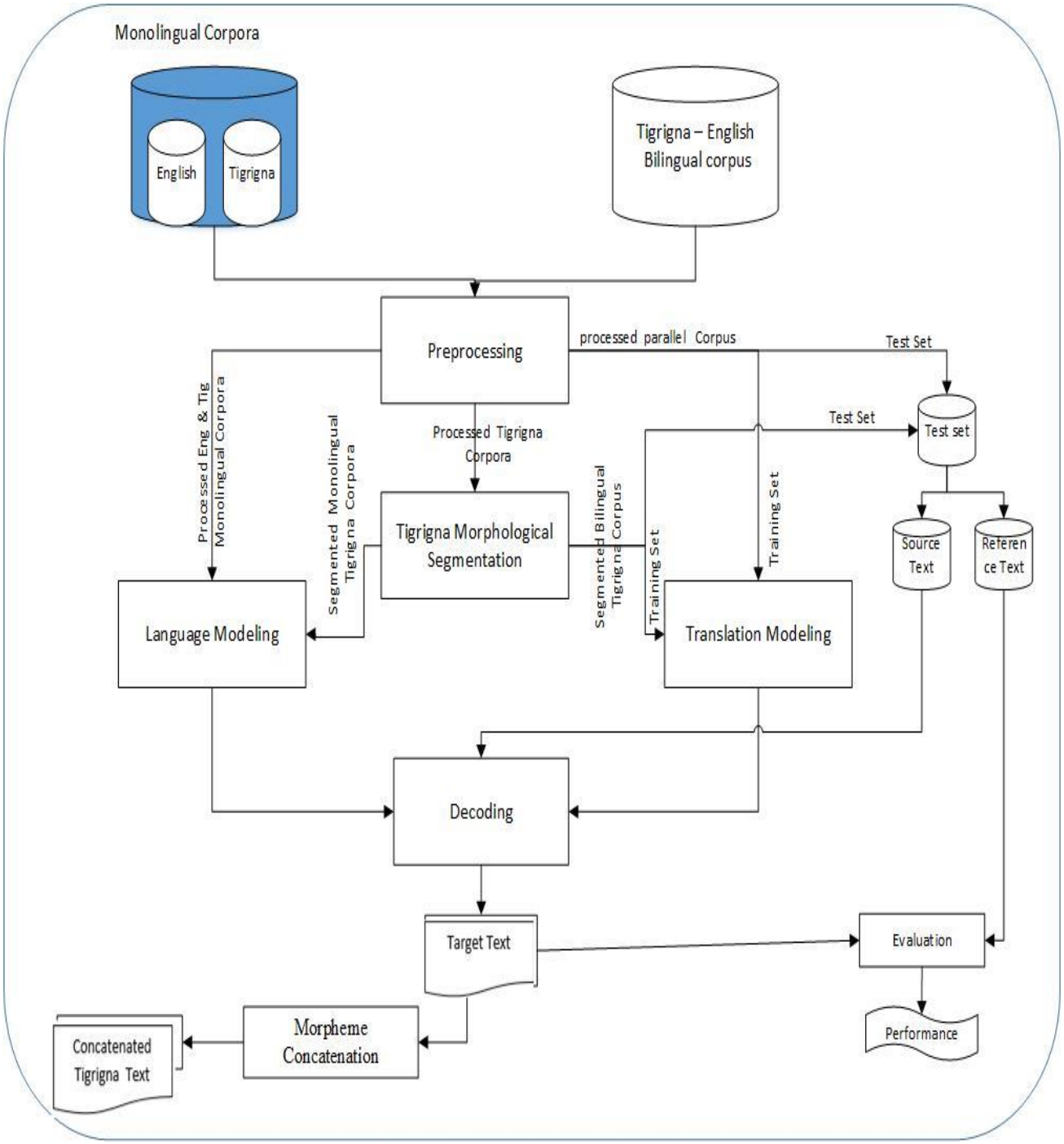


Figure 4.1: Architecture of the proposed system

4.2 The Corpora

4.2.1 Preliminary Preparation

The data used for this research were collected from different sources, the collected raw files were in different formats and encodings. The researchers found digitally available Tigrigna and English versions of some chapters of the New Testament of Holly Bible; which is aligned at verse level. While religious material may not contain the day to day language words, it has the advantage of being inherently aligned on the verse level facilitating further alignment. This was found in html file format which is converted manually to Unicode text file. Another bilingual text is simple sentences customized from [13], which was available in English and Amharic language, so manual translation is conducted by the researchers on the Amharic part to corresponding Tigrigna sentences. Fortunately, this text was manually sentence aligned and it was in doc file which is pretty easy for format conversion; converted to Unicode text file. The other bilingual text is from constitution of FDRE in PDF format; which demanded a lot of time. The first pre-process is to convert the corpus from PDF to RTF then to Unicode text. The most challenging task was converting from the RTF to Unicode text file. The good quality converter tool found was the Power Geez 2010 version. The Tigrigna fonts recognized by the converter include: VG Geez Numbers, VG2000 main, Ge'ez-1, VG2 main, VG2 Title, Visual Geez Unicode, Visual Geez Unicode Title, and Power Geez Unicode1. The converter does not convert all characters to their respective fonts because of some words contain more than two fonts. After automatic conversion of the full document is complete some words with other fonts will contain odd characters. As a result, it takes time to manually select and reconvert those words that have been wrongly converted. Since the sentences prepared from [13] was translated by the researcher, so it needed to be checked for correctness by a linguist. The other corpora are done by professionals since it is made for the use of the public.

4.2.2 Corpora Preparation

Five different corpora have been developed for the purpose of this study. The first corpus was taken from the New Testament of English and Tigrigna Bible; it contains around 6106 relatively long sentences for each language which is aligned at verse level manually. Since religious materials may contain words that are different from the day-to-day language words, preparing additional corpus is required. The second corpus is prepared by customizing simple sentences from [13] which are manually aligned at sentence level; it contains around 1160 sentences. but the

sentences obtained was very small which is not enough for SMT. This motivates to prepare the next corpus. The third corpus is prepared from two different source; by combining the customized simple sentences from [13] and sentences from constitution of the FDRE. The combination made for this two different sources are because of their size and domain. The size obtained was 1982 which is small relative to other systems, and to solve this the fourth corpus is prepared. The fourth corpus is prepared by combining Corpus I; sentences from the Holly Bible, and Corpus III; sentences from constitution of FDRE and customized simple sentences. The purpose of this combination is to solve the issue of domain specific and to increase the size of the corpora. And their size is 6106 sentence pairs from Corpus I , 1982 sentence pairs from Corpus III. But the size form Corpus I shows dominancy since the source for this corpus is form Holy Bible which pose problem for the training and translation. From this, the fifth corpus is prepared from Corpus I and Corpus III by considering their size. The size from each corpus is : 1982 sentence pairs from Corpus III, and randomly selected 1018 sentences pairs from Corpus I. Since Corpus III is prepared from two domains; the amount of sentences selected from this corpus is almost twice of Corpus I. Sample text from the corpora prepared for this study is illustrated on Appendix V and VI.

In addition to bilingual corpora, SMT requires monolingual corpora. There are many options to access the monolingual corpora. One of the options is to take the monolingual corpora of the target language from the parallel corpus that are not used in the translation. Since our study is bidirectional, two monolingual corpora are prepared; one for Tigrigna and the other for English. In addition to the sentences extracted from the bilingual corpora, additional sentences are added from different source; with the objective of increasing the size of the monolingual corpus that is used to estimate the fluency of a sentence in the language. Additional sentences for English monolingual corpora are from criminal code of Ethiopia, and English version of Tigray regional state constitution. For Tigrigna monolingual corpora, additional sentence are added from the web. The detailed size of the corpora are described in Section 5.1.

4.3 Preprocessing the Corpora

There are a number of preprocessing involved to get a cleaned corpora. These preprocessing includes sentence alignment, splitting punctuation marks, and true-casing.

Sentence Level Alignment : The alignment at the sentence level has been done using a sentence aligner called Hunalign developed by [52]. Alignment tasks are performed on each corpus except corpus prepared from Bible; because, sentences from the Holy Bible are aligned at verse level manually. Since the corpora are aligned before this step using manual alignment during the preliminary preparation of the corpora, the aligner was able to align 1160 sentence pairs for simple sentences and 822 sentence pairs for the constitution of FDRE independently.

Splitting Punctuation Marks : This is the process of separating words from punctuation and symbols into tokens. These tokens are words, punctuations, and numbers. Tokenizing tasks are performed on each corpus in order to convert them into a valid format suitable for the translation system. This tokenization is performed before aligning sentence pairs. Tokenization is performed using a Perl script available in [53] and developed for English language. However, this does not work for Tigrigna; because, the punctuation used in Tigrigna and English is different. So, customization of the Perl scrip is needed in the case of Tigrigna sentences tokenization. This is done by converting English punctuation to each corresponding Tigrigna punctuations in the script. The converted English punctuations are (‘,’, ’:’, ’;’, ’.’) to their Tigrigna counterparts (‘፣’, ‘፥’,’ ፤ ’, ‘።’)

True-casing : True-casing is the process of converting the initial words in each sentence to their most probable casing. True-casing tasks are performed on each corpus in order to convert them into their most probable casing. This helps to reduce data being sparse. This is only for English language; because, there is no casing in Tigrigna language. Similar to the previous one, this was also conducted using a Perl script [53]

4.4 Morphological Segmentation

This morphological segmentation focuses on segmenting morphemes of Tigrigna words. Morphemes are divided into two classes called roots and affixes [45]. The core meaning bearing units are roots, and affixes are added or inserted into roots. Morphemes are considered to be the smallest meaning-bearing elements of language. Any word form can be expressed as a

combination of morphemes, for example: “ብትሕተይውን” as “ብ + ትሕተይ + ውን”, “ንኣደይ” as “ን + ኣደይ”, “እናኣረድኦምን “as “እና + ኣረድኦም + ን “, “ከይረአዩ “as “ከይ + ረአዩ “, “ንሕጻናት “as “ን + ሕጻናት “, “ምድሪ ውን” as “ምድሪ + ውን”, “እናተጋደልኩ” as “እና + ተጋደልኩ”, “ንሙሴ “ as “ን + ሙሴ “, “ንኹላቶምውን “ as “ን + ኹላቶም + ውን “, “ዘረባውን “ as “ዘረባ + ውን “, “ንማርያምን “ as “ን + ማርያም +ን “.

Many existing applications make use of words as vocabulary base units. However, for highly-inflected languages like Tigrigna, this is not feasible, as the number of possible word forms is very high. Literatures shows that, word segmentation plays an important role for morphologically rich languages in many NLP applications. Tigrigna is a morphologically rich language; so, in this research we investigate Tigrigna as the source and target language for segmentation. Although Tigrigna is morphologically rich language, there is no segmentation system developed for the language yet. Because of this scarcity, we look at unsupervised word segmentation systems to see how well they perform word segmentation without relying on any linguistic information about the language. As indicated in [54], Morfessor 1.0 works best with SMT from other unsupervised segmentation tools. Morfessor is a general tool for the unsupervised induction of a simple morphology from raw text data, and it has been designed to cope with languages having predominantly a concatenative morphology and where the number of morphemes per word can vary much and is not known in advance [55].

The Morfessor program requires a data file. The data file must be a word list with one word per line. The word may be preceded by a word count (frequency) separated from the word by whitespace; otherwise a count of one is assumed. If the same word occurs many times, the counts are accumulated. The program does not perform any automatic character conversion except it removes superfluous whitespace and carriage return characters. For instance, equivalent characters like “ ‘ጸ’, ‘ፀ’ (ጸልማት, ፀልማት) “ entails to be treated as the same character, this needs to perform the desired conversion and filtering prior to using the Morfessor program. So in this research we use the segmentation learner, to segment the Tigrigna sentences, it includes the training set, the test set and the language model of the Tigrigna language. This process is described in figure 4.2:

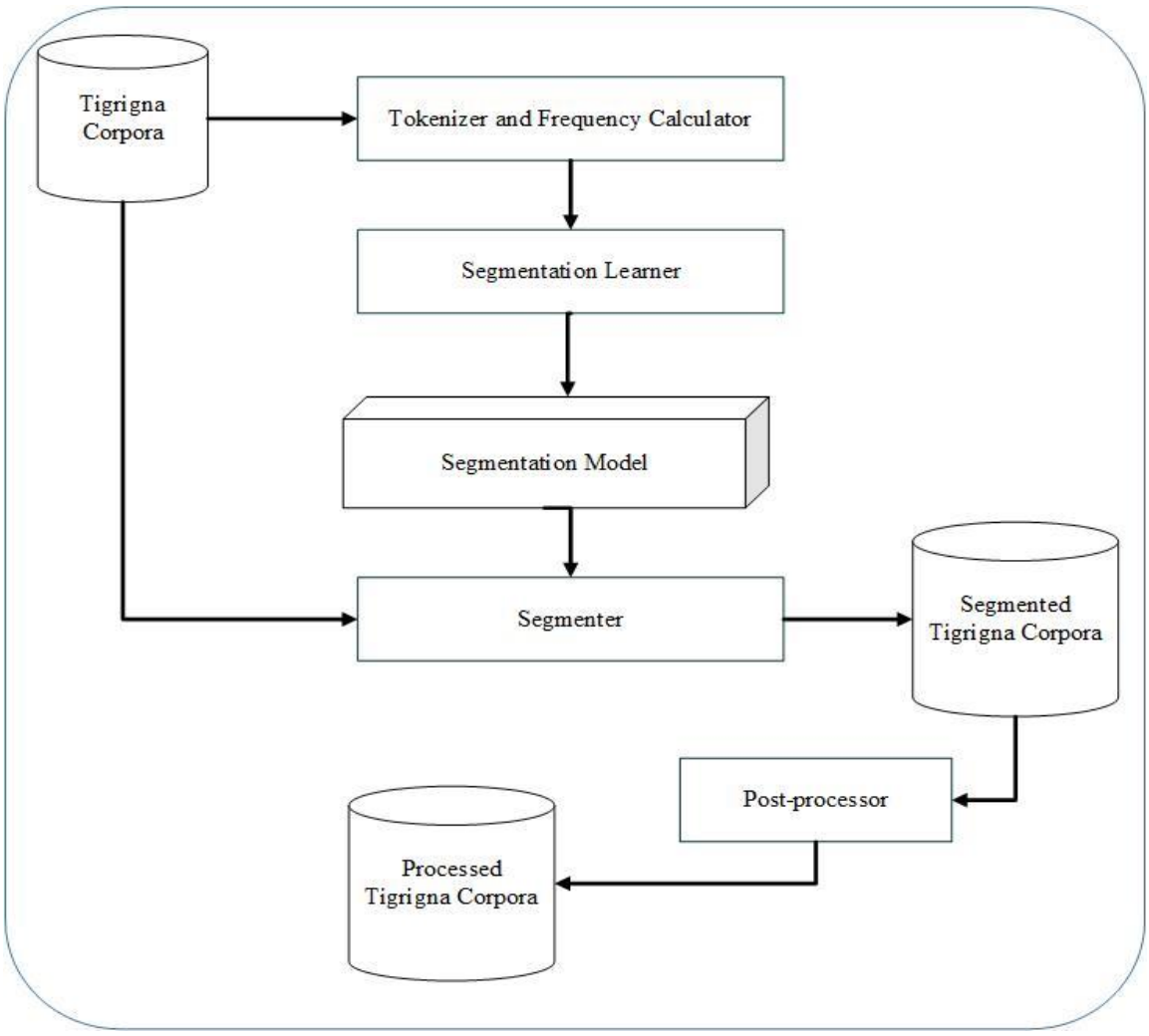


Figure 4.2 : Architecture of the Morphological Segmenter

4.4.1 Tokenizer and Frequency Calculator

Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The Tigrigna corpora becomes input to this tokenizer and frequency calculator component, and the component generates list of words with their frequency of occurrence. Punctuation, digits and whitespaces are not included in the resulting list of tokens. The list of tokens becomes input to the segmentation learner for further processing. The algorithm used for the Tokenizer and Frequency Calculator is shown below:

1. Load the sentences from the Tigrigna monolingual corpora
2. Load all punctuation list // list of punctuation
3. Store all sentences in S
4. Store all words in W
5. Initialize dictionary {}
6. for each s_i Ɛ S do, where i = 1,2,3,4 …. k
7. for each w_j in s_i do, where j = 1,2,3,4, … k
8. if w_x is digit
9. Remove it
10. else if w_x is found in the punctuation list
11. Remove it
12. else if w_j found in dictionary
13. increment the index by 1
14. else
15. store the word in dictionary
16. End for
17. Write the dictionary to a file

Algorithm 4.1: Algorithm for word tokenizing and frequency calculator

4.4.2 Segmentation Learner

The segmentation learner that we adopt for this research is unsupervised segmentation tool which is called Morfessor 1.0. This unsupervised word segmentation system learns the segmentation from a list of words with their frequency that are not annotated or pre-processed in any way that helps the system to predict the correct segmentation. The output from this segmentation learner is morph segmentations of the words in the input. This morph segmentation output becomes a segmentation model that can be used as a model to segment Tigrigna sentences. The segmentation was performed on the Ge'ez syllables without transliterating to Latin characters. This was done using Morfessor1.0-utf8.perl; because, this version supports Unicode characters. Sample output from the system are shown in Figure 4.3.

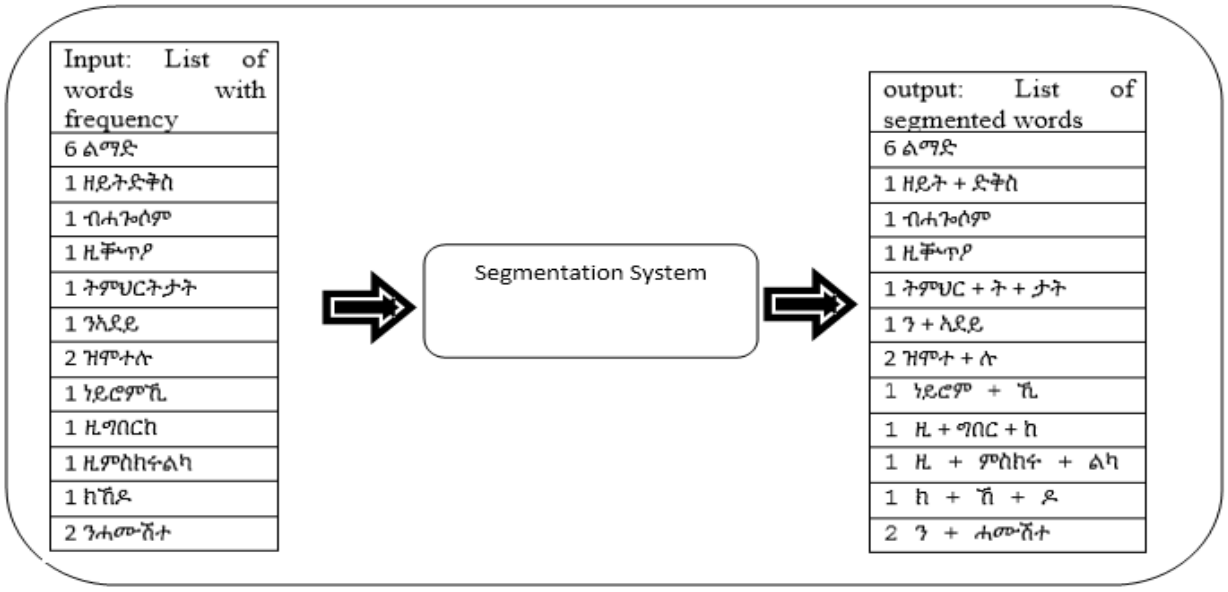


Figure 4.3: Sample output from the Segmentation learner

The table in the left side of the figure is a list of input words to the segmentation system, and the system produces the segmentation model i.e. similar to the table in the right side of the figure, and the details about the segmentation model is described in the next sub topic.

Once a morph segmentation model has been built from the text data set (monolingual corpora), it can be used for segmenting new word forms. In this segmentation model of the Morfessor program no model learning takes place. Each input word is segmented into morphs by the Viterbi algorithm, which finds the most likely segmentation of the word into a sequence of morphs that are present in the existing model. The segmentation model is the model which is produced by the segmentation learner component, by taking a list of words with their frequency similar to the table in the right side of figure 4.3 i.e. the result becomes a model which is used as input to the segmenter.

4.4.3 Segmenter

We use the resulting segmentation model as input to the unsupervised segmenter to segment the available Tigrigna parallel corpus (both the training and test set) and the language model text. For example, in the following figures (figure 4.4 and figure 4.5) shows list of sentences that are input and output of the segmentation learner.

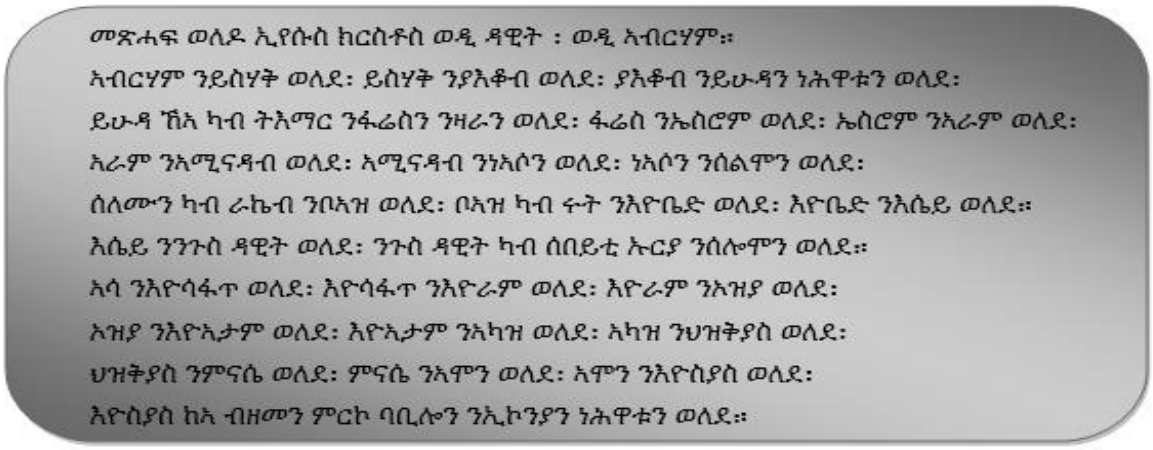


Figure 4.4. Input to the unsupervised segmenter

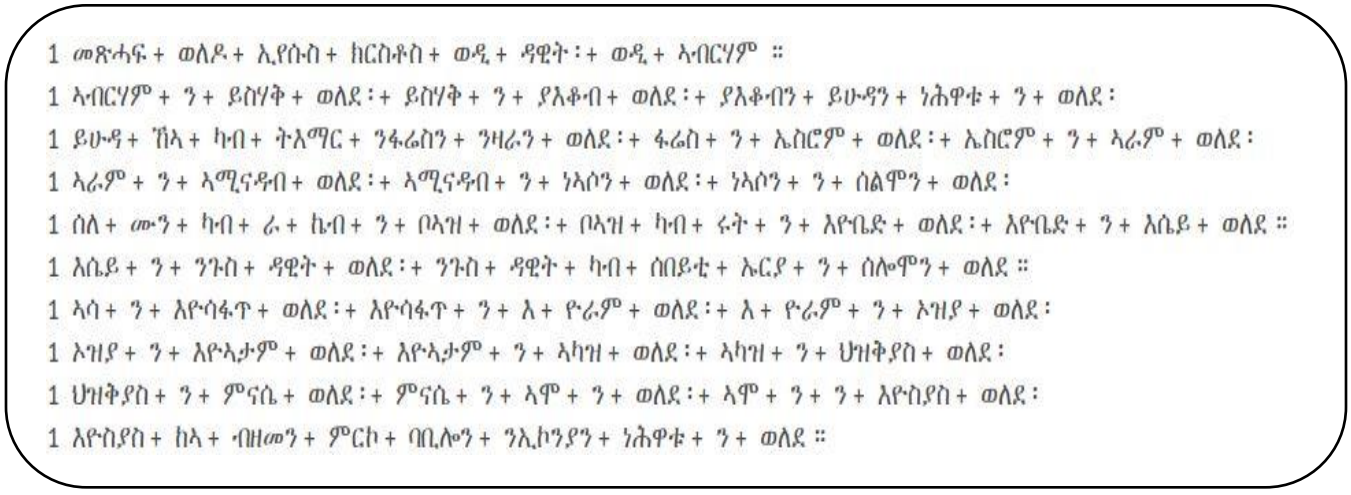


Figure 4.5. Output from the post-segmentation learner

Figure 4.5 shows segmented sentences which are output of the system, but the sentences are preceded by a sentence count (frequency) separated by whitespace, and the segmented morphemes are separated by ' + ' . so postprocessing is required.

4.4.4 Post-processor

Once the Tigrigna sentences are segmented using the post-segmentation learner, it generates segmented sentences but the sentences contain unnecessary symbol ' + ' and '1' as it is shown in Fig 4.5. So, the post-processor is used to remove this symbols. We designed and wrote python program to remove this symbol from the Tigrigna sentences. Finally the post processor facilitated to get sentences that are ready for the language model and translation model like sentences in Figure 4.6.

```
1.    Load all Segmented sentences
2.    Store all sentences in S
3.        for each si Ɛ S do, where i =1,2,3,4 …
5.            for each wj in si do, where j = 1,2,3,4 …
6.                delete wj[0]
7.                if wj == " + "
8.                    delete  wj
9.                End if
10.           End for
11.           Store si in S'
12.       End for
13.   Write S' to file
```

Algorithm 4.2 : Algorithm for Post processing the segmented Tigrigna sentences

Output from the post-processor that are ready for training

```
መጽሓፍ ወለዶ ኢየሱስ ክርስቶስ ወዲ ዳዊት ፡ ወዲ ኣብርሃም ።
ኣብርሃም ን ይስሃቅ ወለደ ፡ ይስሃቅ ን ያእቆብ ወለደ ፡ ያእቆብን ይሁዳን ነሕዋቱ ን ወለደ ፡
ይሁዳ ኸኣ ካብ ትእማር ንፋሬስን ንዛራን ወለደ ፡ ፋሬስ ን ኤስሮም ወለደ ፡ ኤስሮም ን ኣራም ወለደ ፡
ኣራም ን ኣሚናዳብ ወለደ ፡ ኣሚናዳብ ን ነኣሶን ወለደ ፡ ነኣሶን ን ሰልሞን ወለደ ፡
ሰለ ሙን ካብ ራ ኬብ ን ቦኣዝ ወለደ ፡ ቦኣዝ ካብ ሩት ን እዮቤድ ወለደ ፡ እዮቤድ ን እሴይ ወለደ ።
እሴይ ን ንጉስ ዳዊት ወለደ ፡ ንጉስ ዳዊት ካብ ሰበይቲ ኤርያ ን ሰሎሞን ወለደ ።
ኣሳ ን እዮሳፋጥ ወለደ ፡ እዮሳፋጥ ን እ ዮራም ወለደ ፡ እ ዮራም ን ኦዝያ ወለደ ፡
ኦዝያ ን እዮኣታም ወለደ ፡ እዮኣታም ን ኣካዝ ወለደ ፡ ኣካዝ ን ህዝቅያስ ወለደ ፡
ህዝቅያስ ን ምናሴ ወለደ ፡ ምናሴ ን ኣሞ ን ወለደ ፡ ኣሞ ን ን እዮስያስ ወለደ ፡
እዮስያስ ከኣ ብዘመን ምርኮ ባቢሎን ንኢኮንያን ነሕዋቱ ን ወለደ ።
```

Figure 4.6: output from the post processor

4.5 Language Model

The function of the language model in the SMT Architecture is to give an estimate of how probable a sequence of words is to appear in the target language. The language model helps the translation system with selecting words or phrases appropriate for the local context and with combining them in a sequence with better word order. N-gram model was used for this study. In order to have a

reliable estimate for the language model probabilities, the context of the language model is usually restricted to a few words. Another restriction on the context of language models is that the probability of words is computed within the boundaries of a sentence. The probabilities obtained from the N-gram model could be unigram, bigram, trigram or higher order N-grams. For example given the following Tigrigna sentences:

- ✓ ሓነስ ቡን ሰትዩ ።
- ✓ ሓነስ ቡን ገዚኡ።
- ✓ ሓነስ ሻሂ ሰትዩ ።
- ✓ ለምለም ምሳሓ ተመሲሓ ።
- ✓ ሰላም ሕምባሻ ገዚአ ።

The unigram probability can be calculated by:

$$P(w1) = \frac{\text{count(w1)}}{\text{total words observed}}$$

$$P(\text{ሓነስ}) = \frac{3}{15} = 0.2$$

Where 3 is the number of times the word "ሓነስ" was used and 15 is the total words in the above example.

The bigram probability can be calculated by:

$$P(w2/w1) = \frac{\text{count(w1w2)}}{\text{count(w1)}}$$

$$P(\text{ቡን}/\text{ሓነስ}) = \frac{\text{count(ሓነስ ቡን)}}{\text{count(ሓነስ)}} = \frac{2}{3} = 0.667$$

Where 2 is the number of times the words "ሓነስ" and "ቡን" have been used together and 3 is the number of times the word "ሓነስ" is used.

And the trigram probability can be:

$$P(w3/w1w2) = \frac{\text{count(w1w2w3)}}{\text{count(w1w2)}}$$

$$P(\text{ሰትዩ}/\text{ሓነስ ቡን}) = \frac{\text{count(ሓነስ ቡን ሰትዩ)}}{\text{count(ሓነስ ቡን)}} = \frac{1}{2} = 0.5$$

Where 1 is the number of times the word "ሓነስ", "ቡን" and "ሰትዩ" have been used together and 2 is the number of times the words "ሓነስ", and "ቡን" were used.

A LM assigns a higher probability to fluent or grammatical sentences estimated using monolingual corpora. The LM component tries to ensure fluent translation by guiding word choice and word order. There are different Language Modeling Toolkits that are open-source. SRILM, IRSTLM, CMU SLM or KenLM are developed and integrated with the SMT tool, Moses. From the above, IRSTLM has been used as a language modeling toolkit for the purpose of this study. This is because it requires less memory than others [17]. The IRSTLM supports three output format of LM [56]. These formats have the purpose of permitting the use of Language Models by external programs, the formats are ARPA format, qARPA format, and iARPA format. External programs could in principle estimate the LM from an n-gram table before using it, but this would take much more time and memory. So, the best thing to do is to first estimate the LM, and then compile it into a binary format (. blm) that is more compact and that can be quickly loaded and queried by the external program.

As described in the beginning, the system is bidirectional and a language model has been developed for both languages i.e. English and Tigrigna by using trigram language model with the help of IRSTLM tool; trigram order is the state-of-the-art for phrase based translation. Three language models were developed using the whole target corpora, one for English language, and the other two are for the Tigrigna language i.e. for the baseline and segmented corpora. For Tigrigna – English translation, a language model has been developed for English language, and for English – Tigrigna translation, a Tigrigna language model has been developed.

Smoothing : To avoid assigning zero or low probabilities to unseen events, the maximum likelihood estimates for the conditional probabilities are adjusted to make the distribution more uniform. By doing so, zero and low probabilities are increased and high probabilities are decreased. By applying this type of technique called smoothing, the estimations are improved and unseen N-grams receive some probability. Kneser-Ney Smoothing – state of the art smoothing technique is used [57]. It is Based on absolute discounting, but uses more sophisticated backoff mechanism. And the basic idea is unigram probability should not be proportional to the frequency of a word, but to the number of different words that it follows (the number of histories it has). The assumption is words that appeared in more contexts are more likely to appear in some other new context as well.

4.6 Translation Model

The translation model looks for the most probable target language sentence given source language sentence P(E|T). The translation model assigns higher probability to sentences that have corresponding meaning estimated using bilingual corpora. Based on the above architecture two bilingual corpora are prepared for the TM. The first bilingual corpus is for the baseline translation which is directly inserted into the TM from the preprocessing component. And the second is from the morphological segmentation which is segmented bilingual corpus i.e. the segmentation is for Tigrigna only. The training corpus for this model is a sentence level aligned parallel corpus of Tigrigna and English languages. We cannot compute P(E|T) from counts of source and targets sentences in the parallel corpus, because sentences are novel, so we could never have enough data to find values for all sentences. The solution is to find (approximate) the sentence translation probability using the translation probabilities of chunks in the sentences, as in language modeling [35]:

$$P(E/T) = \sum_X P(X, E/T) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.1)$$

The variable X represents alignments between the individual chunks in the sentence pair, and chunks in the sentence pair can be words or phrases. The word/chunk alignment is done using the Expectation Maximization (EM) algorithm, because the parallel corpus gives us only the sentence alignments; it does not tell us how the chunks in the sentences are aligned. This means that a group of words in the source language should be translated by a group of words in the target language and there might not be a word-level correspondence between these groups.

4.6.1 Alignment

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

Expectation Maximization (EM) : An expectation maximization algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in

statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected loglikelihood found on the E step. EM algorithm works as follows [57]:

- Start with some uniform word translation probabilities and calculate alignment probabilities, and
- These alignment probabilities to get (hopefully) better translation probabilities, and keep on doing this, we should converge on some good values.

let's take the following phrase as an example:

- ✓ red car the car
- ✓ ቀይሕ መኪና እታ መኪና

The vocabularies for the two languages are E= {red, car, the} and T= {ቀይሕ, መኪና, እታ}. We will start with uniform probabilities:

t (ቀይሕ|red) = 1/3	t (መኪና|red) = 1/3	t (እታ|red) = 1/3
t (ቀይሕ|car) = 1/3	t (መኪና|car) = 1/3	t (እታ|car) = 1/3
t (ቀይሕ|the) = 1/3	t (መኪና|the) = 1/3	t (እታ|the) = 1/3

E-Step 1: Compute the expected counts E [counts (t (f, e))] for all word pairs (f_j, e_{aj})

E-Step 1a: we first need to compute P (a, f |e), by multiplying all the t probabilities,

P (a, f |e) = t (ቀይሕ, red) * t (መኪና, car) = 1/3 * 1/3 = 1/9

P (a, f |e) = t (ቀይሕ, car) * t (መኪና, red) = 1/3 * 1/3 = 1/9

P (a, f |e) = t (እታ, the) * t (መኪና, car)

= 1/3 * 1/3 = 1/9

P (a, f |e) = t (እታ, car) * t (መኪና, the)

= 1/3 * 1/3 = 1/9

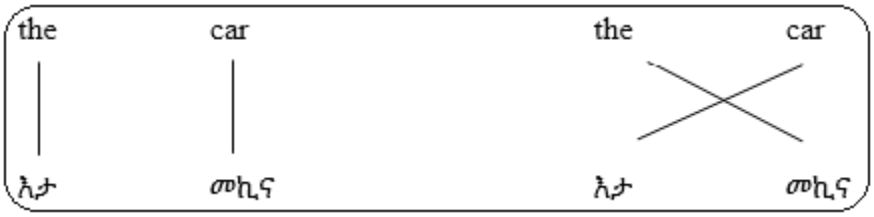
E-Step 1b: normalize P (a, f | e) to get P (a | e, f), using the following

$$P\ (a \mid e, f) = \frac{P\ (a,\ f \mid e)}{\sum_a P\ (a,\ f \mid e)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.2)$$

The resulting values of P (a | f, e) for each alignment are as follows:

P (a | f, e) $= \frac{1/9}{2/9}$ = 1/2

P (a | f, e) $= \frac{1/9}{2/9}$ = 1/2



P (a | f, e) $= \frac{1/9}{2/9}$ = 1/2

P (a | f, e) $= \frac{1/9}{2/9}$ = 1/2

E-Step 1c: compute expected (fractional) counts, by weighting each count by P (a |e, f)

tcount (ቀይሕ | red) = 1/2 tcount (መኪና | car) = ½ + ½ tcount (እታ | red) = 0

tcount (ቀይሕ | car) = 1/2 tcount (መኪና | red) = 1/2 tcount (እታ | car) = 1/2

tcount (ቀይሕ | the) = 0 tcount (መኪና | the) = 1/2 tcount (እታ | the) = ½

total(red)= 1 total(car)= 2 total(the)= 1

M-Step 1: compute the MLE probability parameters by normalizing the tcounts to sum to one

t (ቀይሕ | red) = $\frac{1/2}{1}$ = ½ t (መኪና | car) = $\frac{1}{2}$ = ½ t (እታ | red) = $\frac{0}{1}$ = 0

t (ቀይሕ | car) = $\frac{1/2}{2}$ = ¼ t (መኪና | red) = $\frac{1/2}{1}$ = ½ t (እታ | car) = $\frac{1/2}{2}$ = ¼

t (ቀይሕ | the) = $\frac{0}{1}$ = 0 t (መኪና | the) = $\frac{1/2}{1}$ = ½ t (እታ | the) = $\frac{1/2}{1}$ = ½

Note that each of the correct translations have increased in probability from the assignment; for example, the translation "ቀይሕ" for "red" has increased in probability from 1/3 to 1/2.

E-Step 2a: we re-compute P (a, f | e), again by multiplying all the t probabilities, t (f_j | e_{aj})

$$P(a, f \mid e) = \prod_{j=1}^{j} t(f_i \mid e_{aj}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4.3)$$

P (a, f | e) = t (ቀይሕ, red) * t (መኪና, car) P (a, f | e) = t (ቀይሕ, car) * t (መኪና, red)

= 1/2 * 1/2 = 1/4 = 1/4 * 1/2 = 1/8

P (a, f | e) = t (እታ, the) * t (መኪና, car) P (a, f | e) = t (እታ, car) * t (መኪና, the)

= 1/2 * 1/2 = 1/4 = 1/4 * 1/2 = 1/8

Note that the two correct alignment are now higher in probability than the two incorrect alignments.

4.7 Decoder

The decoder is the major component of the Moses besides LM and TM, and is used as a means for the experiment with Tigrigna - English parallel corpus. As it has been discussed in Section 2.5.3, decoding is a search problem which finds the sentence which maximizes the translation and language model probabilities. Since the system is phrased based translation, the decoder uses the phrase translation table to search for all translations of every source words or phrases, and then it recombines the target language phrases that maximizes the language model probability multiplied by the translation model probability. For example if we want to translate a sentence t in the source language T (Tigrigna) to a sentence e in the target language E (English), the best translation is the probability that maximizes the product of the probability of Tigrigna – English translation model p(t|e) and the language model of English p(e), this is expressed using formula:

$$\overline{e} = \underset{e}{\operatorname{argmax}}\, p(t|e) * p(e)$$

Since the system is bidirectional, there is also translation from English – Tigrigna, the best translation is the probability that maximizes the product of the probability of English – Tigrigna translation model p(e|t) and the language model of Tigrigna p(t), mathematically :

$$\overline{t} = \underset{t}{\operatorname{argmax}}\, p(e|t) * p(t)$$

4.8 Evaluation

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

of the system output is correct, rather than measuring whether the references are fully reproduced in the system output.

4.9 Morphological Concatenation

Since the system is bidirectional the target language is both Tigrigna and English. But in case of English to Tigrigna translation the output is segmented Tigrigna sentences, which is not human readable, so concatenation of morphemes is required in order for the output to be readable by human beings. The concatenation is based on the specific target characters that has been conducted on the segmentation. The algorithm used for the concatenation is shown in Algorithm 4.3 below.

```
1.  Load the translated sentences
2.  Store all sentences in S
3.  Load all List of Tigrigna words
3.  Store all words in X
4.  for each si Ɛ S do, where i =1,2,3,4 …
5.      for each wj in si do, where j = 1,2,3,4 …
6.          if wj  is not found in X
7.              delete  wj+1
8.          End if
9.      End for
10.     Store si in S'
11. End for
12. Write S' to file
```

Algorithm 4.3 : Algorithm for concatenation the Tigrigna target sentences

Based on this design, bidirectional Tigrigna – English Statistical Machine Translation experiments are conducted on the next chapter. And also it includes the evaluation of each experiment.

CHAPTER FIVE

Experiment

Based on the design presented in Chapter Four, bidirectional Tigrigna – English machine translation systems were developed by using SMT approach. This Chapter presents experimental results of our bidirectional Tigrigna – English statistical machine translation.

5.1 Experimental Setup

In this study three groups of experiment conducted to come up with bidirectional SMT for Tigrigna – English language pair. The first group of experiment focuses on baseline statistical machine translation, the second focuses on segmented system and the third group of experiment is with segmented post processed system. Each group of experiment was conducted with Five different corpora; the detail is presented in Section 5.2 . The resources needed for the Tigrigna – English experiment are a monolingual corpus for building the LM and the bilingual Tigrigna – English parallel corpus to build a translation model. Since the translation is bidirectional; a language model is developed for both languages. English monolingual corpora are built from the bilingual corpora and additional sentences are from English version of criminal code of Ethiopia and constitution of Tigray regional state are used. And the Tigrigna monolingual corpora are prepared from the bilingual corpora similar to English, and some texts from the web. The following table summarizes the amount of bilingual and monolingual data used in this research.

<table>
<tr><td rowspan="2">Corpus Name</td><td rowspan="2">Units</td><td colspan="2">Bilingual</td><td colspan="2">Monolingual</td></tr>
<tr><td>Tigrigna</td><td>English</td><td>Tigrigna</td><td>English</td></tr>
<tr><td>Corpus I</td><td rowspan="5">Sentences</td><td>6106</td><td>6106</td><td rowspan="3">8506 Sentences</td><td rowspan="3">16016 Sentences</td></tr>
<tr><td>Corpus II</td><td>1160</td><td>1160</td></tr>
<tr><td>Corpus III</td><td>1982</td><td>1982</td></tr>
<tr><td>Corpus IV</td><td>8088</td><td>8088</td><td rowspan="2">107657 tokens</td><td rowspan="2">268978 tokens</td></tr>
<tr><td>Corpus V</td><td>3000</td><td>3000</td></tr>
</table>

Table 5.1: summary of the size of the total data used in the research

The Tigrigna – English SMT systems include, baseline, morph-based, and post-processed segmented systems, which are all developed using the freely available open-source toolkits. These

tools are: Moses used to train the system in both directions, IRSTLM for building the language model, GIZA++ for word alignment, Morfessor 1.0 for morpheme segmentation, and BLEU score for evaluation.

5.1 Parallel corpora

Corpus I : As it has been described in the above table, Corpus I was made of about 6106 relatively long sentences; because they are aligned at verse level, that had been prepared from the New Testament of Tigrigna and English Bible. Corpus I was divided into two training set and test set. Out of 6106 sentences 90% were used for the training set, and the rest 10 % of the corpus is used for testing. This random classification is done using Perl script. The translation system was trained both ways, that is from Tigrigna to English and vice versa. The same steps were followed for the translation process. The same training and test set are used to test both systems.

Corpus II : Corpus II includes 1160 simple sentences, prepared from [13]. The experimentation process for Corpus II is similar to Corpus I; the corpus is classified in to training set and test set. Out of 1160 sentences 90 % of the corpus were used for the training set, and the rest 10 % were used for testing. Similarly, the translation system was trained in both directions, that is from Tigrigna to English and English to Tigrigna.

Corpus III : Corpus III includes 1982 sentences, prepared from the Constitution of FDRE and simple sentences of corpus II. The experimentation process for Corpus II is also similar to Corpus I and II; the corpus is classified in to training set and test set. Out of 1982 sentences 90 % of the corpus were used for the training set, and the rest 10 % were used for testing. Similarly, the translation system was trained in both directions, that is from Tigrigna to English and English to Tigrigna.

Corpus IV : Corpus IV was made of 8088 sentences, prepared from Corpus I and Corpus III, with size of 6016 and 1982 sentences respectively. The experimentation process is similar to the other corpora. The corpus is divided into training and testing set. Out of 8088 sentences 90% of the corpus is used for training set and 10 % of the corpus were used for testing. Like the other corpora the translation system was trained both directions. The same steps were followed for the translation process, and the same training and test set have been used in both translation systems.

Corpus V : Corpus V includes 3000 sentences, prepared from Corpus I and Corpus II, with size of 1018 and 1982 sentences respectively. The experimentation process is similar to the other corpora. The corpus is divided into training and testing set. Out of 3000 sentences 90% of the corpus is used for training set and 10 % of the corpus were used for testing. Like the other corpora the translation system was trained both directions. The same steps were followed for the translation process, and the same training and test set have been used in both translation systems.

5.2 Training the System

The training process includes creating language model, translation model, and conducting decoding using the help of GIZA++ and IRSTLM. The created language model is built with the target language model, that is, for Tigrigna as well as English separately since it is bidirectional translation system, both the languages become a target language at some point. As it has been stated, IRSTLM toolkit was used. A trigram language model, smoothing with Kneser-Ney and adding sentence boundary symbols were built.

The word alignment, phrase extraction and scoring were used and lexicalized reordering tables and Moses configuration files were created with the training translation system. Mainly this step creates a "moses.ini" file, which is used for decoding and the phrase table is also created which basically contains the probabilities of a word following another. But there is a couple of problems with the file "moses.ini". It is slow to load and also not optimized. The first problem fixed by binarizing (a format that can be load quickly); the phrase table and reordering table. The second problem is solved by tuning, the translation system is tuned.

As mentioned above, while training the translation system, a "moses.ini" file were created which is used for decoding. The querying process could be started right away but the weights used by Moses to weight the different models against each other are not optimized. Therefore, to find better weights, the translation systems needs to be tuned. This process produces another "moses.ini" file that is used for decoding.

5.3 Experiment I: Baseline SMT

This is the first group of experiment conducted on Tigrigna – English language pair for each corpus by using a statistical approach. As it was mentioned in Chapter Two, statistical machine translation is an approach which tries to generate translations using statistical methods based on bilingual text

corpora. The baseline systems have been trained using the five different Tigrigna – English parallel training sets and tested using Tigrigna and English source sentences as new sentences that gives an output target Tigrigna and English sentences of . The output then has been checked with the reference sentences, and scored using the BLEU score. The result obtained from the BLUE score is shown in Table 5.2 and snapshot of the result from the system is described in Appendix II.

	Corpus Name	Translation Accuracy in BLUE score	
		English - Tigrigna	Tigrigna - English
Baseline Phrase-based Translation	Corpus I	10.59	18.09
	Corpus II	20.55	42.75
	Corpus III	18.65	36.40
	Corpus IV	12.41	24.69
	Corpus V	13.94	23.81

Table 5.2: Summary of BLUE score from the Baseline System

The result obtained for each corpus was different, and this is because of the size and type of the data. The data in Corpus I is from Bible which is difficult for machine translation, because the words and phrases in Bible are less probable to found in other sources i.e. the frequency of phrases or words in the corpus are rare. This affects the language model to compute the probability of the next word to follow by predicting one word at a time and the translational model; word translation probabilities. The training and testing sentences were relatively long sentences. This affects the alignment; because as the length of the sentence increase, the possible alignments between source and target sentences increases exponentially. The BLUE score recorded for Corpus II is relatively better, because this corpus contains only simple sentence and the sentence pairs contain almost the same number of words; which is pretty easy for one-to-one alignment. The other corpus is Corpus III prepared from the Constitution of FDRE and simple sentences customized by the researcher from [13], and the result recorded from this corpus was lower than Corpus II and better than Corpus I. This was because of the data; the phrases and words of this corpus are more frequent than Corpus I. The average length of this corpus is almost shortest than Corpus I and longest than Corpus II, and this affects the alignment. Moreover, the fact that most of the sentences are from Corpus II; which are simple sentences, has also effect on the performance. The accuracy recoded from Corpus IV was lower than Corpus II and III, and better than Corpus I, the performance degradation is due

to the data from Corpus I, since the corpus is prepared from bible. Finally the score from Corpus V is recoded and it outperforms Corpus I and IV, this shows preparing the corpus from the average each source is good instead of using all from bible.

The result recorded for English – Tigrigna translation was relatively poor compared to the Tigrigna – English translation. This is because of the following problems observed during the translation such as: non translated words, deletion, wrongly translated, and morphological variations. Sample of the errors are described in Table 5.3. However, this does not mean the errors are not occurred in Tigrigna – English translation, rather the coverage of the errors in the English – Tigrigna translation is high. The errors are occurred because of the following reasons; low training data, morphological complexity of Tigrigna language, and sparse training data. And literatures shows that segmentation of the morphological rich languages solves the problem to some extent. So the second experiment is based on segmentation; described below.

Type	**Sample from the English – Tigrigna translation: using Corpus IV**	**Error Type**
Input	And when you go in, say, may peace be on this house.	Wrongly translated
Output	**ኽሎኹም ። እሐድገልኩም እዚ ።**	
Reference	**ናብ ቤት ምስ ኣቶኹም፡ ሰላም በሉ።**	
Input	The grace of the lord jesus christ be with your spirit.	Deletion
Output	ጸጋ ጐይታና ኢየሱስ ክርስቶስ ምስ ።	
Reference	ጸጋ ጐይታና ኢየሱስ ክርስቶስ ምስ **መንፈስኩም ይኹን። ኣሜን።**	
Input	He was **sentenced** 20 years of **jail**.	Non translated
Output	ንሱ **sentenced** ካብ 20 ዓመት **jail** ።	
Reference	ንሱ ን20 ዓመታት ማእሰርቲ ተፈሪድዎ ።	
Input	Ethiopian airlines is comfy.	Morphological error
Output	መንገዲ ኣየር ኢትዮጵያ **ሙቹውቲ እያ** ።	
Reference	መንገዲ ኣየር ኢትዮጵያ **ሙቹው እዩ** ።	

Table 5.3 Sample of translation errors

5.4 Experiment II: Morph-based System

This is the second group of experiment which was conducted by segmenting the Tigrigna sentences of the training set, test set and language modelling text. The segmentation is performed using a segmentation model which was created for this purpose with the help of Morfessor 1.0 tool. The segmentation model was created using 24251 Tigrigna vocabularies. The segmentation of the Tigrigna corpora has been performed before training and testing steps, using the created segmentation model as input for the unsupervised segmentation model. Table 5.4 describes counts for vocabulary and tokens of the baseline system (unsegmented) and morph-based systems :

Corpus	Size of Tigrigna Tokens		Size of Tigrigna Vocabulary	
	Baseline	Segmented	Baseline	Segmented
Corpus I	86523	107084	20973	11889
Corpus II	4982	5334	1461	1351
Corpus III	15134	16584	4016	3192
Corpus IV	101657	123668	24251	13955
Corpus V	37831	44377	11011	8154

Table 5.4 : Summary counts of token and vocabulary

As it is shown in Table 5.4 the morph-based system contributes to the overall reduction of vocabulary size and increasing the token size of the corpora. The decrease in vocabulary size adds to the overall reduction of rare or unknown words so that we finally get richer vocabulary by solving the data sparsity problem. The percentage decrease of the corpora is 43.4%, 7.52%, 20.5%,, 42.45 % and 26 % for Corpus I, Corpus II, Corpus III, Corpus IV, and Corpus V respectively. Since the segmentation tool is language independent which is unsupervised, the only properties stored for a morph in the lexicon was the frequency (number of occurrences) of the morph in the corpora and the string of letters that the morph consists of. The representation of the string of letters naturally incorporates knowledge about the length of the morph, i.e., the number of letters in the string. This causes a problem of over segmentation and under segmentation.

The translation system have been trained using five segmented different Tigrigna – English training set, and tested using English and segmented Tigrigna source sentences as new sentences that gives an output target Tigrigna and English sentence of. The output has been checked with the

reference sentences, and scored using the BLEU score. The result obtained from the BLUE score is shown in Table 5.5 and snapshot of the result from the system is described in Appendix III.

	Corpus Name	Translation Accuracy in BLUE score	
		English - Tigrigna	Tigrigna – English
Morph-based Translation	Corpus I	2.50	3.37
	Corpus II	18.25	38.85
	Corpus III	13.44	35.66
	Corpus IV	12.52	23.68
	Corpus V	12.55	26.48

Table 5.5 : Summary of BLUE score from the Morph-based System

The reason behind the difference between the records was that there is a difference in the type and size of data. The accuracy obtained using this system shows some degradation. Over and under segmentation problems are observed in the translated output in addition to the problems observed in baseline system. For Example:

No.	**Sentences with over-segmentation**	**Sentences with under-segmentation**
1	ኣስቴር **ሓ ሺ ሸ** ትጥቀም ኣላ ።	እዚ ሰብኣይ እዚ **ብሓቂ** ወዲ ኣምላኽ ነይሩ፡ በለ።
2	መንገዲ ኣየር ኢትዮጵያ **ሙ ቹ ው** እዩ።	ኢየሱስ ከም **ብሓድሽ ብዓብዩ** ድምጺ ጨርሑ፡ ነፍሱ ድማ ወጸት።
3	ንሳ **ዓ ርቢ ተ መሊሳ** ትመጽእ እያ።	እቶም ሰባት **ብምኽንያት** ሓደጋ መኪና ሞይቶም ።
4	**ፌ ሸ ን** ኩዕሶ ትርኢ እያ።	**ሰብኣውን ዴሞክራሲያውን መሰላት** ።
5	እቶም ተማህሮ የጽንዑ **ኣሎ ው** ።	ኣሌክስ **ንኽናደድ** ምኽንያት ነይርዎ ።

Table 5.6 Sample of over and under segmented morphems

From the above table almost the Tigrigna prepositions and conjunctions found in affix are unsegmented. Since the segmentation learner is unsupervised, the segmentation is based on frequency of the morphs. Because of its language independency, those problems were observed during translation other than problems in Experiment I. Over and under segmentation rises the problems observed in Experiment I, and this makes lower the accuracy of the system compared to

the Baseline system. Based on this result the next experiment is conducted, and the main target of the experiment is to solve under and over segmentation of morphemes.

5.5 Experiment III: Post-Processed Segmented System

This is the third group of experiment conducted on Tigrigna – English language pair by using Post-Processed segmentation model. The segmentation model used in the Morph-based experiment pose problem of over and under segmentation. As it is already discussed in Chapter three, there are Tigrigna Prepositions and Conjunctions that are found in the form of affixes in Tigrigna words, but this are found separately in English. The post-processing focused on segmenting these Tigrigna conjunctions and prepositions by post processing the segmentation model created in the above experiment.

5.5.1 Post-Processed Segmentation Model

The post-processing focused on segmenting these Tigrigna conjunctions and prepositions that are found in the form of affixes in Tigrigna words, and words other than these forms are dissegmented to their original form. Segmenting the conjunctions and prepositions, and dissegmenting words other than these, solves the issue of over-segmentation and under-segmentation to some extent. List of Tigrigna prepositions and conjunctions that are found in affixes forms with other words and their examples are shown on table 5.7:

Tigrigna lists	English Meaning
ን	For
ብ	By
ክ	While
ን ን	And
ውን	Also

Table 5.7 : List of Tigrigna conjunctions and prepositions found in affixes form

And the following bullets shows the relationship between the languages :

- **“ ን ን“ “and”** is the most coordinate common coordinate conjunction.it is placed at the end of each word that is to be joined:
    - ጠረቢዛን ወንበርን ኣምጽእ ። Bring a table and chair.
- “**ውን** “ is used to mean “**also**”

- ፈረሰይ‘ውን ሞይቱ ። My horse also is dead.

➢ “ክ” “**while**” for an action that is simultaneous with the main verb

- ክበልዕ ከለኾ ርእዩኒ ። while I was eating he saw me.

➢ “ብ” “**by**” (mainly instrumental, but notice the examples)

- ብበቅሊ መጺኡ ። He came by mule.

➢ “ ን “ “**for**” (to the advantage or disadvantage of something or someone)

- እቲ መግቢ ንኽልቢ እዩ ። The food is for the dog

To process the model using the above list of conjunctions and prepositions Algorithm 5.1 is developed. The algorithm first accepts the list of words from the segmentation model. And check for the target characters weather segmented or not. Characters other than the target characters are dissegment to their original form.

1. Load the words from the segmentation model
2. Store all words in W
3. for each w_j in W do, where j=1,2,3, 4…k
4. if w_j starts with ‘ን’ or ‘ብ’ or ‘ክ’
5. Insert ‘ * ‘ at w_j [1] // after the first character
6. Store w_j in W’
7. End if
8. if w_j ends with ‘ውን’
9. Insert ‘ * ‘ at w_j [n-3] // before the last two characters
10. Store w_j in W’
11. if w_j ends with ‘ን’ and w_j [n-2] is not equal to ‘ው’
12. Insert ‘ * ’ at w_j [n-2] // before the last characters
13. Store w_j in W’
14. End if
15. End for
16. Write W’

Algorithm 5.1: Algorithm for processing the segmentation model

Table 5.8 shows list of input words and output words. The input words are from the segmentation model which are in the left-hand side of the table and the other side of the table shows the output from the system. The “*” was used to identify the character form the word, and a number preceding the word and characters indicates the frequency of the words :

Input: List of segmented words (before processed)	**Output**: List of segmented words (after processed)
6 ልማድ	6 ልማድ
1 ዘይት + ድቅስ	1 ዘይትድቅስ
1 ብሓጐሶም	1 ብ * ሓጐሶም
1 ዚቕጥዖ	1 ዚቕጥዖ
1 ትምህር + ት + ታት	1 ትምህርትታት
1 ን + ኣደይ	1 ን * ኣደይ
2 ዝሞተ + ሉ	2 ዝሞተሉ
1 ነይሮም + ኺ	1 ነይሮምኺ
1 ዚ + ግበር + ከ	1 ዚግበርከ
1 ዚ + ምስክሩ + ልካ	1 ዚምስክሩልካ
1 ክ + ኸ + ዶ	1 ክ * ኸዶ
2 ን + ሓሙሽተ	2 ን * ሓሙሽተ

Table 5.8 : Sample of the processed segmentation model

Once the model is post-processed, sentences are segmented using the post-segmentation learner, it generates segmented sentences but the sentences contain unnecessary symbol ‘ * ’ and ‘1’ as it. So, the post-processor algorithm developed in section 4.4.2.5 implemented to remove this symbols.

The objective of this experiment is to solve the problems occurred in Experiment II. Because the Post Segmented system uses some linguistic information of the Tigrigna language which are conjunctions and prepositions. Since the system is bidirectional the target language is both Tigrigna and English. But in case of English to Tigrigna translation the output is segmented Tigrigna sentences, which is not human readable, so concatenation of morphemes is required in order for the output to be readable by human beings. The concatenation is based on the specific prepositions and conjunctions morphemes that has been conducted on the segmentation. The algorithm used for the concatenation is shown in Algorithm 5.2 below.

1. Load the translated sentences
2. Store all sentences in S
3. for each $s_i \varepsilon S$ do, where i =1,2,3,4 …
4. for each w_j in s_i do, where j = 1,2,3,4 …
5. if w_j == 'ን' and w_{j+4} == 'ን'
6. delete w_{j-1} and w_{j+3}
7. Else if w_j == 'ን'
8. delete w_{j+1}
9. End if
10. if w_j == 'ብ' or w_j == 'ክ'
11. delete w_{j+1}
12. End if
13. if w_j == "ውን"
14. delete w_{j-1} form the word
15. End if
16. End for
17. Store s_i in S'
18. Write S' to file

Algorithm 5.2 : Algorithm for concatenating the Tigrigna target sentences

The post-processed segmented experiment normalize number of vocabulary of the baseline and morph-based experiments. Table 5.9 describes counts for vocabulary and tokens of the baseline (unsegmented), Morph-based and the Post processed segmented corpora.

Corpus Name	Size of Tokens			Size of Vocabulary		
	Baseline	Segmented	Post-Segmented	Baseline	Segmented	Post-Segmented
Corpus I	86523	107084	92080	20973	11889	20104
Corpus II	4982	5334	5299	1461	1351	1417
Corpus III	15134	16584	16870	4016	3192	3626
Corpus IV	101657	123668	108950	24251	13955	22962

Corpus V	37831	44377	40875	11011	8154	10283

Table 5.9: Summary counts of token and vocabulary

As it is shown in Table 5.9: the counts of tokens and vocabulary size of the post-processed segmented corpora is found at an average level. This solves the problem of over segmentation and under segmentation. The data sparseness problem has significantly improved as the vocabulary size decreases.

The trainings have been conducted on Tigrigna – English language pairs by segmenting the Tigrigna sentences includes the training set, test set, and language model using the post-processed segmentation model. And tested using source sentences of English, and Tigrigna sentences segmented using the post-segmentation model as new sentences that gives an output target Tigrigna and English sentences of. The output has been checked with the reference sentences, and scored using the BLEU score. The result obtained from the BLUE score is shown in Table 5.10 and snapshot of the result from the system is described in Appendix IV.

	Corpus Name	Translation Accuracy in BLUE score	
		English - Tigrigna	Tigrigna - English
Post-Segment-based Translation	Corpus I	12.92	22.38
	Corpus II	22.46	53.58
	Corpus III	20.05	39.76
	Corpus IV	14.84	25.54
	Corpus V	15.44	31.03

Table 5.10 : Summary of BLUE score from the Post-processed segmented system

The final experiment shows an improvement in BLUE score for each corpus of the Post-Segmented System over the other two experiments as it is shown in the above table. This clearly shows that the Post segmented system outperforms all the other systems.

The experiments are conducted by using three different systems with similar corpora. From the result of the experiments we can see that the result recorded from a BLEU score shows that the result returns by the post processed segmented system is better than the baseline and segmented systems for Tigrigna – English language pair. We can also see that better translation accuracy is

acquired when Tigrigna is used as a source sentences and English is used as a target sentence in each experiment, this is because of the morphology. Tigrigna is a language that is morphologically rich and it could produce more accurate results if it is the source language rather than being the target language. Since translating to a more morphologically complex is a more difficult task, where there is a higher chance of translating word inflection incorrectly. Comparison between Tigrigna – English and English – Tigrigna translation using the same sentence; from the post-processed segmentation system:

- English – Tigrigna translation
    - Input sentence: He has an exam.
    - Output sentence: ንሱ ፈተና ኣለዋ። // this is for women
    - Correct translation: ንሱ ፈተና ኣለዎ። // this is for men
- Tigrigna – English translation
    - Input sentence: ንሱ ፈተና ኣለዎ ።
    - Output sentence: He has an exam.
    - Correct translation: He has an exam.

5.6 Finding

We understand that the vocabulary size reduction from the baseline system and increment of vocabulary size from the segmented system has contributed for the overall performance of the post segmented system that has shown better performance compared to the other systems. Post processing has played a major role for the improved performance of the post segmented system by reducing the vocabulary size and increasing the frequency of the tokens. And the reduction of vocabulary size contributes to address the data sparseness and absence of one to one correspondences between source and target words.

Despite the fact that there is not any existing MT system for Tigrigna – English language pair with which one can compare the result of this system, the researchers believes that comparing the score among the three systems (fair comparison). Accordingly, the result

obtained from the post processed experiment using corpus II has outperformed the other experiments. And Tigrigna – English translation shows better accuracy than English – Tigrigna

translation. Therefore, one can conclude that translation based on processed segmentation can achieve better translation accuracy.

CHAPTER SIX

Conclusion and Recommendation

6.1 Conclusion

The Statistical machine translation research for many languages is getting more attention than the other machine translation approaches. Since the statistics based approach requires limited computational linguistic resources compared to the other approaches that might take much time to develop some linguistic resources. So it is recommended for resource scarce languages like Tigrigna. The purpose of this study is to investigate the development of a bidirectional Tigrigna – English machine translation system using statistical approach. Therefore, this research has been conducted by developing thirty types of experiments all based on Tigrigna - English and English – Tigrigna Statistical based Machine Translation.

In this research, the basic research questions have been addressed and the output results have been reported. A Tigrigna – English and English – Tigrigna translation using a baseline experiment was conducted by collecting parallel corpora in both Tigrigna and English language. In order to answer the research questions, this experiment has been conducted and served as the basis for the other experiments.

The first research question was " Does the type and size of corpus has impact on the performance of the machine translation system of these language pairs ? ". To look the size and type of the corpus has an impact on the translation system, Tigrigna – English parallel corpora have been collected from different sources. The collected sentences were 6016 from bible, 1160 from [13], and 822 from the constitution of FDRE. From this five parallel corpora have been prepared; by combining one with the other. The corpus prepared from simple sentences perform better than the other; this shows that the type of the corpus has an impact on the translation. Corpus IV and V are prepared from similar sources with different size. And the result obtained from Corpus IV was promising than Corpus V.

This research question was the basis to answer the other research questions. Accordingly, more experimentation on the application of different methods to these corpora includes morph-based, and post-processed segmented systems have answered the rest research questions. We also see potential of unsupervised word segmentation to improve the translation, and is very close to

outperform the baseline system. To improve this post-processing was applied on Tigrigna morphemes incudes; prepositions and conjunctions. The result obtained from this post-processing method outperforms the other systems. And better translation accuracy achieved when Tigrigna used as a source sentences and English as a target sentences. Since the post-processing method only segments prepositions and conjunctions of Tigrigna, we propose that semi-supervised segmentations for Tigrigna has more potential to improve the Tigrigna – English translation.

The challenge that faces in this work was to get Tigrigna – English bilingual corpus. As Tigrigna is one of the language that suffer severely from lack of computational resources. And it also hurdle to process the available Tigrigna resources because of the format; the most challenging task was converting from the RTF to Unicode text file due to the characteristics of the Tigrigna writing system that uses the Ge'ez script.

Finally, it is important to give a direction for future experiment. The direction should be in the application of methods that help to get semi-supervised segmentation model to segment Tigrigna morphology as the processed segmentation experiment outperformed the other experiments (baseline and morph-based experiments). Since this method only segments prepositions and conjunctions, there should be a mechanism to apply more techniques to segment the other parts-of-speech as well.

6.2 Recommendation

Based on the findings of this study and the knowledge obtained from the literature, the following recommendations are forwarded for future work.

- The segmentation of only preposition and conjunctions has led to a huge gain in BLUE score. Supervised segmentation of other derivational and inflectional morphs of Tigrigna language may lead to further improvement of the translation quality. This can be an area of study towards improving performance of a translation system for this language pair.
- Extending the approach using factored based translation. Factored based translation models helps to represent phrases not simply as sequences of fully inflected words, but instead as sequences containing multiple levels of information, and this is better for languages like Tigrigna to handle the effect of morphology.
- Further results can be accomplished by increasing the size and domain of the data set used for training the system. Due to the challenge in preparing the corpus, we have prepared a

corpus of limited size, on the other hand, statistical approaches are data greedy. So, by increasing the size of the training data set and integrating the linguistic information one can develop a full-fledged bidirectional Tigrigna -English machine translation.

- Further researches in machine translation on Tigrigna to other languages, even using languages in Ethiopia such as Amharic, Affan Oromo or so could be performed while preparing a large corpus. Because the Tigrigna part is available, especially for Tigrigna – Amharic and Amharic Tigrigna translation.

References

[1] ED. Liddy, "Natural Language Processing", *In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc*, 2001.

[2] S. J. Karen, "Natural Language Processing, a historical review," *in W. Bright (ed.) International encyclopedia of linguistics, New York: Oxford University Press*, 1992, Vol. 3, 53–59.

[3] F. J. Och, "Challenges in Machine Translation", in*: Proceeding ISCSLP'06 Proceedings of the 5th international conference on Chinese Spoken* , Singapore, December, 2006 .

[4] P. Lewis, F. Simon and D. Fennin, "Ethnologue: Language of the world," *in 17th edn,SIL international, Dallas,* 2013.

[5] Tigrigna, (ትግርኛ), "http://www.omniglot.com/writing/tigrinya.htm," Last Accessed on February 28,2017.

[6] M. Osborne, "Statistical Machine Transilation," Unpublished Masters Thesis, Department of Computer Science, University of Edinburgh, Edinburgh.

[7] A. Lopez and P. Post, "Human Language Technology Center of Excellence," In Twenty Years of Bitext, 2013.

[8] R. Bear, "M.ED. In TESEL Program, language specific informational reports," Rhode Island College, 2011.

[9] S. Mydans, "Acorss culture, English is the Word", New York Times, September 2011.

[10] A. Mouiad, O. Nazlia and M. Tengku, "Machine Translation from English to Arabic", *IPCBEE vol.11,* 2011.

[11] S. Takahashi, H. Wada, R. Tadenuma and S. Watanabe, "Machine translation systems developed by the Electrotechnical Laboratory of Japan", Future computing systems 2.3 (1989): 275-291.

[12] H. Schwenk, JB. Fouet and J.Senellart, "First Steps towards a general purpose French/English Stastical Machine Translation System," *In: proceedings of the third workshop on statistical machine translation, Ohio,* 2008.

[13] Eleni Teshome, "Bidirectional English - Amharic Machine Translation: An Experiment using Constrained Corpus ", Unpublished Masters Thesis, Departement of Computer Science, Addis Ababa University, 2013 .

[14] Mulu Gebreegziabher and L. Besacier. " English - Amharic Statistical Machine Translation", SLTU, 2012.

[15] Sisay Adugna "English – Oromo Machine Translation: An Experiment Using a Statistical Approach," Unpublished Masters Thesis, School of information Science, Addis Ababa University, 2009.

[16] Jabesa Daba and Yaregal Assibe , "A Hybrid Approach to the Development of Bidirectional English - Oromiffa", *In Proceedings of the 9th International Conference on Natural Language Processing (PolTAL2014),, Warsaw, Poland*, 2014.

[17] N. Bertoldi and FBK-irst , "IRSTLM Toolkit, Fifth MT Marathon, Le Mans, France," 13-18 September 2010.

[18] K. Philipp, F. Marcello, S. Wade, B. Nicol, B. Ondrej, CB. Chris, B. Cowan, C. Alexandra, C.C. Moran and H. Evan, "Final Report of the 2006 Language Engineering Workshop," Johns Hopkins University: Center for Speech and Language Processing, September 3, 2007.

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

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

[21] M.D. Okpor, "Machine Translation Approaches: Issues and Challenges," *International Journal of Computer Science Issues*, vol. Vol. 11, no. Issue 5, September 2014.

[22] D. Jurafsky and J.H. Martin, "Speech and language processing", International Edition 710 (2000).

[23] C. Michael and A. Way, "Recent advances in example-based machine translation", Springer Science & Business Media, Vol. 21, 2003.

[24] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J.Della Pietra, F. Jelinek, J.D. Lafferty, R.L.Mercer and P.S. Roossin, "A Statistical Approach To Machine Translation", *Computational Linguistics*, pp. 79-85, June 1990.

[25] P.F. Brown, V.J. Della Pietra, S.A. Della Pietra, and R L. Mercer, "The Mathematics of Statistical Machine: Parameter Estimation," *Association for Computational Linguistics*, 1993.

[26] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?", *Proceedings of the IEEE 88*, pp. 1270-1278, 2000.

[27] A.E. Axelrod, "Factored Language Models for Statistical Machine Translation", Unpublished Masters Thesis, University of Edinburgh, 2006.

[28] C.D Manning and H. Schiitze, "Foundations of statistical natural language processing ", Cambridge: MIT press, Vol.999, 1999.

[29] J.A. Bilmes and K. Kirchhoff. , "Factored Language Models and Generalized Parallel Backoff ", in *Proceedings of the Conference of the North America Chapter of the Association for Computational lingustics on Human Language Technology,* Canada, 2003.

[30] G. Jianfeng "A Quirk Review of Translation Models," 2011.

[31] A. . L. and P. R. , "Word-Based Alignment, Phrase-Based Translation: What's the Link?," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* Cambridge, August 2006.

[32] K. Yamada and K. Knight , "A Syntax-based Statistical Translation Model," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2001.*

[33] P. Koehn and H. Hoang. , "Factored Translation Models," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational,* 2007.

[34] A. Ramanathan. " Simple Syntactic and Morphological Processing Can Help English-Hindi Statistical Machine Translation ", IJCNLP. 2008.

[35] JJJ. Brunning "Alignment Models and Algorithms for Statistical Machine Translation", Diss. University of Cambridge, 2010.

[36] S.Vogel, H. Ney and C. Tillmann, "HMM-Based Word Alignment in Statistical Translation", *In Proceedings of COLING,* pp. 836-841, 1996.

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

[38] S.Sereewattana, "Unsupervised Segmentation for Statistical Machine Translation", Unpapulishied Master Thesis, School of Informatics, University of Edinburgh, 2003.

[39] I Badr, R Zbib, and J Glass , "Segmentation for English-to-Arabic statistical machine translation." *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Association for Computational Linguistics*, 2008.

[40] K.Sachdeva, R.Srivastava and S.Jain, "Hindi to English Machine Translation: Using Effective Selection in Multi-Model SMT", LREC. 2014.

[41] O.Bojar, "English-to-Czech Factored Machine Translation", *Association for Computational Linguistics*, p. 232–239, June 2007.

[42] P. Porkaew, A. Takhom and T. Supnithi, "Factored Translation Model in English-to-Thai Translation", *Eighth International Symposium on Natural Language Processing*, 2009.

[43] CSA, "Population and Housing Census 2007 Report, Tigray Region," Central Statistical Agency, Ethiopia, 2007.

[44] Yonas Fisaha, "Development of Stemming Algorithm For Tigrigna Text," Unpublished Masters Thesis, School of Information Science, Addis Ababa University, 2011.

[45] Daniel Tekilu, "ዘበናዊ ሰዋስው ቋንቋ ትግርኛ ", Mega printing, Mekelle, 2008.

[46] J. Manson, "Tigrinya Grammar", The Red Sea Press, 1996.

[47] Hailay Beyene, "Design And Development Of Tigrigna Search Engine", Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, 2013.

[48] Teklay Gebregzabiher "Part of Speech Tagger For Tigrigna Language," Unpublished Masters Thesis, Department of Computer Science, Addis Ababa University, 2010.

[49] H. Rodney and GK. Pullum, " The cambridge grammar of english ", Cambridge, Cambridge University Press, 2002.

[50] TK. Pratt, "The Cambridge Encyclopedia of the English Language", Cambridge, Cambridge University Press, 1995.

[51] H. Rodney. and GK. Pullum, " The cambridge grammar of english ", Cambridge: Cambridge University Press, 2002.

[52] V. Dániel, N. László, H. Péter, K. András, T. Viktor and N. Viktor , "Parallel corpora for medium density languages.," in *In Proceedings of the Recent Advances in Natural Language Processing*, Borovets, Bulgaria, 2005.

[53] MOSES, "Moses: open source toolkit for statistical machine translation system," in http://www.statmt.org/moses/; 2017.

[54] H. Alshikhabobakr, "Unsupervised Arabic Word Segmentation and Statistical Machine Translation," Qatar Foundation Annual Research Conference. No. 2013. 2013.

[55] M. Creutz and K. Lagus, "Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0", Computer and Information Science, Report A81, Helsinki University of Technology, 2005.

[56] M. Federico, N. Bertoldi and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," *In Interspeech, pp. 1618-1621. 2008.*

Appendices

Appendix I : Tigrigna Alphabet (Fidel)

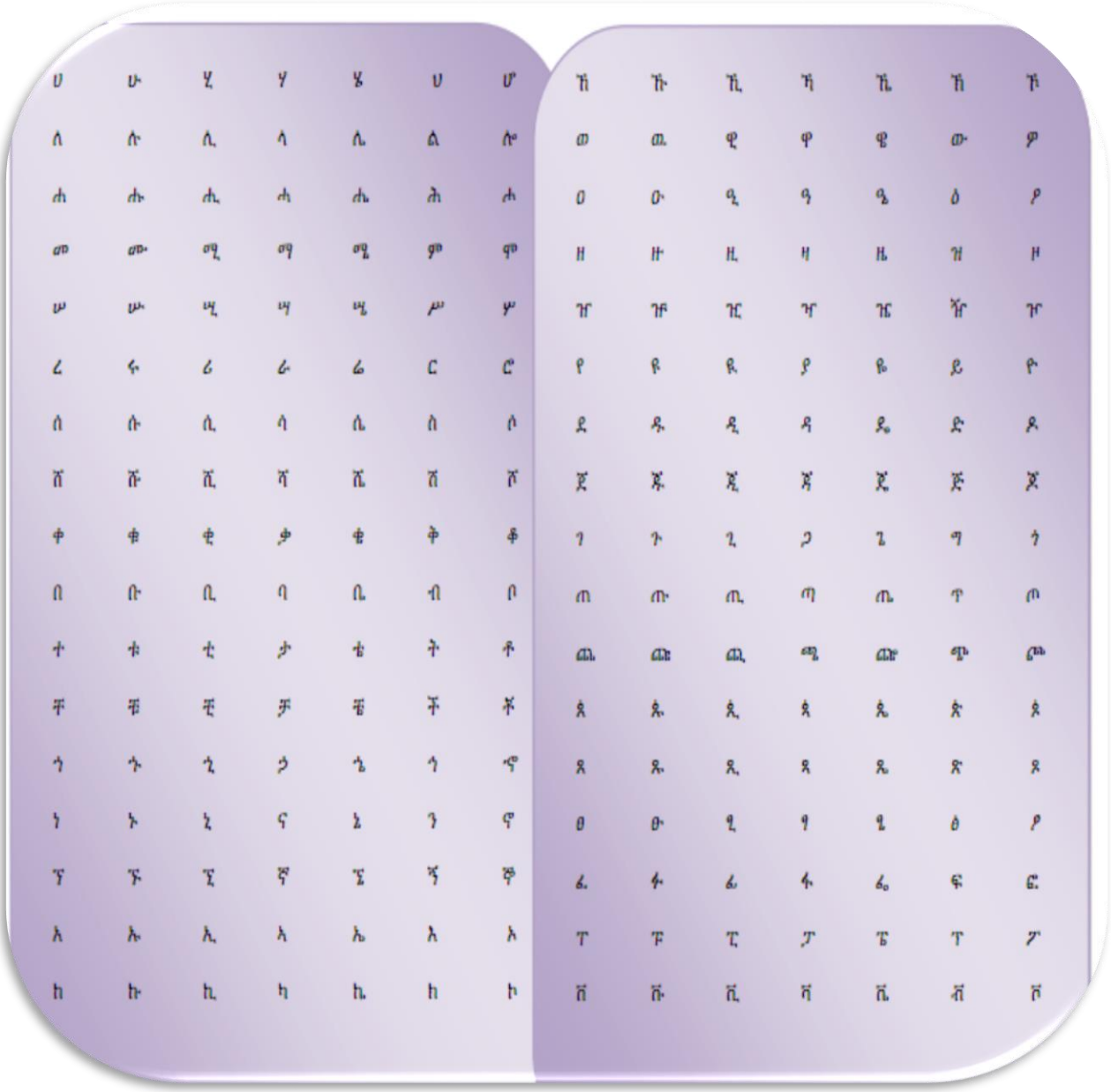


Figure I.1: Tigrigna Alphabet (Fidel)

Appendix II : Snapshot Result From Baseline System

Corpus I

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  1.6914   1.7834   1.7903   1.7903   1.7903   1.7903   1.7903   1.7903   1.7903  "moses"

 BLEU:  0.2137   0.1425   0.1059   0.0800   0.0589   0.0454   0.0359   0.0268   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 17:24:48
========================================================================
```

Figure II.1: English – Tigrigna BLEU score result from Corpus I

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  1.7173   2.0100   2.0434   2.0469   2.0488   2.0488   2.0488   2.0488   2.0488  "moses"

 BLEU:  0.3359   0.2370   0.1809   0.1450   0.1188   0.0981   0.0799   0.0621   0.0494  "moses"
MT evaluation scorer ended on 2017 May 10 at 17:05:40
========================================================================
```

Figure II.2: Tigrigna – English BLEU score result from Corpus I

Corpus II

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.5874   2.6819   2.6819   2.6819   2.6819   2.6819   2.6819   2.6819   2.6819  "moses"

 BLEU:  0.3816   0.2455   0.2055   0.0748   0.0000   0.0000   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 Jun 8 at 02:15:06
========================================================================
```

Figure II.3: English – Tigrigna BLEU score result from Corpus II

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  3.8046   4.3284   4.3643   4.3696   4.3778   4.3778   4.3778   4.3778   4.3778  "moses"

 BLEU:  0.6202   0.4978   0.4275   0.3686   0.3113   0.2542   0.1793   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 Jun 8 at 23:02:50
========================================================================
```

Figure II.4: Tigrigna – English BLEU score result from Corpus II

Corpus III

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.3441   2.4522   2.4522   2.4522   2.4522   2.4522   2.4522   2.4522   2.4522  "moses"

 BLEU:  0.3286   0.2119   0.1865   0.0850   0.0592   0.0441   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 17:51:06
=====================================================================
```

Figure II.5: English – Tigrigna BLEU score result from Corpus III

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  3.3004   3.7346   3.7578   3.7776   3.8045   3.8045   3.8045   3.8045   3.8045  "moses"

 BLEU:  0.5028   0.4160   0.3640   0.3151   0.2626   0.2079   0.1553   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 17:45:45
=====================================================================
```

Figure II.6 Tigrigna – English BLEU score result from Corpus III

Corpus IV

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.1026   2.2148   2.2209   2.2224   2.2240   2.2261   2.2261   2.2261   2.2261  "moses"

 BLEU:  0.2526   0.1650   0.1241   0.0895   0.0704   0.0551   0.0413   0.0296   0.0000  "moses"
MT evaluation scorer ended on 2017 Jun 9 at 17:12:15
=====================================================================
```

Figure II.7: English – Tigrigna BLEU score result from Corpus IV

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.3815   2.8947   2.9494   2.9619   2.9712   2.9737   2.9737   2.9737   2.9737  "moses"

 BLEU:  0.3866   0.2997   0.2464   0.2071   0.1750   0.1481   0.1251   0.1068   0.0918  "moses"
MT evaluation scorer ended on 2017 Jun 9 at 08:56:14
=====================================================================
```

Figure II.8 Tigrigna – English BLEU score result from Corpus IV

Corpus V

```
# ------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  1.5419   1.5934   1.5934   1.5934   1.5934   1.5934   1.5934   1.5934   1.5934  "moses"

 BLEU:  0.2034   0.1571   0.1394   0.0853   0.0508   0.0429   0.0374   0.0318   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 18:19:40
==================================================================
```

Figure II.9: English – Tigrigna BLEU score result from Corpus V

```
# ------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  1.9747   2.2535   2.2714   2.2811   2.2942   2.2942   2.2942   2.2942   2.2942  "moses"

 BLEU:  0.3336   0.2714   0.2381   0.2119   0.1847   0.1593   0.1358   0.1144   0.0934  "moses"
MT evaluation scorer ended on 2017 May 10 at 18:10:00
==================================================================
```

Figure II.10: Tigrigna – English BLEU score result from Corpus V

Appendix III: Snapshot Result from the Morph-based System

Corpus I

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  0.8876   0.9390   0.9390   0.9390   0.9390   0.9390   0.9390   0.9390   0.9390  "moses"

 BLEU:  0.1187   0.0545   0.0250   0.0000   0.0000   0.0000   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 19:32:08
=====================================================================
```

Figure III.1: English – Tigrigna BLEU score result from Corpus I

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  0.2438   0.2686   0.2701   0.2701   0.2701   0.2701   0.2701   0.2701   0.2701  "moses"

 BLEU:  0.1249   0.0599   0.0337   0.0222   0.0143   0.0099   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 19:12:47
=====================================================================
```

Figure III.2: Tigrigna – English BLEU score result from Corpus I

Corpus II

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.2203   2.3195   2.3195   2.3195   2.3195   2.3195   2.3195   2.3195   2.3195  "moses"

 BLEU:  0.3174   0.2393   0.1825   0.0997   0.0000   0.0000   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 Jun 10 at 01:42:23
=====================================================================
```

Figure III.3: English – Tigrigna BLEU score result from Corpus II

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  3.6488   4.1810   4.1974   4.1974   4.1974   4.1974   4.1974   4.1974   4.1974  "moses"

 BLEU:  0.5293   0.4520   0.3885   0.3181   0.3404   0.2698   0.1914   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 Jun 10 at 01:39:50
=====================================================================
```

Figure III.4: Tigrigna – English BLEU score result from Corpus II

Corpus III

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  1.9489   2.0156   2.0156   2.0156   2.0156   2.0156   2.0156   2.0156   2.0156  "moses"

 BLEU:  0.2536   0.1591   0.1344   0.0493   0.0000   0.0000   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 22:46:22
========================================================================
```

Figure III.5: English – Tigrigna BLEU score result from Corpus III

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  3.2117   3.6686   3.6887   3.6975   3.7088   3.7088   3.7088   3.7088   3.7088  "moses"

 BLEU:  0.5283   0.4241   0.3566   0.2946   0.2356   0.1846   0.1513   0.1255   0.1051  "moses"
MT evaluation scorer ended on 2017 May 10 at 22:40:51
========================================================================
```

Figure III.6: Tigrigna – English BLEU score result from Corpus III

Corpus IV

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  3.2093   3.8719   3.9326   3.9458   3.9526   3.9553   3.9553   3.9553   3.9553  "moses"

 BLEU:  0.3924   0.3116   0.2368   0.1959   0.1438   0.1593   0.1322   0.1126   0.0975  "moses"
MT evaluation scorer ended on 2017 Jun 10 at 09:21:17
========================================================================
```

Figure III.7: English – Tigrigna BLEU score result from Corpus IV

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.3061   2.4677   2.4727   2.4742   2.4759   2.4780   2.4780   2.4780   2.4780  "moses"

 BLEU:  0.2615   0.1727   0.1252   0.0927   0.0703   0.0525   0.0339   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 Jun 10 at 09:44:30
========================================================================
```

Figure III.8: Tigrigna – English BLEU score result from Corpus IV

Corpus V

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  1.9014   1.9970   1.9970   1.9970   1.9970   1.9970   1.9970   1.9970   1.9970  "moses"

 BLEU:  0.2392   0.1553   0.1255   0.0648   0.0470   0.0329   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 23:10:49
=====================================================================
```

Figure III.9: English – Tigrigna BLEU score result from Corpus V

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.8925   3.2964   3.3139   3.3193   3.3259   3.3259   3.3259   3.3259   3.3259  "moses"

 BLEU:  0.4181   0.3272   0.2648   0.2099   0.1602   0.1207   0.0950   0.0756   0.0607  "moses"
MT evaluation scorer ended on 2017 May 10 at 22:59:32
=====================================================================
```

Figure III.10: Tigrigna – English BLEU score result from Corpus V

Appendix IV: Snapshot Result From Post-processed segmented system

Corpus I

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  1.8411   1.9663   1.9735   1.9755   1.9755   1.9755   1.9755   1.9755   1.9755  "moses"

 BLEU:  0.2530   0.1603   0.1292   0.0630   0.0373   0.0000   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 23:47:33
========================================================================
```

Figure IV.1: English – Tigrigna BLEU score result from Corpus I

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  1.8719   2.2532   2.2931   2.2998   2.3028   2.3051   2.3051   2.3051   2.3051  "moses"

 BLEU:  0.3624   0.2755   0.2238   0.1850   0.1562   0.1338   0.1148   0.0993   0.0872  "moses"
MT evaluation scorer ended on 2017 May 10 at 23:31:59
========================================================================
```

Figure IV.2: Tigrigna – English BLEU score result from Corpus I

Corpus II

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.4801   2.6278   2.6278   2.6278   2.6278   2.6278   2.6278   2.6278   2.6278  "moses"

 BLEU:  0.3577   0.2877   0.2246   0.1415   0.0000   0.0000   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 Jun 10 at 16:39:35
========================================================================
```

Figure IV.3: English – Tigrigna BLEU score result from Corpus II

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  4.0087   4.5754   4.5937   4.6017   4.6132   4.6132   4.6132   4.6132   4.6132  "moses"

 BLEU:  0.6692   0.5976   0.5358   0.4681   0.3814   0.2899   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 Jun 10 at 16:37:32
========================================================================
```

Figure IV.4: Tigrigna – English BLEU score result from Corpus II

Corpus III

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.6070   2.8220   2.8220   2.8220   2.8220   2.8220   2.8220   2.8220   2.8220  "moses"

 BLEU:  0.3442   0.2322   0.2005   0.1329   0.0849   0.0715   0.0634   0.0554   0.0000  "moses"
MT evaluation scorer ended on 2017 May 11 at 00:05:05
========================================================================
```

Figure IV.5: English – Tigrigna BLEU score result from Corpus III

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  3.5912   4.0691   4.0944   4.1134   4.1388   4.1388   4.1388   4.1388   4.1388  "moses"

 BLEU:  0.5531   0.4598   0.3976   0.3402   0.2794   0.2216   0.1727   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 10 at 23:59:36
========================================================================
```

Figure IV.6: Tigrigna – English BLEU score result from Corpus III

Corpus IV

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.0524   2.2156   2.2218   2.2233   2.2250   2.2271   2.2271   2.2271   2.2271  "moses"

 BLEU:  0.2609   0.1869   0.1484   0.0832   0.0619   0.0486   0.0389   0.0325   0.0277  "moses"
MT evaluation scorer ended on 2017 Jun 10 at 16:17:11
========================================================================
```

Figure IV.7: English – Tigrigna BLEU score result from Corpus IV

```
# ------------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.4844   3.0013   3.0544   3.0647   3.0711   3.0736   3.0736   3.0736   3.0736  "moses"

 BLEU:  0.3910   0.3019   0.2554   0.2038   0.1697   0.1416   0.1193   0.1012   0.0861  "moses"
MT evaluation scorer ended on 2017 Jun 10 at 15:56:06
========================================================================
```

Figure IV.8: Tigrigna – English BLEU score result from Corpus IV

Corpus V

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  2.0615   2.2103   2.2103   2.2103   2.2103   2.2103   2.2103   2.2103   2.2103  "moses"

 BLEU:  0.2739   0.1925   0.1544   0.1011   0.0000   0.0000   0.0000   0.0000   0.0000  "moses"
MT evaluation scorer ended on 2017 May 11 at 00:28:14
=====================================================================
```

Figure IV.9: English – Tigrigna BLEU score result from Corpus V

```
# ---------------------------------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram   6-gram   7-gram   8-gram   9-gram
        ------   ------   ------   ------   ------   ------   ------   ------   ------
 NIST:  3.1433   3.6405   3.6679   3.6786   3.6919   3.6919   3.6919   3.6919   3.6919  "moses"

 BLEU:  0.4538   0.3685   0.3103   0.2601   0.2103   0.1679   0.1333   0.1069   0.0827  "moses"
MT evaluation scorer ended on 2017 May 11 at 00:20:19
=====================================================================
```

Figure IV.10: Tigrigna – English BLEU score result from Corpus V

Appendix V: Sample Parallel Corpus for training from Corpus I

መጽሓፍ ወለዶ ኢየሱስ ክርስቶስ ወዲ ዳዊት፡ ወዲ ኣብርሃም።	The book of the generations of Jesus Christ, the son of David, the son of Abraham
ኣብርሃም ንይስሃቅ ወለደ፡ ይስሃቅ ንያእቆብ ወለደ፡ ያእቆብ ንይሁዳን ነሕዋቱን ወለደ፡	The son of Abraham was Isaac; and the son of Isaac was Jacob; and the sons of Jacob were Judah and his brothers;
ይሁዳ ኸኣ ካብ ትእማር ንፋሬስን ንዛራን ወለደ፡ ፋሬስ ንኤስሮም ወለደ፡ ኤስሮም ንኣራም ወለደ፡	And the sons of Judah were Perez and Zerah by Tamar; and the son of Perez was Hezron; and the son of Hezron was Ram;
ኣራም ንኣሚናዳብ ወለደ፡ ኣሚናዳብ ንነኣሶን ወለደ፡ ነኣሶን ንሰልሞን ወለደ፡	And the son of Ram was Amminadab; and the son of Amminadab was Nahshon; and the son of Nahshon was Salmon;
ሰለሙን ካብ ራኬብ ንቦኣዝ ወለደ፡ ቦኣዝ ካብ ሩት ንእዮቤድ ወለደ፡ እዮቤድ ንእሴይ ወለደ።	And the son of Salmon by Rahab was Boaz; and the son of Boaz by Ruth was Obed; and the son of Obed was Jesse;
እሴይ ንንጉስ ዳዊት ወለደ፡ ንጉስ ዳዊት ካብ ሰበይቲ ኡርያ ንሰሎሞን ወለደ።	And the son of Jesse was David the king; and the son of David was Solomon by her who had been the wife of Uriah;
ሰሎሞን ንሮብኣም ወለደ። ሮብኣም ንኣብያ ወለደ፡ ኣብያ ንኣሳ ወለደ፡	And the son of Solomon was Rehoboam; and the son of Rehoboam was Abijah; and the son of Abijah was Asa;
ኣሳ ንእዮሳፋጥ ወለደ፡ እዮሳፋጥ ንእዮራም ወለደ፡ እዮራም ንኡዝያ ወለደ፡	And the son of Asa was Jehoshaphat; and the son of Jehoshaphat was Joram; and the son of Joram was Uzziah;
ኡዝያ ንእዮኣታም ወለደ፡ እዮኣታም ንኣካዝ ወለደ፡ ኣካዝ ንህዝቅያስ ወለደ፡	And the son of Uzziah was Jotham; and the son of Jotham was Ahaz; and the son of Ahaz was Hezekiah;
ህዝቅያስ ንምናሴ ወለደ፡ ምናሴ ንኣሞን ወለደ፡ ኣሞን ንእዮስያስ ወለደ፡	And the son of Hezekiah was Manasseh; and the son of Manasseh was Amon; and the son of Amon was Josiah;
እዮስያስ ከኣ ብዘመን ምርኮ ባቢሎን ንኢኮንያን ነሕዋቱን ወለደ።	And the sons of Josiah were Jechoniah and his brothers, at the time of the taking away to Babylon.
ድሕሪ ምርኮ ባቢሎን ድማ ኢኮንያ ንሰላትኤል ወለደ፡ ሰላትኤል ንዘሩባቤል ወለደ፡	And after the taking away to Babylon, Jechoniah had a son Shealtiel; and Shealtiel had Zerubbabel;
ኣዛር ንሳዶቅ ወለደ፡ ሳዶቅ ንኣኪም ወለደ፡ ኣኪም ንኤልዩድ ወለደ፡	And Azor had Zadok; and Zadok had Achim; and Achim had Eliud;
ኤልዩድ ንኣልኣዛር ወለደ፡ ኣልኣዛር ንማታን ወለደ፡ ማታን ንያእቆብ ወለደ፡	And Eliud had Eleazar; and Eleazar had Matthan; and Matthan had Jacob;
ያእቆብ ንዮሴፍ ሕጹይ ማርያም ወለደ፡ ካብኣ ድማ እቲ ክርስቶስ ዚበሀል ዘሎ ኢየሱስ ተወለደ።	And the son of Jacob was Joseph the husband of Mary, who gave birth to Jesus, whose name is Christ.
እምበኣርሲ ኹሉ ወለዶ ኻብ ኣብርሃም ክሳዕ ዳዊት ዓሰርተው ኣርባዕተ ወለዶ፡ ካብ ዳዊት ክሳዕ ምርኮ ባቢሎን ዓሰርተ ኣርባዕተ ወለዶ፡ ካብ ምርኮ ባቢሎን ክሳዕ ክርስቶስ ድማ ዓሰርተው ኣርባዕተ ወለዶ እዩ።	So all the generations from Abraham to David are fourteen generations; and from David to the taking away to Babylon, fourteen generations; and from the taking away to Babylon to the coming of Christ, fourteen generations.
ልደት ኢየሱስ ክርስቶስ ከምዚ እዩ፡ ኣዲኡ ማርያም ንዮሴፍ ተሐጽያ ኸላ፡ ከይተራኸቡ፡ ብመንፈስ ቅዱስ ጠኒሳ ተረኽበት።	Now the birth of Jesus Christ was in this way: when his mother Mary was going to be married to Joseph, before they came together the discovery was made that she was with child by the Holy Spirit.
ዮሴፍ ሕጹያ፡ ጻድቕ ሰብኣይ ነበረ እሞ፡ ብሕቡእ ኪሐድጋ ሓሰበ እምበር፡ ኪገልጻ ኣይፈተወን።	And Joseph, her husband, being an upright man, and not desiring to make her a public example, had a mind to put her away privately.

ንሱ እዚ ኺሐስብ ከሎ፡ እንሆ፡ መልኣኽ እግዚኣብሄር ብሕልሚ ተራእዩ፡ ከምዚ ኢሉ፡ ዮሴፍ ወዲ ዳዊት፡ እቲ ሕጽይትኻ ማርያም ጠኒሳቶ ዘላ ኻብ መንፈስ ቅዱስ እዩ እሞ፡ ንምውሳዳ ኣይትፍራህ።	But when he was giving thought to these things, an angel of the Lord came to him in a dream, saying, Joseph, son of David, have no fear of taking Mary as your wife; because that which is in her body is of the Holy Spirit.
ወዲ ኽትወልድ እያ፡ ንሱ ኽኣ ንህዝቡ ኻብ ሓጢኣቶም ኬድሕኖም እዩ እሞ፡ ስሙ ኢየሱስ ክትሰምዮ ኢኻ።	And she will give birth to a son; and you will give him the name Jesus; for he will give his people salvation from their sins.
እዚ ዅሉ እቲ እግዚኣብሄር ብነብዩ እተዛረቦ ኺፍጸም ኮነ፡ ከምዚ ዚብል፡	Now all this took place so that the word of the Lord by the prophet might come true,
እንሆ፡ ድንግል ክትጠንስ ወዲውን ክትወልድ እያ፡ ስሙ ኽኣ ኣማኑኤል ኪብልዎ እዮም፡ ትርጉሙ፡ ኣምላኽ ምሳና እዩ።	See, the virgin will be with child, and will give birth to a son, and they will give him the name Immanuel, that is, God with us.
ዮሴፍ ካብ ድቃሱ ተንሲኡ፡ መልኣኽ እግዚኣብሄር ከም ዝኣዘዞ ገበረ፡ ንማርያም ሕጽይቱ ወሰዳ።	And Joseph did as the angel of the Lord had said to him, and took her as his wife;
በዃሪ ወዳ ኽሳዕ እትወልድ ከኣ ኣይፈለጣንኡ። ስሙ ድማ ኢየሱስ ሰመዮ።	And he had no connection with her till she had given birth to a son; and he gave him the name Jesus.
ንጉስ ሄሮዶስ ሰሚዑ ሰምበደ፡ ኵላ ኢየሩሳሌምውን ምስኡ።	And when it came to the ears of Herod the king, he was troubled, and all Jerusalem with him.
ድሕርዚ ሄሮድስ ንሰብ ጥበብ ብሕቡእ ጸውዖም፡ ካባታቶም ድማ ዘመን እቲ እተራእዮም ኰኽብ ኣጸቢቑ ተረድኤ።	Then Herod sent for the wise men privately, and put questions to them about what time the star had been seen.
ንሳቶም ከኣ ካብ ንጉስ ሰሚዖም ከዱ። እንሆ፡ እቲ ኣብ ምብራቕ ዝረኣይዎ ኰኽብ ኣብ ልዕሊ እታ እቲ ሕጻን ዘለዋ ቦታ መጺኡ ደው ክሳዕ ዚብል መርሖም።	And after hearing the king, they went on their way; and the star which they saw in the east went before them, till it came to rest over the place where the young child was.
ነቲ ኰኽብ ምስ ረኣይዎ፡ ኣዝዮም ብዙሕ ሓጎስ ተሐጎሱ።	And when they saw the star they were full of joy.
ናብ ቤት ኣትዮም፡ ነቲ ሕጻን ምስ ማርያም ኣዲኡ ረኣይዎ፡ ፍግም ኢሎም ድማ ሰገድሉ። ሳጽናም ከፊቶም ከኣ ወርቅን ዕጣንን ከርበን ገጸ በረኽተ ሀብዎ።	And they came into the house, and saw the young child with Mary, his mother; and falling down on their faces they gave him worship; and from their store they gave him offerings of gold, perfume, and spices.
ናብ ሄሮዶስ ከይምለሱ፡ ብሕልሚ ምስ ተራእዮም ድማ፡ ብኻልእ መገዲ ናብ ሃገሮም ተመልሱ።	And it was made clear to them by God in a dream that they were not to go back to Herod; so they went into their country by another way.
እቲ እግዚኣብሄር ብነብዩ፡ ንወደይ ካብ ግብጺ ጸዋዕኩዎ፡ ዝበሎ ምእንቲ ኺፍጸም ድማ፡ ሄሮዶስ ክሳዕ ዚመውት ኣብኡ ጸንሐ።	And was there till the death of Herod; so that the word of the Lord through the prophet might come true, Out of Egypt have I sent for my son.
ሄሮዶስ ምስ ሞተ፡ እንሆ ኣብ ግብጺ መልኣኽ እግዚኣብሄር ንዮሴፍ ብሕልሚ ተራእዮ፡	But when Herod was dead, an angel of the Lord came in a dream to Joseph in Egypt,
እቶም ህይወት እዚ ሕጻን ዚደልዩ ሞይቶም እዮም እሞ፡ ተንስእ፡ ነዚ ሕጻንን ነዲኡን ሒዝካ፡ ናብ ምድሪ እስራኤል ኪድ፡ በሎ።	Saying, Get up and take the young child and his mother, and go into the land of Israel: because they who were attempting to take the young child's life are dead.
ንሱ ድማ ተንሲኡ፡ ነቲ ሕጻንን ነዲኡን ሒዙ፡ ናብ ምድሪ እስራኤል መጸ።	And he got up, and took the young child and his mother, and came into the land of Israel.

Table V.1: sample parallel data from Corpus I

Appendix VI: Sample Parallel Corpus for training from Corpus II

ፌዴራል ነጋሪት ጋዜጣ ፌዴራላዊ ዴሞክራስያዊ ሪፖብሊክ ኢትዮጵያ	Federal negarit gazeta of the federal democratic republic of Ethiopia
ሓፈሻዊ ሕጋገታት	General provisions
እዚ ሕገ መንግስቲ ንኣታሃናንፃ ፌዴራላውን ዴሞክራሲያውን መንግስቲ ይሕግግ ።	This Constitution establishes a Federal and Democratic State structure.
በዚ መሰረት መንግስቲ ኢትዮጵያ ፌዴራላዊ ዴሞክራሲያዊ ሪፖብሊክ ኢትዮጵያ ብዝብል ስያመ ይፅዋዕ ።	Accordingly, the Ethiopian state shall be known as the Federal Democratic Republic of Ethiopia.
ዶብ ግዝኣት ኢትዮጵያ	Ethiopian territorial jurisdiction
ዶብ ግዝኣት ኢትዮጵያ ንዶብ ኣባላት እቲ ፌዴራል ዘጠቃልል ኮይኑ እቲ ብመሰረት ዓለም ለኻዊ ስምምዕ ዝተወሰነ እዩ ።	The territorial jurisdiction of Ethiopia shall comprise the territory of the members of the Federation and its boundaries shall be as determined by international agreements.
ሰንደቅ ዕላማ ኢትዮጵያ	The Ethiopian flag
ሰንደቅ ዕላማ ኢትዮጵያ ኣብ ላዕሊ ቐፅለዋይ ኣብ ማእኸል ብጫ ፣ ኣብ ታሕቲ ቀይሕ ኮይኑ ኣብ ማእኸሉ ሃገራዊ ኣርማ ይህልዎ ።	The Ethiopian flag shall consist of green at the top, yellow in the middle and red at the bottom, and shall have a national emblem at the center.
ሰለስቲኣም ሕብርታት ብማዕረ ተገዲሞም ይቅመጡ ።	The three colors shall be set horizontally in equal dimension.
ኣብቲ ሰንደቅ ዕላማ ዝቅመጥ ሃገራዊ ኣርማ ብሄራት ፣ ብሄረሰባት ፣ ህዝብታትን ሃይማኖታትን ኢትዮጵያ ብሓባርን ብማዕርነትን ንምንባር ዘለዎም ተስፋ ዘንፀባርቅ ይኸውን ።	The national emblem on the flag shall reflect the hope of the Nations, Nationalities, Peoples as well as religious communities of Ethiopia to live together in equality and unity.
ኣባላት እቲ ፌዴራል ነናይ ባዕሎም ሰንደቅ ዕላማን ኣርማን ክህልዎም ይኽእል ። ነቲ ዝርዝር ኣብ ነናይ ባዕሎም ቤት ምኽሪ ይውስኑ ።	Members of the Federation may have their respective flags and emblems and shall determine the details thereof through their respective legislatures.
ሃገራዊ መዝሙር ኢትዮጵያ	National Anthem of Ethiopia
ሃገራዊ መዝሙር ኢትዮጵያ ንዕላማታት እዚ ሕገ መንግስትን ህዝብታት ኢትዮጵያ ኣብ ዴሞክራሲያዊ ስርዓት ንምንባር ዘለዎም እምነት ከምኡ እውን ናይ መፃኢ ሓበራዊ ዕድሎም ዘንፀባርቅ ኮይኑ ክቅረፅ ብሕጊ ይውሰን ።	The national anthem of Ethiopia, to be determined by law, shall reflect the ideals of the Constitution, the Commitment of the Peoples of Ethiopia to live together in a democratic order and of their common destiny.
ኩሎም ቋንቋታት ኢትዮጵያ ብማዕረ መንግስታዊ ኣፍልጦ ይህልዎም ።	All Ethiopian languages shall enjoy equal state recognition.
ኣምሓርኛ ፣ ቋንቋ ስራሕ መንግስቲ ፌዴራል ይኸውን ።	Amharic shall be the working language of the Federal Government.
ኣባላት እቲ ፈዴሬሽን ነናይ ባዕሎም ቋንቋ ስራሕ ብሕጊ ይውስኑ ።	Members of the Federation may by law determine their respective working languages.
ዜጋታት ወፃኢ ሃገር ዜግነት ኢትዮጵያ ክጐናፀፉ ይኽእሉ ።	Foreign nationals may acquire Ethiopian nationality.
ዜግነት ብዝምልከት እቲ ዝርዝር ብሕጊ ይውሰን ።	Particulars relating to nationality shall be determined by law.
ናይ ፆታ ኣገላልፃ	Gender reference
ሓወይ ጋዘጠኛ እዩ ።	My brother is a journalist.
ወርቁ ንዓይ ኣይተረደእንን	Worku do not understand me.

ቴድሮስ ምስ ሰበይቱ ይበኣስ ኣሎ ።	Theodros is fighting with his wife.
ሙሉ ትእዛዝ ትቐበል ኣላ ።	Mulu is taking an order.
ሳራ መንሽሮ ኣለዋ ።	Sara got cancer.
ሚኪኤለ ይዛረብኣሎ ።	Micheal is talking.
ኢብራሂም ስርሑ ክለቅቕ ኣለዎ	Ibrahim need to quit his job.
ንሱ ንወዱ ደጊሙ ክሪኦ ኣይኽእልን	He will never see his kid again.
ሃናንሽጋራ ተትክኽ ኣላ ።	Hanan is smoking cigarette.
ጥላሁንሓሳዊ እዩ ።	Tilahun is a lair.
ኣነ ነቲ ሓቂ ክፈልጥ የብለይን።	I do not need to know the truth.
ሳሮንክልቢ ትፈርሕ እያ ።	Saraon is afraid of dogs.
ሀይሉ ተመሊሱ ናብቤት ማእሰርቲ ኣትዩ ።	Hailu is back to the prison.
ሳሙኤል ፀሊም ኣሜሪካዊ እዩ ።	Samuel is black American.
ንሕና ኢትዮጵያውያንኢና ።	We are Ethiopian.
ሰላም ናይ ዝኾነ ሰብነብሲ ኣጥፊኣ ።	Selam murdered a guy.
ማትያስ ንበቀለ ምኽያድ ንኽቁም ነጊርዎ ።	Matyas told Bekelle to stop walking.
ተስፋሁንንጡፍ ሰራሕተኛ እዩ ።	Tesfahun is a hard worker.
ጌታሁንሬድዮ የዳምፅ ኣሎ ።	Getahun is listening to the radio.
ገብረ ይንቅጥቀጥ ኣሎ ።	Gebre is shaking.
ተስፉ ፍቕሪ ሒዝዎ ።	Tesfu is in love.
ኣበበ ዝተናደደ ይመስል ።	Abebe looks angry.
ኣለሙ ናብቤት ትምህርቲ ከይዱ ።	Alemu went to school.
ኣልማዝ ነዋሕ እያ ።	Almaz is tall.
ንሱ ክምርያ እዩ ።	He is getting married.
ንሳ እንጀራ ትፈቱ እያ ።	She loves enjera.
ወልዴ ህያብሂቡኒ ።	Wolde gave me a gift.
ኣለሙ ቀታሊ ነብሲ እዩ ።	Alemu is a murderer.
ወልዴ ወደለ እዩ ።	Wolde is fat.
ኣማረ ንናቱ ኣዶ ኣዛሪቡዋ ።	Amare talked to his mother.
ኣበበ ነዓኣ ኣሚኑዋ ።	Abebe trusted her.
ኣልማዝ ቑማርተኛ እያ ።	Almaz is a gambler.
ኣለሚቱ መፅሓፍ ፅሒፉ ።	Alemitu wrote a book.
ተስፉ ተማሃራይ እዩ ።	Tesfu is a student.
ንሳቶም ነታ ሰበይቱ ኣብቅድሚኡ ቀቲሎማ ።	They killed his wife in front of him.
ሳምራዊት መኪና ምዝዋር ትኽእል እያ ።	Samrawit can drive a car.
ዳዊት ነቲ ዓርኩ ካብመኪና ንኽይወፅእ ነጊርዎ ።	Dawit told his friend not to get out of the car.
ፍቓዱ ኣብውሽጢ ፀገም ኣሎ ።	Fikadu is in trouble.
በላይ ነቲ ሓው ቐቲሉዎ ።	Belay killed his brother.
ሰበይቲ ሞገስ ነብሰ-ፀር እያ ።	Moges's wife is pregnant.
ማርታ ተፀሊእዋ ኣሎ ።	Marta is sick.
መስፍንብመዓልቱ ይሰክር እዩ ።	Mesfin gets drunk everyday.
ሳሙኤል ንመንግስቲ ይሰርሕ እዩ ።	Samuel works for the government.

Table VI.1: Sample Parallel data from Corpus II

Appendix VII: Sample translated output from Post Segmented System

English – Tigrigna Translation		
N.O	**English Input Sentences**	**Tigrigna Output Sentences**
1.	Ebriham is building a house.	ኣብራሃም ገዛ ይሰርሕ ኣሎ ።
2.	The doctor saved aster's life.	እቲ ዶክተር ኣስቴር ህይወት ኣትሪፉላ ን ።
3.	She will come back on friday.	ንሳ ተመሊሳ ክግፈፍ ኣይኽእልን come ኣብ friday ።
4.	The students are preparing for presentation.	እቶም ንፕረዘንቴሽን ይዘጋጀዉ ኣለዉ
5.	He is working on the company.	ንሱ ቋንቋ ኣብ ።
6.	Alex had a reason to get mad.	ኣሌክስ ኽናደድ ምኽንያት ነይርዎ ።
7.	My sister leaves outside addis ababa.	ናሃተይ ሓፍቲ ዝዋሃቡ outside ኣዲስ ኣበባ እዩ ።
8.	She is an information science student.	ንሳ ናይ ኢንፎርሜሽን ሳይንስ ተማሃራይ እዩ ።
9.	The flood has left a lot of people homeless.	እቲ flood ሓንፃጊ ካብ ተዛራባይ ንህዝቢ homeless ።
10	He was sentenced 20 years of jail.	ንሱ ን 20 sentenced ነይሩ ።
11	The wedding is tomorrow.	እቲ መርዓ ቐፃላይ ዕባሕ እዩ ።
12	Her fiancé was angry with her.	ናታ fiancé ብኣኣ ተናዲዱ ነይሩ ።
13	The economy has gone down because of corruption.	እቲ ኢኮኖሚ ኣንቆልቁሉ ልፍንቶም ን corruption ።
14	The movie is not scary.	እቲ ፊልሚ ዘፍርሕ ኣይኮነን ።
15	People died because of the car accident..	ህዝቢ ብልፍንቶም ብሓደጋ መኪና..
16	Sorry i have to go, you can come in.	ይቅሬታ ኣነ ክኸይድ ኣለኒ ፣ ንስኻ ክትኣቱ ትክእል እኻ ።
17	My family are calling me.	ናተይ ስድራ ንዓይ ይጽውዕኒ ኣለው ።
18	Meaza enjoys watching football.	Meaza ኩዕሶ ምርኣይ የሕጉሳ ።
19	Haymanot wants to be a president.	ሀይማኖት ፕሬዘዳንት ምኻን ይደሊ እዩ ።
20	Addis wants to be a teacher.	ኣዲስ መምህር ምኻን ትደሊ እያ ።
21	Human and democratic rights.	ሰብኣውን ዴሞክራሲያውን መሰላት ።
22	Human rights and freedoms, emanating from the nature of mankind, are inviolable and inalienable.	ሰብኣዊ መሰላትን ናፅነታትን emanating ካብ እታ mankind ፣ ናይ inviolable ን inalienable ።
23	Separation of state and religion.	Separation ንመንግስት ንሃይማኖት ።
24	Religion shall not interfere in state affairs.	ሃይማኖት ካብ ኣብ ጉዳይ interfere ክልል ።
25	Every person has the right to life.	ዝኾነ ሰብ በህይወት ናይ ምንባር መሰል ኣለዎ ።
26	Rights of persons accused.	ሰባት መሰል ቀሪቦም ።
27	Right to privacy.	መሰል ህይወት ።

Table VII.1: Sample English – Tigrigna Translation

Appendix VIII: Sample translated output from Post Segmented System

Tigrigna – English Translation		
N.O	**Tigrigna Input Sentences**	**English Output Sentences**
1.	ኣብራሃም ገዛ ይሰርሕ ኣሎ ።	Ebriham is building a house.
2.	እቲ ዶክተር ናይ ኣስቴር ህይወት ኣትሪፉላ ።	The doctor saved aster for life.
3.	ንሳ ዓርቢ ተመሊሳ ትመጽእ እያ ።	She ዓርቢ will come back on tuesday.
4.	እቶም ተማህሮን ፕረዘንቴሽን ይዘጋጀዉ ኣለዉ ።	The ተማህሮ are preparing for presentation..
5.	ንሱ ኣብቲ ድርጅት ይሰርሕ ኣሎ ።	He is in ድርጅት is working on it.
6.	ኣሌክስ ንቕናደድ ምኽንያት ነይርዎ ።	ኣሌክስ had a reason to get mad.
7.	ናሃተይ ሓፍቲ እትነብረሉ ካብ ኣዲስ ኣበባ ወጻኢ እዩ ።	My sister እትነብረሉ from addis ababa ወጻኢ.
8.	ንሳ ናይ ኢንፎርሜሽን ሳይንስ ተማሃሪት እያ ።	She is an information science is a student.
9.	እቲ ውሕጅን ብዙሓት ሰባት ገዛ ኣልባ ገይሩ ።	The a and ብዙሓት people house ኣልባ peoples.
10	ንሱ ን 20 ዓመታት ማእሰርቲ ተፈሪድዎ ።	He 20 ዓመታት wards ተፈሪድዎ.
11	ስነስርዓት እቲ መርዓ ፅባሕ እዩ ።	For the the መርዓ is tomorrow.
12	ናታ ሕፁይ ብኣኣ ተናዲዱ ነይሩ ።	Her candidate was angry with her.
13	እቲ ኢኮኖሚ ብምኽንያት ግዕዝይና ኣንቆልቁሉ ።	The ኢኮኖሚ a ግዕዝይና ኣንቆልቁሉ.
14	ኣማረ ንናቱ ኣቦ ኣዛሪቡዎ ።	Amare talked to his father ኣዛሪቡዎ.
15	እቲ ፊልሚ ዘፍርሕ ኣይኮነን።	The movie is not scary.
16	እቶም ሰባት ብምኽንያት ሓደጋ መኪና ሞይቶም ።	The people a of a car accident.
17	ይቅሬታ ኣነ ክኸይድ ኣለኒ: ንስኻ ክትኣቱ ትክእል እኻ ።	Sorry i: you can come in.
18	ስድራይ ንዓይ ይጽውዕኒ ኣለው ።	ስድራይ are calling me.
19	ማዓዛ ኩዕሶ ምርኣይ የሕጉሳ ።	ማዓዛ enjoys watching football.
20	ሀይማኖት ፕሬዝዳንት ምኻን ትደሊ እያ ።	Haymanot wants to be a president.
21	ኣዲስ መምህር ምኻን ትደሊ እያ ።	Addis wants to be a teacher.
22	ሰብኣውን ዴሞክራሲያውን መሰላት ።	Human and democratic rights.
23	ሰብኣዊ መሰላትን ናፅነታት ን ካብ ተፈጥሮ ዘርኢ ሰብ ዝፍልፍሉ ዘይጣሓሱ ን ዘይጋሃሱ ን እዮም ።	Human rights and freedoms of natural race and collect ዘይጣሓሱ and ዘይጋሃሱ and are
24	ምፍልላይ መንግስትን ሃይማኖትን ።	ምፍልላይ state and religion.
25	ሃይማኖት እውን ኣብ ጉዳይ መንግስቲ ኢዱ ኣየእትውን ።	Religion, the internal affairs of government ኢዱ ኣየእትውን.
26	ዝኾነ ሰብ በህይወት ናይ ምንባር መሰል ኣለዎ ።	Every person has the right to life.
27	መሰል ዝተኸሰሱ ሰባት ።	The ዝተኸሰሱ persons.
28	መሰል ምኽባርን ሓለዋን ውልቀ ህይወት ።	To the right to protection to.
29	ዝኾነ ሰብ ብዘይ ዝኾነ ኢድ ኣእታውነት ዝመሰሎ ኣረኣእያ ክሕዝ ይኽእል ።	Any without any interference ዝመሰሎ opinion ክሕዝ may
30	ዓንቀፅ 38	Article 38

Table VIII.1: Sample Tigrigna – English Translation

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 all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name : *Mulubrhan Hailegebreal*

Signature :_____

Date : June, 2017

Confirmed by advisor :

Name : *Martha Yifiru (PhD)*

Signature : _____

Date : June, 2017